Aplicações em R:

Encurtando Distâncias nas Ciências

Organização

Dra. LUCIANE FERREIRA ALCOFORADO
Dr. JOÃO PAULO MARTINS DOS SANTOS
Dr. ARIEL LEVY
Dr. ORLANDO CELSO LONGO
Dr. JUAN LÓPEZ LINARES

Textos motivados a partir de palestras do VII Seminário Internacional de Estatística com R

ORGANIZADORES
LUCIANE FERREIRA ALCOFORADO
JOÃO PAULO MARTINS DOS SANTOS
ARIEL LEVY
ORLANDO CELSO LONGO
JUAN LÓPEZ LINARES

Aplicações em R: Encurtando Distâncias nas Ciências

DOI: 10.11606/9786587023397

Pirassununga - SP
FACULDADE DE ZOOTECNIA E ENGENHARIA DE ALIMENTOS (FZEA)
UNIVERSIDADE DE SÃO PAULO (USP)
2024

UNIVERSIDADE DE SÃO PAULO

Reitor: Prof. Dr. Carlos Gilberto Carlotti Junior

Vice-Reitora: Profa. Dra. Maria Arminda do Nascimento Arruda

FACULDADE DE ZOOTECNIA E ENGENHARIA DE ALIMENTOS

Avenida Duque de Caxias Norte, 225 - Pirassununga, SP

CEP 13.635-900

http://www.fzea.usp.br

Diretor: Prof. Dr. Carlos Eduardo Ambrósio

Vice-Diretor: Prof. Dr. Carlos Augusto Fernandes de Oliveira

Capa: Ariel Levy

Diagramação: João Paulo Martins dos Santos

Dados Internacionais de Catalogação na Publicação

Serviço de Biblioteca e Informação da Faculdade de Zootecnia e Engenharia de Alimentos da Universidade de São Paulo

```
          Alcoforado, Luciane Ferreira (org.)
A354a       Aplicações em R : encurtando distâncias nas
          ciências / Luciane Ferreira Alcoforado (org.), João
          Paulo Martins dos Santos (org.), Ariel Levy (org.),
          Orlando Celso Longo (org.), Juan López Linares (org.).
          -- Pirassununga : Faculdade de Zootecnia e Engenharia de
          Alimentos da Universidade de São Paulo, 2024.
             286 p.

             ISBN 978-65-87023-39-7 (e-book)
             DOI: 10.11606/9786587023397


             1. Linguagem R. 2. Visualização. 3. Estatística.
          4. Matemática. I. Santos, João Paulo Martins dos (org.).
          II. Levy, Ariel (org.). III. Longo, Orlando Celso
          (org.). IV. López Linares, Juan (org.). V. Título.
```

Ficha catalográfica elaborada por Girlei Aparecido de Lima, CRB-8/7113

"Que cada linha escrita aqui seja um passo mais próximo do entendimento e da inovação, iluminando o caminho da descoberta. Com gratidão, dedicamos esta obra àqueles que nunca cessam de questionar, aprender e avançar."

AGRADECIMENTOS

Agradecemos a todos que incentivaram de forma direta ou indireta na produção deste E-book. Essa obra, assim como outras, é uma concretização de inúmeras colaborações de pessoas e instituições. Dessa forma, deixamos registrado agradecimentos a todos os autores que empenharam esforços para a produção do capítulo, como também agradecemos as respectivas instituições em que atuam. De forma específica, os organizadores registram os agradecimentos a Orlando Celso Longo (Universidade Federal Fluminense-UFF/RJ), Ariane Hayana Thomé de Farias (Tribunal de Justiça do Estado de Roraima - TJRR), Marcus Antônio Cardoso Ramalho (PPGAd da Universidade Federal Fluminense - UFF/RJ), Manuel Febrero-Bande (Departamento de Estadística, Análisis Matemático y Optimización, Universidade de Santiago de Compostela), João Paulo Martins dos Santos (Academia da Força Aérea, AFA/SP), Luciane Ferreira Alcoforado (Academia da Força Aérea, AFA/SP), Marco Aurélio Chaves Ferro (Universidade Federal Fluminense-UFF/RJ), Felipe Rafael Ribeiro Melo (Departamento de Métodos Quantitativos da Universidade Federal do Estado do Rio de Janeiro (DMQ/UNIRIO)), María José Ginzo Villamayor (Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela), Thiago de Oliveira Pires (F&O/CIO/International Business Machines Corporation) e Ariel Levy (Universidade Federal Fluminense-UFF/RJ). Os organizadores agradecem o apoio, direto ou indireto, recebido de cada uma das instituições associadas: Ariel Levy (Universidade Federal Fluminense), João Paulo Martins dos Santos (Academia da Força Aérea), Luciane Ferreira Alcoforado (Academia da Força Aérea), Orlando Celso Longo (Universidade Federal Fluminense-UFF/RJ), Juan López Linares (Faculdade de Zootecnia e Engenharia de Alimentos (FZEA) da Universidade de São Paulo-FZEA/USP/SP). Por fim, agradecemos a Profa Dra. Maysa Sacramento de Magalhães da Escola Nacional de Ciências Estatísticas (ENCE/RJ), ao Prof. Dr. Wenceslao González Manteiga do Departamento de Estadística, Análisis Matemático y Optimización da Universidade de Santiago de Compostela e ao Prof. Dr. Steven Dutt Ross da Universidade Federal do Estado do Rio de Janeiro (UNIRIO/RJ).

ORGANIZADORES

Dra. LUCIANE FERREIRA ALCOFORADO

https://orcid.org/0000-0002-9504-8087. Possui graduação em Licenciatura em Matemática pela Universidade Federal de Santa Maria (1994), mestrado em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (1998) e Doutora em Engenharia Civil pela Universidade Federal Fluminense (2009). É professora na Academia da Força Aérea em Pirassununga/SP e professora colaboradora no Programa de Pós-Graduação em Engenharia Civil da UFF. Tem experiência na linguagem R, análise e visualização de dados e em problemas de otimização. Sua atuação na área de Matemática e Estatística destaca-se nos seguintes temas: regressão linear, regressão logística, testes de hipóteses, análise multivariada, métodos de apoio à tomada de decisão como Simplex, AHP entre outros. É líder do grupo de Pesquisa Estatística com R, coordenadora do SER - Seminário Internacional de Estatística com R, autora de diversos livros sobre a linguagem R e dos pacotes disponíveis no CRAN, MandalaR e AHPWR. Textos completos e gratuitos das publicações da autora podem ser encontrados aqui.

Dr. JOÃO PAULO MARTINS DOS SANTOS

https://orcid.org/0000-0002-0957-7119. Possui graduação em Licenciatura em Matemática pela Universidade Estadual Paulista Júlio de Mesquita Filho (2006), mestre em Matemática pela Universidade Estadual Paulista Júlio de Mesquita Filho (2009) e Doutor em Ciências pela Escola de Engenharia de São Carlos - EESC-USP. É professor na Academia da Força Aérea em Pirassununga/SP. Tem experiência na área de Sistemas Dinâmicos não lineares e não ideais com pesquisa desenvolvida em métodos de perturbação. Tem experiência na área de Matemática e interesse nos seguintes temas: método numéricos para solução de equações diferenciais ordinárias e parciais, estimador de erro do tipo residual para a equação do transporte de poluentes, linguagem Python de programação, Computação Científica em Python e métodos numéricos para solução de sistemas lineares. Textos completos e gratuitos das publicações do autor podem ser encontrados aqui.

Dr. ORLANDO CELSO LONGO https://orcid.org/0000-0002-0323-473X. Graduado em Engenharia Civil, Mestrado em Engenharia Civil e Doutorado em Engenharia de Transportes. Atualmente é Professor Titular da Universidade Federal Fluminense. Coordenador do Programa de Pós-graduação em Engenharia Civil da Universidade Federal Fluminense no período 2005 – 2013 e 2017 até a data atual. Diretor do DATAUFF de 2019 até data atual. Coordenou vários eventos nacionais e internacionais tais como IV Semana de Engenharia - III Seminário Fluminense de Engenharia, 4th International Conference on the Behaviour of Damaged Structures, VII Seminário Internacional de Estatística com R. Autor do pacote disponível no CRAN AHPWR. Tem experiência na área de Engenharia Civil e ambiente construído com ênfase em

Construção Civil, atuando principalmente nos seguintes temas: construção civil, custos, gerenciamento / acompanhamento fiscalização, orçamento, administração de projetos e elaboração e desenvolvimento de projetos de infraestrutura para cidades inteligentes.

Dr. ARIEL LEVY https://orcid.org/0000-0003-3557-1201. Doutor em Economia (Universidade Federal Fluminense - 2013), mestre em Administração (IBMEC -2003) e engenheiro eletricista (Universidade Federal Fluminense - 1982). Professor Associado da Universidade Federal Fluminense vinculado ao Departamento de Administração da Faculdade de Administração e Ciências Contábeis e fui coordenador do Curso de Graduação em Administração (2016-2021). Professor do quadro permanente do PPGAd - UFF e colaborador no MBA de Logística (LOGEMP - UFF), no MBA de Finanças (UFF), no MBA de Marketing (UFF) e dos cursos de Extensão em Ciências dos Dados (UFF) onde atua como coordenador. Possui experiência em Administração, com ênfase em Finanças Quantitativas; Finanças públicas; Planejamento e Controle; na linguagem R e na análise e visualização de dados. Organizador dos Seminários de Estatística R - Evento Internacional de Divulgação de Aplicações e Desenvolvimento de Linguagens R. Coordenador do Grupo de Pesquisa (CNPQ/UFF) - Métodos Quantitativos Aplicados à Administração.

Dr. JUAN LÓPEZ LINARES https://orcid.org/0000-0002-8059-0631. Professor Associado do Departamento de Ciências Básicas (ZAB) da Faculdade de Zootecnia e Engenharia de Alimentos (FZEA) da Universidade de São Paulo (USP). Atualmente ministra as disciplinas de Cálculo II e IV para estudantes de engenharias e os cursos de "Treinamento Olímpico em Matemática para estudantes do Ensino Fundamental e Médio" e "Geometria olímpica com GeoGebra" para professores. Desenvolve projetos de pesquisa nas áreas de ensino de Cálculo e na resolução de problemas de Olimpíadas. Graduação e Mestrado em Física na Universidade da Havana, Cuba, em 1994 e 1996, respetivamente. Curso de Diploma da Matéria Condensada no Centro Internacional de Física Teórica Abdus Salam, em Trieste, na Itália em 1997-1998. Estágio no Instituto de Espectroscopia Molecular (CNR), Bolonha, Itália em 1998-1999. Doutor em Física pela Universidade Federal de São Carlos (UFSCar) em 1999-2001. Pós-doutorado de 4 anos (2002-2005) na Universidade Estadual de Campinas (Unicamp). Mestre Profissional em Matemática em Rede Nacional (PROFMAT) pela UFSCar em 2019 e Livre Docente na área de Ensino de Matemática Olímpica na FZEA USP em 2022. Textos completos e gratuitos das publicações do autor podem ser encontrados aqui.

AUTORES

ORLANDO CELSO LONGO (*orlandolongo@id.uff.br*)

https://orcid.org/0000-0002-0323-473X. Graduado em Engenharia Civil, Mestrado em Engenharia Civil e Doutorado em Engenharia de Transportes. Atualmente é Professor Titular da Universidade Federal Fluminense. Coordenador do Programa de Pós-graduação em Engenharia Civil da Universidade Federal Fluminense no período 2005 – 2013 e 2017 até a data atual. Diretor do DATAUFF de 2019 até data atual. Coordenou vários eventos nacionais e internacionais tais como IV Semana de Engenharia - III Seminário Fluminense de Engenharia, *4th International Conference on the Behaviour of Damaged Structures*, VII Seminário Internacional de Estatística com R. Autor do pacote disponível no CRAN AHPWR. Tem experiência na área de Engenharia Civil e ambiente construído com ênfase em Construção Civil, atuando principalmente nos seguintes temas: construção civil, custos, gerenciamento / acompanhamento fiscalização, orçamento, administração de projetos e elaboração e desenvolvimento de projetos de infraestrutura para cidades inteligentes.

ARIANE HAYANA THOMÉ DE FARIAS (*ariane.hayana@gmail.com*)

https://orcid.org/0000-0003-1571-8739. Graduada em Estatística e Economia, ambas pela Universidade Federal do Amazonas (UFAM), com MBA em Perícia e Auditoria Econômico-Financeira pelo Instituto de Pós-Graduação (IPOG). Atua como Assessora Estatística no Tribunal de Justiça do Estado de Roraima (TJRR) e possui conhecimentos em linguagens de programação R e Python, com proficiência no desenvolvimento de aplicativos em R/Shiny, bem como na elaboração de relatórios reprodutíveis com R Markdown/Quarto. Tem interesse em Jurimetria, Processamento de Linguagem Natural (PLN) e Machine Learning.

MARCUS ANTONIO CARDOSO RAMALHO (*marcusantonio@id.uff.br*)

https://orcid.org/0009-0003-9282-7098. Possui graduação em Administração pela Universidade Federal Fluminense (2020) e é candidato ao título de mestre em Administração pelo programa de pós-graduação em administração da UFF (PPGAd-UFF). É professor convidado dos MBA's de Ciências de Dados e de Finanças Corporativas e Mercados de Capitais na UFF. Tem experiência em ciência de dados com R e Python, programação funcional e desenvolvimento de bots, mapeamento e automação de processos administrativos. Tem interesse em Administração da Informação, Gestão do Conhecimento Pessoal, Economia Política, Finanças, R e Python.

MANUEL FEBRERO-BANDE (*manuel.febrero@usc.es*)

https://orcid.org/0000-0002-9536-2973. Manuel Febrero-Bande é Professor de Estatística e Pesquisa Operacional na Universidade de Santiago de Compostela onde se licenciou em Ma-

temática (1990) e defendeu a sua tese de doutoramento (1995). Publicou mais de 70 artigos em revistas internacionais em diversas áreas, embora fundamentalmente relacionados com Séries de Tempo, Estatística Espacial, Dados Funcionais, Estatística Computacional e, em geral, Métodos Estatísticos aplicados ao meio ambiente, Bioestatística e finanças. Orientou 5 teses de doutorado (mais duas em andamento) e foi coordenador acadêmico do Mestrado Interuniversitário em Técnicas Estatísticas (2008-2012) e do Doutorado Interuniversitário em Estatística e Pesquisa Operacional (2013-2016), em ambos os casos organizados pelas universidades de Santiago de Compostela, Vigo e A Coruña. Realizou visitas de pesquisa ministrando cursos em universidades e centros em mais de 10 países. É um programador especialista em R e *Shiny* e, em particular, é coautor da biblioteca *fda.usc* dedicada à análise de dados funcionais.

JOÃO PAULO M. DOS SANTOS (*jp2@alumni.usp.br*)

https://orcid.org/0000-0002-0957-7119. Graduado em Licenciatura Plena em Matemática pela Universidade Estadual Paulista Júlio de Mesquita Filho (2006), mestre em Matemática pela Universidade Estadual Paulista Júlio de Mesquita Filho (2009) e Doutor em Ciências pela Escola de Engenharia de São Carlos - EESC-USP. Docente na Academia da Força Aérea em Pirassununga/SP, e colaboradora no Programa de Pós-Graduação em Eng. Civil (UFF). Tem interesse em Matemática Aplicada e Estatística.

LUCIANE FERREIRA ALCOFORADO (*lucianea@id.uff.br*)

https://orcid.org/0000-0002-9504-8087. Graduada em Matemática (UFSM) , Mestre em Engenharia de Sistemas e Computação (UFRJ) e Doutora em Engenharia Civil (UFF). Docente na Academia da Força Aérea, e colaboradora no Programa de Pós-Graduação em Eng. Civil (UFF), autora de diversos livros sobre a linguagem R e pacotes publicados no CRAN como o MandalaR e o AHPWR. Atua na disseminação da linguagem R no Brasil.

MARCO AURÉLIO CHAVES FERRO (*marcoferro@id.uff.br*)

https://orcid.org/0000-0002-8198-8668. Professor dos Cursos de Graduação e Pós-Graduação em Engenharia Civil da Universidade Federal Fluminense (UFF). Graduado em Engenharia Civil pela Universidade Federal do Rio de Janeiro (UFRJ) em 1987, Mestre pela COPPE/UFRJ em 1997, Doutor pela COPPE/UFRJ em 2002 e Pós-Doutor pela COPPE/UFRJ em 2008 e pela FGV/EBRAPE em 2012. Atua nas áreas de simulação numérica e análise e cálculo estrutural. Possui interesse em Métodos Numéricos e Inteligência Artificial em Engenharia.

FELIPE RAFAEL RIBEIRO MELO (*felipe.ribeiro@uniriotec.br*)

https://orcid.org/0000-0002-1482-8533. Doutor em Estatística pela Universidade Federal do Rio de Janeiro (UFRJ) e, desde 2014, professor adjunto do Departamento de Métodos Quantitativos da Universidade Federal do Estado do Rio de Janeiro - DMQ/UNIRIO, tem

interesse nas áreas de Estatística e de Probabilidade, sobretudo em temas voltados ao ensino destas áreas, além de interesse permanente na linguagem de programação R. Em sua atuação docente, ministra disciplinas de Probabilidade para os cursos de bacharelado em Engenharia de Produção e Sistemas de Informação ninterruptamente desde 2019 e coordena o projeto de pesquisa envolvendo análise de dados provenientes da avaliação da gestão coletiva do trabalho dos servidores técnico-administrativos da UNIRIO, além de ser autor de duas apostilas do pacote R *Commander* do software R.

MARÍA JOSÉ GINZO VILLAMAYOR (*mariajose.ginzo@usc.es*)

https://orcid.org/0000-0001-6392-3812. María José Ginzo é professora assistente doutora do Departamento de Estatística, Análise Matemática e Otimização (USC) desde 2023 e pesquisadora desde 2008. Licenciada em Matemática com especialização em Estatística e Pesquisa Operacional, pela USC, fez pós-graduação em Estatística pela Universidade do Porto (Portugal), mestrado interuniversitário em Técnicas Estatísticas (USC) e obteve o doutoramento em Estatística e Investigação Operacional em maio de 2022 com a tese intitulada "Técnicas Estatísticas em Geolinguística. Modelagem Onomástica"sob a supervisão da Profa. Dra. Rosa Mª Crujeiras Casais. Ministrou cursos de Estatística com R em instituições como AGACA, Arcelor, Consello de Contas de Galicia, Misión Biológica de Galicia (CSIC), Tecnocom ou na própria USC, entre outras. Participa da comissão organizadora e científica das Jornadas de Usuários de R na Galícia, https://www.r-users.gal/ desde 2015. É coautora do pacote R *FORTLS: Automatic Processing of Terrestrial-Based Technologies Point Cloud for Forestry Purposes*.

THIAGO DE OLIVEIRA PIRES (*thop100@hotmail.com*)

https://orcid.org/0000-0003-4535-5537. Tenho graduação em Estatística (IME/UERJ), MSc. em Epidemiologia (ENSP/FIOCRUZ) e DSc. em Engenharia Biomédica (PEB/COPPE/UFRJ). Atualmente tenho atuado como Cientista de Dados na IBM. Tenho interesses em estatística, psicometria, otimização, cloud, sistemas embarcados e linguagens de programação (R e Python).

ARIEL LEVY (*alevy@id.uff.br*)

https://orcid.org/0000-0003-3557-1201. Doutor em Economia (Universidade Federal Fluminense - 2013), mestre em Administração (IBMEC -2003) e engenheiro eletricista (Universidade Federal Fluminense - 1982). Professor Associado I da Universidade Federal Fluminense vinculado ao Departamento de Administração na Faculdade de Administração e Ciências Contábeis e fui coordenador do Curso de Graduação em Administração (2016-2021). Professor do quadro permanente do PPGAd - UFF e colaborador nos Curso de Especialização em Administração Pública da UFF (CEAP), no MBA de Logística (LOGEMP - UFF), no MBA de Finanças (UFF), no MBA de Marketing(UFF) e no MBA de Ciências dos Dados (UFF) onde atuo como

coordenador. Possuo experiência em Administração, com ênfase em Finanças Quantitativas; Finanças Públicas; Planejamento e Controle. Organizador dos Seminários de Estatística com R - Evento internacional de divulgação de aplicações e desenvolvimento da linguagem R. Coordenador do grupo de pesquisa (CNPQ/UFF) - Métodos Quantitativos Aplicados à Administração.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Título

Aplicações em R: Encurtando distâncias nas Ciências

Prefácio

As mudanças do mundo atual têm ocorrido cada vez mais rapidamente, as quais têm uma relação direta com o desenvolvimento tecnológico. A tecnologia tem permitido o acesso à informação em frações de segundo, a coleta, o armazenamento, a análise de bases de dados enormes, a disseminação de informação e de conhecimento, em diferentes meios e formas, e estes são apenas alguns dos papéis que a mesma tem impactado nossas vidas.

Neste contexto de mundo veloz, de tecnologia trazendo avanços, mas também riscos, impondo desafios e demandando soluções, nasce o SER – Seminário Internacional de Estatística com R.

A proposta de criação de um evento que reunisse profissionais e estudantes de distintas áreas do conhecimento que fizessem uso da linguagem R, tratando de questões de análise de dados e que tivesse um caráter internacional foi uma visão atualizada e de futuro do que vinha acontecendo no mundo da Estatística com a utilização dessa linguagem de programação livre e de código aberto.

O SER já não nasceu timidamente, pois no primeiro evento que ocorreu em maio de 2016 foram mais que duzentos participantes, mostrando assim a necessidade no mundo acadêmico quanto profissional dessa reunião da comunidade usuária do R, para discussão, apresentação, disseminação e conhecimento dos seus trabalhos bem como de aprendizagem com outros profissionais.

A concepção de um e-Book composto por textos provenientes de palestras apresentadas no VII Seminário Internacional de Estatística com R - SER que ocorreu em 2023 é muito feliz, pois celebra um projeto vitorioso ao longo desses oito anos de sua existência[1].

O título do livro, “Aplicações em R: Encurtando Distâncias nas Ciências” foi muito bem escolhido e é muito bonito, pois em poucas palavras resume o

[1]Em 2020, o SER não ocorreu devido à pandemia de Covid-19

cerne do conteúdo da obra. Os textos do livro fazendo uso da linguagem R apresentam aplicações e abordagens em diversos campos de conhecimento, tais como, engenharia, estatística, gestão do conhecimento, matemática, métodos de apoio à decisão, probabilidade, web scraping.

O livro é composto por 11 capítulos, reunindo contribuições de 11 autores de instituições diversas do país como do exterior.

O primeiro capítulo, de autoria de Orlando Longo, nos traz toda a beleza da história do SER, a sua origem, o seu desenvolvimento, o seu crescimento, bem como a importância do mesmo no cenário brasileiro e da América Latina. Nos traz o papel deste evento que abrange várias dimensões, seja na formação de pessoas, na colaboração, na disseminação e atualização do conhecimento sobre a linguagem de programação R em suas diversas aplicações em várias áreas do conhecimento.

Tanto o capítulo 2, escrito por Ariane de Farias, quanto o capítulo 3, de autoria de Marcus Ramalho e Ariel Levy, fazem uso do Quarto, um sistema de código aberto para publicação científica e técnica. O capítulo 2 traz a construção de um livro desenvolvido no Quarto, apresentando um exemplo de construção de um livro específico e a sua publicação no Quarto Pub. Já no capítulo 3, o Quarto é empregado como uma ferramenta de gestão de conhecimento pessoal na pesquisa científica. Os autores tendo como base o modelo "seek, sense and share" de Harold Jarche introduzem os conceitos de gestão de conhecimento pessoal e, como os mesmos podem ser considerados nas fases de desenvolvimento de uma pesquisa científica.

No capítulo 4, Manuel Febrero-Bande apresenta uma introdução a alguns modelos clássicos de Séries Temporais (ST), bem como as etapas de modelagem de uma ST utilizando a linguagem R.

O capítulo 5 de João Paulo Martins dos Santos apresenta o método de coloração gradiente e o método de coloração por rotações sucessivas para a coloração de figuras construídas por meio de movimentos rígidos de rotação, translação e homotetias. Para tal utiliza o R/Rposit.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

No capítulo 6, Luciane Alcoforado apresenta o método Analytic Hierarchy Process (AHP) como uma técnica de apoio à decisão, apresentando os passos para a aplicação do mesmo. O pacote AHPWR, desenvolvido pela autora, Sousa e Longo, foi utilizado para implementar o AHP na linguagem R, apresentando também as funções disponíveis e os exemplos de código.

Metodologias de Inteligência Artificial (IA) empregadas na Engenharia, em especial na Engenharia Civil, são apresentadas no capítulo 7, cujo autor é Marco Ferro. O autor também apresenta diversas aplicações de IA e um exemplo da predição da resistência de concreto utilizando Redes Neurais Artificiais com implementação no código R.

No capítulo 8 de autoria de Felipe Ribeiro, a solução do Problema de Aniversário, um problema clássico em Probabilidade, é apresentado de forma didática. O autor apresenta os pacotes IPSUR e RcmdrPlugin.IPSUR, uma vez que trazem funções associadas ao problema estudado, bem como apresenta também a interface R Commander. Uma dinâmica em sala de aula para alunos do ensino médio e do ensino superior de como apresentar e solucionar o problema pelo professor também é proposta pelo autor.

De autoria de Maria José Villa Mayor, o capítulo 9 tem como objetivo classificar os sobrenomes da Galícia em uma das três categorias - apelativos, toponímicos, patronímicos, utilizando técnicas de web scraping. No trabalho foi utilizado a linguagem R para a extração dos dados.

No capítulo 10, Thiago Pires apresenta recursos do DuckDB, um sistema de gerenciamento de banco de dados de código aberto, que de acordo com o autor é adequado para lidar com grandes bases de dados, e sua interação com a linguagem R. O autor apresenta também alguns exemplos de uso do DuckDB, tais como, para mineração de dados, em dados de Covid-19, em dados de serviço de armazenamento na nuvem, em análise de dados espaciais.

Escrito por Ariel Levy e Marcus Ramalho, o capítulo 11 apresenta o planejamento de uma pesquisa em escala Likert desde o seu início até a análise dos

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

resultados. Os autores utilizam como base de dados o PISA 2009 (em português, Programa Internacional de Avaliação de Estudantes) e empregam o Quarto na construção do documento e o pacote Likert do R, os pacotes tydeverse e gt para a análise dos resultados, geração de gráficos e tabelas.

Assim, este livro organizado por Luciane Alcoforado, João Paulo Martins dos Santos, Orlando Longo, Ariel Levy e Juan López Linares, traz uma contribuição valorosa aos usuários da linguagem R de distintas áreas de conhecimento que trabalham com análise de dados.

Maysa Sacramento de Magalhães
Pesquisadora do Programa de Mestrado e Doutorado da ENCE
Coordenadora - Geral da ENCE (de agosto de 2014 a setembro de 2023)

Palavras-chave: Linguagem R, Visualização, Estatística, Matemática.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Lista de Figuras

1.1 Ranking das linguagens em 2015. 27
1.2 Início do R Consortium em 2015. 28
1.3 Livro de Introdução ao R. 29
1.4 Projeto Estatística é com R!. 29
1.5 Homenagem ao prof. Djalma Pessoa. 36
1.6 Canal de vídeos do "SER". 39

2.1 Estrutura. 46
2.2 Tela do `_quarto.yml`. 48
2.3 Títulos/Subtítulos ou Seções/Subseções. 49
2.4 Estruturas para o texto. 50
2.5 Nomenclaturas e referências. 50
2.6 Resultado exemplificativo. 52
2.7 Configurações de blocos de código. 53
2.8 Configuração no terminal. 54
2.9 Capa do modelo. 55
2.10 Capítulo *Introdução*. 58
2.11 Exemplo de *Referências* no *Quarto Book*. 59

3.1 Instalação do QUARTO em ambiente Linux. 65
3.2 Exemplo de cabeçalho YAML. 66
3.3 Fluxo de renderização de um documento usando o QUARTO. . . . 66
3.4 Modelo *seek sense share*. 68

3.5 Menu para acessar a janela de citações. 71
3.6 Criando um projeto ou arquivo QUARTO. 72
3.7 Exemplo de uso de pacotes Latex no YAML. 73
3.8 Fluxo sugerido de criação e publicação de um blog com QUARTO e GitHub Pages. 75
3.9 Configuração da pasta raiz para renderização no GitHub Pages. . . 76

4.1 Promedio mensual de la cotización del BitCoin 90
4.2 Cierre diario de GOOG con bandas de Bollinger y ajuste por medias móviles 92
4.3 Promedio mensual cierre de GOOG (formato xts) 94
4.4 Gráfico de Google 96
4.5 Últimos dos años del precio de apertura de GOOG 96
4.6 Descomposición clásica de la serie log(*GOOG*) 100
4.7 Serie log(*GOOG*) y una estimación por media móvil (izda) 102
4.8 ACF y PACF de la serie *AirPassengers* transformada por logaritmos 108
4.9 ACF y PACF de la serie log(*GOOG*) (primera fila) y de su diferencia (segunda fila) 109
4.10 Gráficos de diagnóstico para ARIMA(2,1,0) 112
4.11 Salida de la función `tsdisplay` 115
4.12 Predicciones a $h = 24$ con el modelo ARIMA(2,1,0) (stats) y el ARIMA(1,1,2)x(1,0,0) (forecast) 116
4.13 Cuantiles de predicción obtenidos a partir de la aproximación normal (izda) y por Bootstrap (dcha) 118
4.14 Retornos de log(*GOOG*) 120
4.15 ACF y PACF de la serie de retornos (primera fila) y de la serie de retornos al cuadrado (segunda fila) 122
4.16 Gráfico de los retornos con el intervalo dado por la desviación condicional (izda) 128

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

4.17 Gráfico de la densidad de los residuos estandarizados (izda) 129
4.18 Predicción de la serie con sus intervalos de confianza (izda) y de $\hat{\sigma}_t$ (dcha) 130
4.19 Gráficos de predicción obtenidos con Bootstrap 131

5.1 Gráfico da função $f : \mathbb{R} \longrightarrow \mathbb{R}$ $f(x, y) = \frac{\text{sen}\left(\sqrt{x^2+y^2}\right)}{\sqrt{x^2+y^2}}$. 136
5.2 Curvas planas: circunferência, elipse, astroide, cardioide, lemniscata de Gerono e lemniscata de Bernoulli. 139
5.3 Composição de rotações, e homotetias. 140
5.4 Construções com base em curvas planas. 141
5.5 Ilustração do método de coloração denominado sequencial. 143
5.6 Construções coloridas utilizando modelo de cores RGB (*Red, Green, Blue*) com o método sequencial. 144
5.7 Amostras aleatórias com reposição das cores $colors()[1 : 27]$ e $colors()$. 145
5.8 Cores gradientes utilizando o mapa $rainbow()$ com distintas quantidades de pontos. 146
5.9 Cores gradientes utilizando o mapa $heat.colors()$ com distintas quantidades de pontos. 147
5.10 Método sequencial com cores aplicadas às homotetias de razão crescente e decrescente, respectivamente. 149
5.11 Rotações sucessivas e respectivas homotetias de um cardioide com cores determinadas em $colors()[1 : 53]$ e amostra aleatória de 73 elementos com reposição da lista de cores $colors()$. 150
5.12 Rotações sucessivas com cores $colors()[i], i = 1, \cdots, 90$ e $t \in [0, 4\pi]$ (esquerda) e rotações sucessivas com homotetias com $colors()[i], i = 1, \cdots, 90$ e $t \in [0, 2\pi]$ (direita). 152
5.13 Rotações sucessivas da Lemniscata de Bernoulli para $t \in [0, \pi]$. . . 153

6.1 Estrutura Hierárquica de um problema. 159

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

6.2 Escala Fundamental de Saaty. 160
6.3 Árvore Hierárquica do Exemplo 1. 163
6.4 Console do R para obter matriz de julgamento. 164
6.5 Arquivo contendo as matrizes de julgamento em planilhas. 169
6.6 Tabela Final com os pesos globais, versão padrão. 170
6.7 Tabela Final com os pesos globais, versão GRAY. 171
6.8 Tabela Final com os pesos globais, versão WHITE. 171
6.9 Árvore Hierárquica do Exemplo 2. 172
6.10 Tabela Final com os pesos globais. 181

7.1 Multidisciplinaridade da IA na Engenharia Civil. 186
7.2 Tripé da Inteligência Artificial. 187
7.3 Gêmeos Digitais: saiba o que é. 189
7.4 Computador Quântico. 189
7.5 Metaverso. 190
7.6 Assentamento de alvenaria com robô. 191
7.7 Execução de amarração de armadurta de laje com robô. 191
7.8 Transporte de material pesado em obra por robô. 192
7.9 Serviço automatizado de terraplenagem. 192
7.10 Fiscalização de obra tradicional. 193
7.11 Óculos especial para medição de obras. 194
7.12 Verificação de instalações. 194
7.13 Robô para fiscalizar obras. 195
7.14 Fiscalização de obras com robô. 195
7.15 4 primeiras linhas do arquivo de dados. 196
7.16 Rede Neural com um neurônio. 197

8.1 Evolução da probabilidade de interesse do Problema do Aniversário conforme n cresce. 215
8.2 Janela R Commander. 217

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

8.3 Menu do R Commander para o Problema do Aniversário, adicionado pelo *plugin* `RcmdrPlugin.IPSUR` 217

9.1 HDiR from toponymic, patronymic and apelative surnames. 234
9.2 Left: councils for $\tau = 0.8$ (left connected components - Figure 9.1). Middle: councils for $\tau = 0.8$ (left and right connected components - Figure 9.1). Right: all councils in Galicia. 235
9.3 Spherical cluster results considering all data (left); Spherical cluster results filtering by population born before 1965 (center); Spherical cluster results filtering by population born before 1945 (right). . . 236
9.4 Spherical cluster results considering all data (and removing those ones below and above the 5% and 95% quantiles of the distribution of number of councils) (left); Spherical cluster results filtering by population born before 1965 (and removing those ones below and above the 5% and 95% quantiles of the distribution of number of councils) (center); Spherical cluster results filtering by population born before 1945 (and removing those ones below and above the 5% and 95% quantiles of the distribution of number of councils) (right). 236

10.1 Pontos de embarque e desembarque. 260

11.1 Fluxo do trabalho. 268
11.2 Gráfico após a tradução. 275
11.3 Alteração de cores e legendas. 277
11.4 Tradução do nome dos países. 278
11.5 Gráfico de calor com média. 279
11.6 Gráfico de calor traduzido e sem média. 281
11.7 Outra opção para tradução. 283

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Conteúdo

Lista de Figuras

1 INTRODUÇÃO - ORIGEM DO "SER" E SUA CONTRIBUIÇÃO INIGUALÁVEL 26

1.1 Origem do "SER" . . . 27
1.2 O Surgimento Visionário do Seminário de Estatística com R ("SER") 31
1.3 A Continuidade de uma Jornada Pioneira: O Legado do "SER" . . 34
1.4 Homenagem ao Professor Djalma Pessoa precursor da linguagem R no Brasil . . . 36
1.5 A Continuidade da Jornada do "SER": Uma Perspectiva de evolução 37
1.6 A Criação do Canal de Vídeos do "SER": Uma Janela Virtual para o Conhecimento Estatístico com R . . . 38
1.7 O "SER": Diversidade, Inovação e Colaboração na Vanguarda da Estatística com R . . . 39
1.8 Referências . . . 41

2 QUARTO BOOK: CONTANDO HISTÓRIAS COM QUARTO 43

2.1 Introdução . . . 44
2.2 Objetivo . . . 45
2.3 Aplicação . . . 45
2.3.1 Estrutura . . . 46
2.3.1.1 Markdown . . . 49

2.3.1.2 Código . . . 51
2.3.1.3 Quarto Pub . . . 53
2.4 Resultados e Discussão . . . 55
2.5 Conclusão . . . 60
2.6 Referências . . . 60

3 O QUARTO COMO FERRAMENTA DE PKM PARA PESQUISA CIENTÍFICA 61
3.1 Introdução . . . 61
3.2 Objetivo . . . 63
3.3 O QUARTO . . . 63
3.4 Como usar o QUARTO . . . 64
3.5 O PKM para Harold Jarche . . . 67
3.6 O QUARTO como ferramenta de PKM na pesquisa científica . . . 68
3.6.1 seek e sense . . . 69
3.7 Share . . . 70
3.8 Aplicação . . . 70
3.8.1 Criando um documento . . . 70
3.9 Conclusão . . . 75
3.10 Referências . . . 76

4 SERIES DE TIEMPO CON R 78
4.1 Introducción . . . 78
4.2 Fecha y Hora . . . 82
4.2.1 Objetos básicos de Fecha/Hora . . . 82
4.2.2 Otros paquetes . . . 85
4.3 Series de tiempo . . . 87
4.3.1 Importando directamente de Internet . . . 88
4.3.2 Paquetes para importar datos . . . 89
4.3.3 Clases para series de tiempo . . . 91

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

4.4 Modelización clásica de una serie de tiempo 97
4.4.1 Modelización ARMA 102
4.4.1.1 Modelo AR(p) 103
4.4.1.2 Modelo MA(q) 103
4.4.1.3 Modelo ARMA(p,q) 104
4.4.2 Identificación del modelo ARMA 105
4.4.3 Identificación y estimación del modelo 108
4.4.4 Predicción 115
4.4.5 Otros paquetes 117
4.5 Volatilidad condicional 119
4.5.1 GARCH Models 120
4.5.2 Modelos ARIMA–GARCH 122
4.5.2.1 Dinámica 123
4.5.2.2 Volatilidad 123
4.6 Conclusión 130
4.7 Referências 132

5 CURVAS E CORES EM R: MOVIMENTOS RÍGIDOS NO PLANO 134
5.1 Introdução 135
5.2 Objetivos 136
5.3 Aplicação 137
5.4 Conclusões 153
5.5 Referências 154

6 COMO USAR O PACOTE AHPWR PARA A TOMADA DE DECISÃO MULTICRITÉRIO 155
6.1 Introdução 155
6.2 Objetivos 156
6.3 Aplicação 157
6.3.1 A estrutura do método AHP 157

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

6.3.2 Como utilizar o pacote AHPWR . . . 162
6.3.2.1 Exemplo 1 . . . 162
6.3.2.2 Exemplo 2 . . . 171
6.4 Resultados e Discussão . . . 181
6.5 Considerações Finais . . . 182
6.6 Referências . . . 183

7 INTELIGÊNCIA ARTIFICIAL APLICADA À ENGENHARIA 185
7.1 Introdução . . . 186
7.2 Objetivo . . . 187
7.3 Aplicações . . . 190
7.4 Resultados e Discussões . . . 197
7.5 Conclusões . . . 198
7.6 Referências . . . 198

8 O PROBLEMA DO ANIVERSÁRIO, O PACOTE IPSUR E SEU PLUGIN PARA O R COMMANDER: Uma possibilidade para sala de aula 200
8.1 Introdução . . . 201
8.2 Objetivo . . . 202
8.3 Aplicação . . . 203
8.4 Resultados e Discussão . . . 210
8.5 Conclusão . . . 218
8.6 Referências . . . 221

9 CLASSIFICATION OF GALICIAN SURNAMES WITH WEB SCRAPING 223
9.1 INTRODUCTION . . . 224
9.1.1 Surnames, words and language . . . 225
9.2 Objective . . . 229

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

9.3 Application . . . 230
9.3.1 Directional Highest Density Regions . . . 230
9.3.2 Mixtures of von Mises-Fisher Distributions . . . 232
9.4 Results and Discussions . . . 233
9.4.1 Application HDiR to surnames data . . . 233
9.5 References . . . 237

10 USO DO DUCKDB COM R 239
10.1 Introdução . . . 240
10.2 Objetivo . . . 240
10.3 Aplicação . . . 241
10.3.1 Simplicidade . . . 241
10.3.2 Velocidade . . . 241
10.4 Recursos . . . 243
10.4.1 Tipos de dados . . . 243
10.4.2 Tipos de dados aninhados . . . 244
10.4.3 Leitura e escrita de arquivos externos . . . 246
10.4.3.1 **csv** e **parquet** . . . 246
10.4.3.2 **json** . . . 247
10.4.4 Funções . . . 248
10.5 Resultados . . . 249
10.5.1 Mineração de texto . . . 249
10.5.2 Dados de COVID-19 . . . 253
10.5.3 Lendo dados de um serviço de armazenamento na nuvem . . . 255
10.5.4 Análise de Dados Espaciais . . . 257
10.6 Um banco embarcado em uma API . . . 259
10.7 Conclusão . . . 263
10.8 Referências . . . 263

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

11 EDITANDO OS GRÁFICOS DO PACOTE LIKERT **265**

11.1 Introdução 266

11.1.1 Objetivo 267

11.1.2 Aplicação 267

11.2 Resultados e discussão 270

11.2.1 Outra solução 281

11.2.2 Tabelas 284

11.3 Conclusão: 284

11.4 Referências 285

Capítulo 1

INTRODUÇÃO - ORIGEM DO "SER" E SUA CONTRIBUIÇÃO INIGUALÁVEL

Autor: Orlando Celso Longo[1]
Programa de Pós-Graduação em Engenharia Civil Universidade Federal Fluminense
e-mail: orlandolongo@id.uff.br

Este capítulo oferece uma visão histórica do surgimento do Seminário de Estatística com R ("SER"), um evento pioneiro e inovador dedicado à linguagem R no Brasil, destacando o cenário da época de sua criação. Iniciado em 2016 pela Professora Dra. Luciane Ferreira Alcoforado, que introduziu o R na UFF e publicou um dos primeiros livros sobre a linguagem R no Brasil, o "SER" foi criado com o objetivo de promover o uso do R na comunidade acadêmica e profissional. Contou com a colaboração de professores de diversas áreas e instituições, enfrentando desafios e superando obstáculos para se estabelecer como um evento de referência. Com o passar dos anos, o "SER" cresceu, atraindo participantes de diferentes segmentos e áreas do conhecimento. Este crescimento não apenas reflete o impacto do "SER" na carreira de estudantes de graduação e mestrado, mas também destaca a importância do "SER" na formação de uma nova geração de profissionais e acadêmicos proficientes em R.

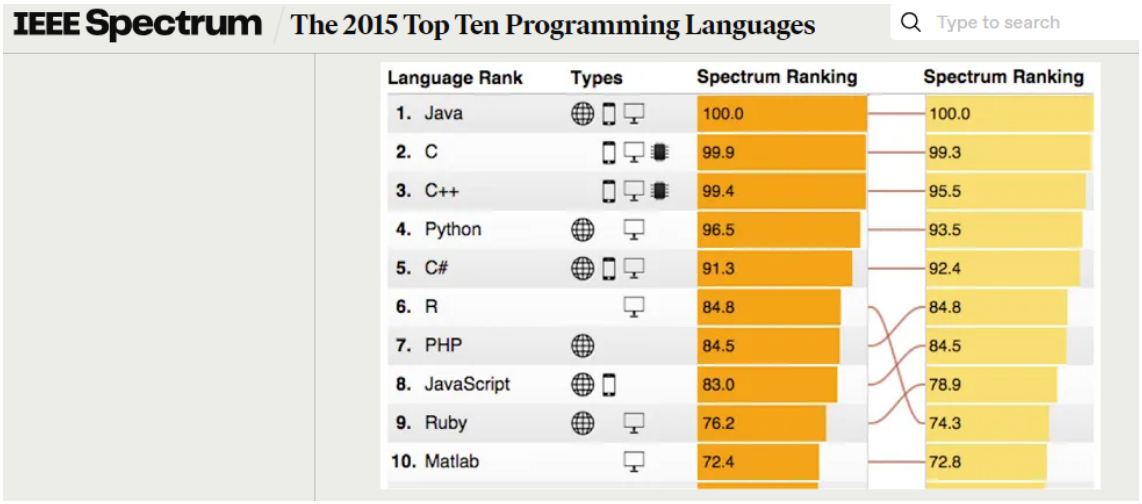
[1]Agradeço a todos os envolvidos na concepção, implementação e manutenção do projeto SER.

Palavras-Chave: Estatística com R; linguagem R; SER.

1.1 ORIGEM DO "SER"

No ano de 2015, a linguagem R conquistava seu lugar de destaque entre as principais ferramentas computacionais do mundo, conforme atestava o ranking publicado pela revista IEEE Spectrum (DIAKOPOULOS, 2015) Figura 1.1.

Figura 1.1: Ranking das linguagens em 2015.



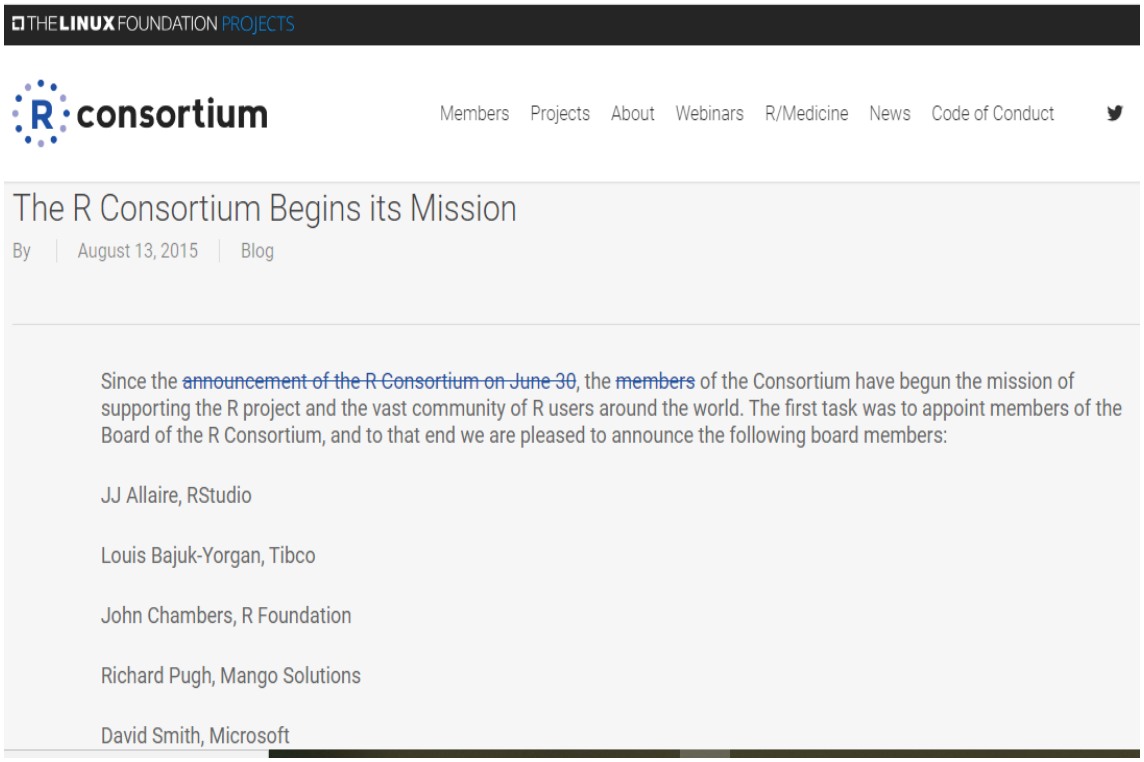
Fonte: (ALLAIRE et al., 2015).

Nesse mesmo período, ganhava forma o *R Consortium* (ALLAIRE et al., 2015), uma coalizão composta por empresas e associações, unidas no propósito de oferecer suporte à vasta comunidade global de usuários, que já ultrapassava a marca impressionante de dois milhões de adeptos em todos os cantos do planeta, Figura 1.2.

Foi nesse cenário efervescente, marcado por sua visão extraordinária, que a Professora Dra. Luciane Ferreira Alcoforado desempenhou um papel crucial na disseminação da linguagem R na Universidade Federal Fluminense (UFF). Sua notável trajetória teve início em 2010, quando introduziu o R no curso de Estatística da UFF, assumindo a posição pioneira de ministrar um curso dedicado a essa linguagem. O fruto desse trabalho incansável começou a se revelar em

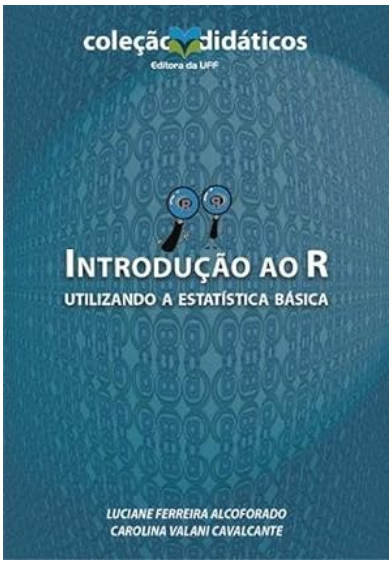
2014, quando, em parceria com a então estudante de Estatística, Carolina Valani Cavalcante, publicou um dos primeiros livros sobre a linguagem R no Brasil. Intitulado "Introdução ao R,", (ALCOFORADO; CAVALCANTE, 2014), Figura 1.3, o livro foi o resultado tangível de uma jornada iniciada com a introdução do R na disciplina de Métodos Computacionais do curso de Estatística. Esse marco histórico não apenas consolidou a posição da Professora como a introdutora do R na UFF, mas também enriqueceu o cenário acadêmico ao estabelecer as bases para o projeto "Estatística é com R!" apresentado no ano de 2015 na UFF Figura 1.4. O cerne desse projeto residia na divulgação das iniciativas em curso nas áreas de ensino, pesquisa e extensão, destacando-se pela integração exemplar com a comunidade acadêmica e profissional na UFF.

Figura 1.2: Início do R Consortium em 2015.



Fonte: (ALLAIRE et al., 2015).

Figura 1.3: Livro de Introdução ao R.



Fonte: O autor.

Figura 1.4: Projeto Estatística é com R!.

Fonte: O autor.

A contribuição notável da Professora Luciane não se limitou apenas à ex-

celência do projeto "Estatística é com R!", mas também à forma cuidadosa e engenhosa com que o conduziu. Ao promover uma sólida conexão entre a teoria estatística e sua aplicação prática, ela ofereceu uma plataforma inovadora para o desenvolvimento acadêmico e profissional, criando um ambiente enriquecedor para estudantes, pesquisadores e profissionais interessados em estatística e suas aplicações.

Mais do que um projeto, a iniciativa não poupou esforços para promover o aprimoramento constante dos conhecimentos estatísticos, evidenciando a dedicação ímpar à disseminação do saber. Através de eventos, cursos e outras atividades, o projeto não apenas enriqueceu o ambiente acadêmico da UFF, mas estabeleceu uma colaboração profícua com setores externos, fomentando uma troca de conhecimentos e experiências enriquecedora para todos os envolvidos.

Em um cenário em que os projetos de Extensão no Brasil, especialmente na área da Estatística, muitas vezes se restringiam a iniciativas tradicionais de assessoria estatística, a proposta inovadora da Professora no projeto "Estatística é com R!" surgiu como uma verdadeira transformação. Distanciando-se das abordagens convencionais, o projeto não se limitava à consultoria estatística comum, mas, de maneira audaciosa, introduzia um ecossistema digital. Esse cenário foi composto por um site dinâmico, com canal de vídeos no *YouTube* e na rede social *Facebook/Meta*, no qual alunos e professores tiveram a oportunidade de contribuir com postagens regulares, além de participarem da criação de vídeos curtos, todos com duração máxima de 3 minutos, coordenados pela professora, (ALCOFORADO; COSTA; FILGUEIRA; SIMÕES, 2017). Essa iniciativa, embora anterior à popularização do termo "*reels*", demonstra claramente a visão precursora da professora em incorporar recursos de mídia modernos para enriquecer a experiência acadêmica e atender às necessidades da comunidade. Dessa forma, o projeto não apenas evoluiu o panorama dos projetos de Extensão na área da Estatística, mas também ofereceu uma resposta inovadora aos desafios contemporâneos, destacando-se como um exemplo inspirador de como a integração entre ensino, pesquisa e exten-

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

são pode ser realizada de maneira eficaz e impactante na comunidade acadêmica e profissional.

1.2 O SURGIMENTO VISIONÁRIO DO SEMINÁRIO DE ESTATÍSTICA COM R ("SER")

No auge do desenvolvimento do projeto "Estatística é com R!", a Professora Luciane percebeu uma lacuna significativa no cenário acadêmico brasileiro: a ausência de um grande evento dedicado à linguagem R que estivesse à altura da sua imensa importância para o avanço das pesquisas em diversas áreas do conhecimento. Com a linguagem R sendo gratuita, mas os eventos especializados em R predominantemente ocorrendo no exterior com custos proibitivos para participantes brasileiros, especialmente na área da estatística, a necessidade de um evento acessível tornou-se evidente.

Foi nesse contexto que nasceu a iniciativa de organizar o primeiro Seminário de Estatística com R no Brasil, carinhosamente chamado de "SER". A proposta era clara: proporcionar um evento de qualidade, acessível a todos os interessados na linguagem R, estimulando a atualização e a troca de conhecimentos na comunidade de usuários.

Desde o início, o Programa de Pós-Graduação em Engenharia Civil desempenhou um papel fundamental no apoio ao projeto "Estatística é com R!", ao disponibilizar um laboratório dedicado para o desenvolvimento das atividades do projeto, bem como para reuniões entre professores, alunos de graduação e de pós-graduação. O programa criou um ambiente colaborativo propício ao crescimento tanto do projeto "Estatística é com R!", quanto do "SER". Essa parceria e suporte contínuos foram essenciais para o sucesso e a expansão dessas iniciativas, destacando o comprometimento do programa com o avanço do conhecimento estatístico na comunidade acadêmica.

Para concretizar esse ambicioso empreendimento, a Professora convocou um conjunto multidisciplinar de professores da UFF, dentre os quais destaco minha

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

contribuição essencial. Como apoiador ativo, não apenas respaldei a inserção da linguagem R no Programa de Pós-Graduação em Engenharia Civil, mas também desempenhei um papel significativo como incentivador e divulgador do trabalho da Professora Luciane. Essa colaboração sinérgica tornou-se um pilar fundamental para o desenvolvimento do projeto "SER", evidenciando a importância da união e apoio entre os membros do grupo organizador. A colaboração com o IMPA, impulsionada pelo empenho do Professor Ariel Levy e o respaldo do Professor Jorge Zubelli, desempenhou um papel fundamental no fortalecimento do projeto, conferindo-lhe maior relevância no cenário acadêmico. Vale ressaltar que a visão audaciosa do Prof. Ariel Levy, aliada ao seu contato com o Prof. Jorge Zubelli do IMPA, e a participação dos professores Manuel Febrero e Wenceslao Manteiga da Universidade de Santiago de Compostela, foram cruciais para elevar o status do evento ao nível internacional, dando origem ao Seminário Internacional de Estatística com R. A Professora Maysa Magalhães, à época coordenadora da ENCE, percebendo o potencial inovador do evento, prontamente uniu-se ao grupo de organizadores, que contava com a liderança da Professora Luciane e a contribuição valiosa de Ariel, José Rodrigo, Steven, e, claro, minha própria participação ativa. Essa diversidade de perspectivas e expertise contribuiu de maneira decisiva para o sucesso do "SER", transformando-o em um espaço de atualização, troca de conhecimentos e inovação, consolidando seu papel essencial no cenário acadêmico brasileiro.

Assim, em maio de 2016 o "SER" emergiu como uma resposta visionária às demandas crescentes da comunidade de usuários da linguagem R no Brasil, inaugurando um espaço de atualização, troca de conhecimentos e inovação, com a missão de impulsionar o desenvolvimento de pesquisas em diversas áreas do conhecimento, (ALCOFORADO; LEVY; LONGO, 2016).

O sucesso do evento reverberou internacionalmente, alcançando reconhecimento pela *R Foundation*, que publicou um artigo enaltecendo o pioneirismo do evento "SER" na América Latina, (LEVY; MAYER et al., 2018), (LEVY; AL-

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

COFORADO; LONGO, 2018). Estes artigos circularam intensamente pelas redes sociais em todo o mundo, consolidando a reputação do evento.

A trajetória do Seminário de Estatística com R (“SER”) foi enriquecida significativamente pela valiosa contribuição de estudantes do grupo da Professora , cujo engajamento foi fundamental para o início e sucesso contínuo do evento. Destacam-se alguns desses alunos que desempenharam papéis essenciais:

1. **Camila Simões, Leonardo Filgueira, Jonatha Azevedo:** Alunos do curso de Estatística em 2015, Camila, Leonardo e Jonatha não apenas participaram ativamente do projeto “Estatística é com R!”, mas também foram peças-chave na organização inicial do “SER”. Sua dedicação e envolvimento contribuíram significativamente para estabelecer as bases do evento.

2. **Noelle Camelo:** Noelle Camelo, aluna do Mestrado em Engenharia Civil, desempenhou um papel crucial na condução do Cerimonial do “SER”, oferecendo sua considerável experiência na organização de eventos na área do turismo. Sua contribuição começou nos estágios iniciais do “SER” e permanece vital até os dias atuais.

3. **Marlon Magalhães:** Marlon Magalhães iniciou sua participação no projeto enquanto ainda estudante do ensino médio e posteriormente ingressou no curso de Desenho Industrial da UFF. Com habilidades notáveis em design, Marlon desempenhou um papel crucial no desenvolvimento de arte para o evento, continuando a ser uma peça essencial no grupo.

4. **Vanessa Manhães:** Vanessa Manhães, aluna do Mestrado em Engenharia Civil, destacou-se pela contribuição para as metodologias de realização do evento em sua forma remota. Além disso, nos anos de 2021 e 2022, impulsionou as redes sociais do Instagram e do canal do *YouTube* do “SER”, ampliando ainda mais a presença online do evento.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

5. **Marcus Ramalho:** Marcus Ramalho, aluno do Mestrado em Administração da UFF, desempenhou um papel crucial na continuidade da transmissão remota do "SER" em 2023. Sua contribuição foi fundamental para a adaptação bem-sucedida do evento a novos formatos.

6. **Raquel dos Santos:** Uma peça essencial na equipe do "SER", Raquel dos Santos, estudante do curso de Turismo, desempenhou um papel crucial na parte de credenciamento e organização da mesa de abertura. Além de liderar uma equipe de alunos do curso de Turismo, sua dedicação e habilidades foram fundamentais para o sucesso e acolhimento eficaz dos participantes do evento. O "SER" beneficiou-se significativamente de sua liderança e comprometimento.

Além de suas importantes responsabilidades na organização do evento, é digno de nota que muitos desses alunos evoluíram para se tornar palestrantes em diversas edições do "SER", destacando-se Camila, Leonardo, Jonatha, Vanessa e Marcus, que não apenas ajudaram a construir o evento, mas também compartilharam seus conhecimentos em diferentes edições, enriquecendo ainda mais o "SER" como um fórum de atualização, troca de conhecimentos e inovação na comunidade acadêmica.

1.3 A CONTINUIDADE DE UMA JORNADA PIONEIRA: O LEGADO DO "SER"

A trajetória do Seminário de Estatística com R no Brasil, transcendeu fronteiras a cada edição, consolidando-se como um marco indelével na comunidade acadêmica e científica. Após uma estreia bem-sucedida em 2016, a segunda edição, em 2017, elevou o evento a patamares ainda mais notáveis, (ALCOFORADO, 2019b), (ALCOFORADO, 2017).

Nesse contexto vibrante, a rica programação envolveu palestras, TEDs e oficinas, com a participação de conferencistas internacionais renomados. Destacaram-

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

se também 27 trabalhos científicos na sessão de pôster e 21 artigos na sessão de comunicação oral, reconhecidos com premiações. O êxito transbordou para além das fronteiras, culminando na publicação de dois volumes de anais, registrados com ISSN e ISBN, pelo IBICT.

A terceira edição, em 2018, consolidou o evento como um espaço inclusivo, acolhendo alunos, professores e profissionais de diversas áreas. A programação diversificada contemplou minicursos, conferências internacionais e homenagens, com destaque para a reverência ao saudoso professor Djalma Pessoa do ENCE/IBGE, pioneiro da linguagem R no Brasil, (ALCOFORADO, 2018)

Em 2019, a quarta edição expandiu ainda mais os horizontes, atraindo participantes de diversas esferas acadêmicas e profissionais. Com uma variedade impressionante de 13 minicursos e um treinamento conduzido pelo renomado professor Dean Attali, a experiência dos participantes foi enriquecida. O evento, além de palestras internacionais e premiações, inovou com a incorporação do encontro R-ladies e o evento satélite EDUCA-SER, ampliando a disseminação do R desde a educação básica até o ensino superior.

A contribuição crucial de apoiadores como UFF, CAPES, ENCE/IBGE, IMPA, SBMAC, UNIRIO e a Prefeitura de Niterói foi fundamental para o sucesso e a abrangência do evento em todas as suas edições. Além disso, minha participação e apoio junto ao grupo de líderes do evento foi essencial para sua continuidade, pois, além de ser o coordenador do programa de pós-graduação em Engenharia Civil, reunia uma vasta experiência na organização de eventos na área de engenharia civil. Assim, conduzi o grupo à tomar providências essenciais, como o registro do ISSN e homenagem ao Prof. Djalma no ano de 2018.

Dessa forma, o Seminário de Estatística com R no Brasil não apenas consolidou-se como um fórum acadêmico de excelência, mas também como um catalisador da disseminação do conhecimento estatístico em âmbito nacional e internacional.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

1.4 HOMENAGEM AO PROFESSOR DJALMA PESSOA PRECURSOR DA LINGUAGEM R NO BRASIL

Em um tributo emocionante e significativo, o Seminário de Estatística com R no Brasil prestou uma homenagem ímpar ao saudoso Professor Djalma Pessoa. O Prof. Djalma, precursor da linguagem R no Brasil na década de 1990, foi homenageado durante a terceira edição do "SER" em 2018, em reconhecimento à sua contribuição pioneira, Figura 1.5.

Figura 1.5: Homenagem ao prof. Djalma Pessoa.

Fonte: Acervo do III SER.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

A atmosfera calorosa e respeitosa do evento reverberou em reconhecimento à intensa produção e ao papel precursor do Prof. Djalma no cenário estatístico brasileiro. Contudo, com pesar, registramos o falecimento do estimado professor em 01 de agosto de 2020, deixando um vazio na comunidade acadêmica e no coração daqueles que tiveram a honra de conhecê-lo.

A experiência e conhecimento que pude oferecer desempenharam um papel crucial no sucesso do evento. A homenagem, inicialmente prestada em vida ao Prof. Djalma, permanece como um reconhecimento duradouro à sua contribuição singular e presença marcante no cenário acadêmico e científico.

1.5 A CONTINUIDADE DA JORNADA DO "SER": UMA PERSPECTIVA DE EVOLUÇÃO

Ao longo dos anos, minha participação ativa na organização do Seminário de Estatística com R ("SER") continuou a ser um elemento crucial, especialmente ao garantir o apoio da CAPES nas edições de 2016, 2018 e 2019. Minha experiência e conhecimento desempenharam um papel fundamental para o sucesso contínuo do evento.

Em 2020, estávamos ansiosos para a quinta edição quando a pandemia da Covid-19 nos forçou a cancelar o evento. No ano seguinte, em 2021, enfrentamos os desafios da pandemia ao realizar a quinta edição de forma remota, conectando aproximadamente 1.500 participantes de diversos setores. A rica programação incluiu 22 palestras, quatro delas internacionais, e dois minicursos online. Os apoiadores fundamentais, como UFF, ENCE/IBGE, UNIRIO, UERJ, AFA e USC, desempenharam um papel vital no sucesso do evento, (ALCOFORADO; LONGO; LEVY, 2021).

Em 2022, a sexta edição também ocorreu de maneira remota, reunindo cerca de 1.300 participantes de diferentes segmentos da área de estatística. A programação manteve seu padrão de excelência, com 22 palestras, incluindo cinco internacionais, e dois minicursos. Destacou-se a segunda edição do evento satélite

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

EDUCA-SER na UNIRIO, focado em difundir o uso do R no ensino de estatística. Os apoiadores, incluindo UFF, ENCE/IBGE, SBMAC, UNIRIO, UERJ, AFA, IME e USC, foram essenciais para o êxito do evento.

Já em 2023, a sétima edição do "SER" foi marcada pela participação expressiva de diversos segmentos e usuários do R, desde alunos de graduação até professores/pesquisadores e profissionais do mercado. A programação abrangente contou com 16 palestras, incluindo duas internacionais. Paralelamente, realizou-se a terceira edição do evento satélite EDUCA-SER na UNIRIO, mantendo o mesmo objetivo e formato das edições anteriores. Os apoiadores, como UFF, UNIRIO, AFA e USC, contribuíram significativamente para o sucesso do evento.

1.6 A CRIAÇÃO DO CANAL DE VÍDEOS DO "SER": UMA JANELA VIRTUAL PARA O CONHECIMENTO ESTATÍSTICO COM R

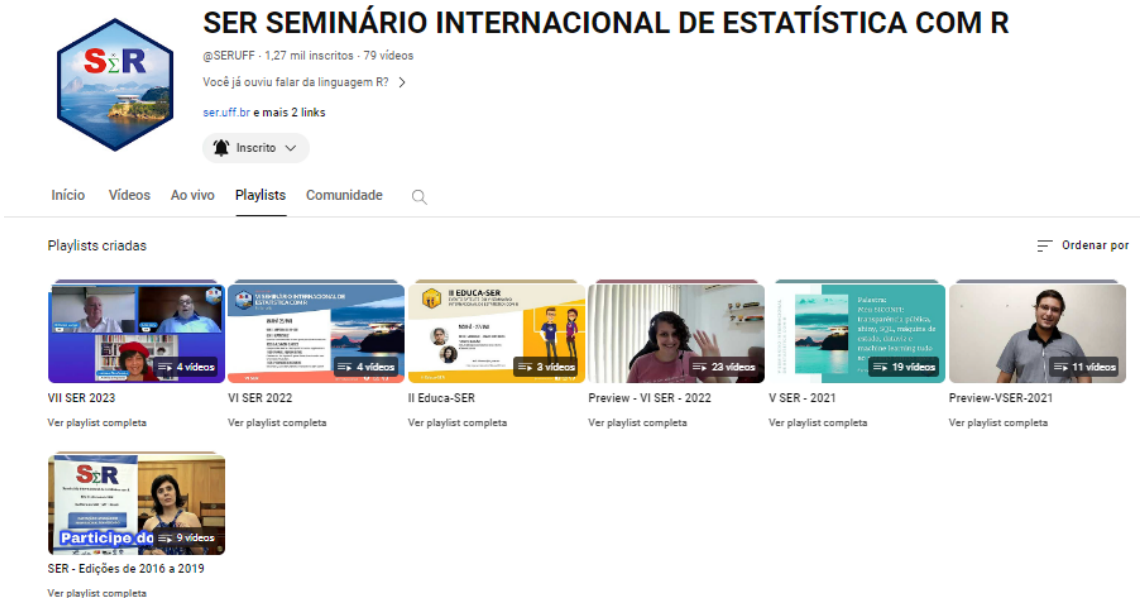
Em resposta à nova dinâmica do Seminário de Estatística com R ("SER") como evento remoto, uma iniciativa significativa surgiu: a criação do canal de vídeos no *YouTube* do "SER" [https://www.youtube.com/@SERUFF]Alcoforado et al. (2015), Figura 1.6. Este canal foi concebido como uma ferramenta inovadora para registrar e compartilhar as valiosas palestras proferidas durante o evento, proporcionando uma rica fonte de conteúdo para os entusiastas da linguagem R.

Ao longo do tempo, o canal evoluiu para se tornar um repositório digital robusto, destacando-se como a principal plataforma para acessar a produção das palestras desde a quinta edição do Seminário Internacional de Estatística com R. Os responsáveis pela produção e manutenção deste canal são a Profa. Luciane, o Prof. Ariel e eu Prof. Orlando.

O conteúdo do canal oferece uma variedade de perspectivas e aplicações envolvendo a linguagem computacional R, sendo que a responsabilidade sobre cada palestra e seu respectivo conteúdo recai sobre o palestrante. Essa abordagem multifacetada reflete a diversidade e a amplitude das discussões promovidas no

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 1.6: Canal de vídeos do "SER".



Fonte: (ALCOFORADO; COSTA; FILGUEIRA; SIMÕES; COLABORADORES, 2015).

"SER", contribuindo para a disseminação abrangente de conhecimento estatístico.

Em última análise, o Canal "SER" no *YouTube* representa uma janela virtual para o mundo do R, proporcionando a estudantes, pesquisadores e profissionais uma oportunidade única de explorar as aplicações práticas e avançadas da linguagem, tudo isso no conforto de suas telas.

1.7 O "SER": DIVERSIDADE, INOVAÇÃO E COLABORAÇÃO NA VANGUARDA DA ESTATÍSTICA COM R

Desde sua primeira edição, o "SER" tem se destacado pela diversidade de sua comissão científica e organizadora, composta não apenas por estatísticos, mas também por matemáticos, engenheiros, administradores e outros profissionais. Essa composição heterogênea resulta em uma programação abrangente, atendendo aos interesses de diversas áreas do conhecimento. Destaco especialmente a oferta de minicursos, tornando a linguagem R acessível a pesquisadores e estudantes que buscam familiaridade, ampliando, assim, o acesso ao conhecimento de maneira

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

conclusiva.

A evolução dinâmica da linguagem R exige encontros periódicos, não superiores a um ano, para garantir o acesso às atualizações de novos pacotes e funções implementadas e aprimoradas diariamente. Essa busca constante por inovação no R não apenas fortalece a capacidade dos profissionais da área de estatística e ciência de dados, mas também promove a colaboração interdisciplinar e o uso de ferramentas tecnológicas avançadas. A inovação no R impulsiona a pesquisa, tornando-a mais eficiente e relevante, ao mesmo tempo em que amplia o alcance da estatística e da ciência de dados em diferentes campos, criando oportunidades para resolver problemas complexos e realizar descobertas valiosas.

Na terceira edição do evento "SER" em 2018, a presença influente das R-Ladies dos Estados Unidos, Julia Silge e Jesse Maegan, trouxe um novo capítulo à história do "SER". Nessa ocasião, ambas incentivaram a organização de um capítulo R-Ladies Niterói, que foi efetivamente estabelecido em julho pela Professora Luciane, a mestranda Noelle Camelo e a estudante de Estatística Julia Ferreira, (GLOBAL et al., 2018). Esse capítulo não apenas enriqueceu a diversidade de gênero na comunidade R, mas também fortaleceu a missão do "SER" de ser um espaço inclusivo e inovador.

O capítulo R-Ladies Niterói foi incluído no âmbito da UFF como um projeto de extensão no ano de 2019, (ALCOFORADO, 2019a). Muitos dos encontros foram realizados nas dependências do Programa de Pós-Graduação em Engenharia Civil, demonstrando o quanto o programa incentivou a disseminação da linguagem R, apoiando a inclusão feminina. Essa integração reflete o compromisso contínuo do programa em promover não apenas a excelência acadêmica, mas também a diversidade e a igualdade de oportunidades.

O convite feito à Professora Luciane para integrar o corpo docente do Programa de Pós-Graduação em Engenharia Civil, em 2010, revelou-se uma decisão extraordinariamente valiosa. Na época, recém-doutora e repleta de dinamismo, ela trouxe uma abordagem inovadora ao ensino da disciplina de Estatística, mui-

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

tas vezes considerada intimidante pelos alunos. Desde então, a Professora não apenas ministrou a disciplina com maestria, introduzindo o uso da linguagem R, algo pouco comum nos cursos de Pós-Graduação da UFF na época, mas também desempenhou um papel crucial na incorporação de outros docentes igualmente inovadores, como o Professor Steven e, mais recentemente, o Professor João Paulo. Essa equipe de professores dinâmicos e comprometidos criou um ambiente propício para o desenvolvimento de mestrandos, que, orientados pelos docentes envolvidos com o programa e a disciplina de Estatística, puderam explorar e apresentar trabalhos de análise de dados usando o R como ferramenta, culminando no enriquecimento do Seminário de Estatística com R. O impacto dessa trajetória colaborativa ressoa não apenas nas salas de aula, mas também nas contribuições significativas para a disseminação do conhecimento estatístico e no fortalecimento do ecossistema acadêmico da UFF.

1.8 REFERÊNCIAS

ALCOFORADO, Luciane Ferreira. **R-ladies UFF**. [S.l.]: Sigproj, 2019. Disponível em: http://sigproj.ufrj.br/apoiados.php?projeto_id=318735.

_____. SER abre o primeiro dia com oferta variada de minicursos. **Boletim Informativo do SER**, 2018.

_____. SER abre o primeiro dia com R-pesquisadores da Espanha, Portugal e do Brasil. **Boletim Informativo do SER**, 2017.

_____. **SER promove a integração entre Universidade e Mercado e supera expectativas**. [S.l.]: Boletim Informativo do SER, 2019. Disponível em: http://sigproj.ufrj.br/apoiados.php?projeto_id=318735.

ALCOFORADO, Luciane Ferreira; CAVALCANTE, Carolina Valani. **Introdução ao R Utilizando a Estatística Básica**. [S.l.]: Eduff - Niterói - RJ, 2014. ISBN 978-8522807659.

ALCOFORADO, Luciane Ferreira; COSTA, Jonatha Azevedo; FILGUEIRA, Leonardo; SIMÕES, Camila S.D. Programa Estatística é com R, uma janela aberta para o aprendizado da Estatística e do software R. **Revista Compartilhar**, 2017.

ALCOFORADO, Luciane Ferreira; COSTA, Jonatha Azevedo; FILGUEIRA, Leonardo; SIMÕES, Camila S.D.; COLABORADORES. **Estatística é com R!** [S.l.: s.n.], 2015. Disponível em: http://www.estatisticacomr.uff.br. Accessed: 2023-14-11.

ALCOFORADO, Luciane Ferreira; LEVY, Ariel; LONGO, Orlando Celso. **SER: Seminário Internacional de Estatística com R**. [S.l.]: UFF, 2016. Disponível em: https://www.ser.uff.br.

ALCOFORADO, Luciane Ferreira; LONGO, Orlando Celso; LEVY, Ariel. SER o maior evento de R da América Latina totalmente online. **Boletim Informativo do SER**, 2021.

ALLAIRE, JJ et al. **R consortium begins its mission**. [S.l.: s.n.], 2015. Disponível em: https://www.r-consortium.org/blog/2015/08/13/the-r-consortium-begins-its-mission. Accessed: 2023-14-11.

DIAKOPOULOS, Nick. the-2015-top-ten-programming-languages. **IEEE Spectrum**, 2015.

GLOBAL, R Ladies et al. **R Ladies Niterói**. [S.l.]: R Ladies Global, 2018. Disponível em: https://www.meetup.com/pt-BR/rladies-niteroi/.

LEVY, Ariel; ALCOFORADO, Luciane Ferreira; LONGO, Orlando Celso. Conference Report: SER III. **R Journal**, 2018.

LEVY, Ariel; MAYER, Fernando et al. **R in Latin America**. [S.l.]: Forwards, 2018. Disponível em: https://forwards.github.io/blog/2018/02/05/r-in-latin-america.

Capítulo 2

QUARTO BOOK: CONTANDO HISTÓRIAS COM QUARTO

Autor: Ariane Hayana Thomé de Farias
Assessora Estatística no Tribunal de Justiça do Estado de Roraima (TJRR)
e-mail: ariane.hayana@gmail.com

Com o avanço das tecnologias, automatizar processos repetitivos tornou-se de fundamental importância para otimizar o tempo e gerar *insights* para uma melhor tomada de decisão. Um ponto que merece destaque: em meio a tantas ferramentas, como comunicar seus resultados? Traduzir dados em informação é uma tarefa que, caso não seja bem executada, pode não expressar o que se deseja. Dentre as possibilidades atuais de desenvolvimento de produtos automatizados, temos o Quarto, a nova geração de `R` *Markdown*. Assim, com objetivo de explorar as funcionalidades e a criação de um livro desenvolvido em *Quarto*, este capítulo abordará conceitos, definições e orientações para o desenvolvimento de um livro desde a instalação do Quarto até a etapa de publicação no Quarto Pub.
Palavras-Chave: Quarto; Livro; R.

2.1 INTRODUÇÃO

Traduzir dados complexos em um conteúdo de fácil entendimento não é uma tarefa trivial considerando que, embora existam diversas ferramentas para a apresentação de resultados, de nada adianta se a abordagem utilizada não for adequada para comunicar com os tomadores de decisão. A imensa quantidade de dados disponível nas mais diversas fontes é uma realidade desafiadora para extrair *insights* e gerar resultados e neste sentido, uma das possibilidades para a contextualização de narrativas embasadas em dados é através da utilização do Quarto. Com ele, é possível criar diversos produtos, tais como relatórios automatizados em `html`, `pdf`, `epub`, assim como possibilita a criação de *blogs* que podem ser atualizados e compartilhados na *web* de forma rápida e em versões multilíngues, com conteúdo textual e em códigos nas linguagens `R`, `Python` e `Julia`, por exemplo. Neste contexto, este capítulo visa explorar as funcionalidades do *Quarto Book* desde a instalação, com orientações de quais são os requesitos necessários para a utilização do Quarto localmente e na nuvem. Também será apresentado a estrutura básica de um documento Quarto e, por fim, o passo a passo dos procedimentos necessários para publicação do *Quarto Book*.

No que se refere à organização deste capítulo, este trabalho encontra-se organizado em seções dispostas da seguinte forma: além desta seção introdutória, o objetivo é apresentado na Seção 2.2. Também são apresentadas as aplicações do Quarto, contextualizando definições e pilares da utilização do Quarto na Seção 2.3. Nesta seção também são abordados os passos necessários para a publicação do *Quarto Book* no Quarto Pub. Na Seção 2.4 temos os resultados de um modelo exemplificativo do livro personalizado. Por fim, a Seção 2.5 são apresentadas as considerações finais e posteriormente, as referências.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

2.2 OBJETIVO

O objetivo principal deste capítulo é apresentar as etapas de criação de um livro desenvolvido com Quarto, trazendo orientações preliminares sobre conceitos básicos para a criação deste, com a exemplificação de um livro e sua publicação no Quarto Pub.

2.3 APLICAÇÃO

Quarto é uma versão multilíngue da próxima geração do `R` Markdown e com ele é possível criar relatórios automatizados tanto com a linguagem R em Python e Julia. Além disso, pode-se criar diversos tipos de documentos, apresentações, livros, *blogs*, entre outros recursos. O objetivo neste capítulo consiste em abordar as etapas de desenvolvimento de um livro com Quarto, remetendo aos procedimentos necessários para a criação do *Quarto book* no RStudio. Para tanto, a primeira etapa metodológica é realizar as instalações necessárias para utilizar o Quarto no RStudio. Ressalta-se que Quarto é um *software* independente de `R` e pode ser utilizado em diversas ferramentas (RStudio, VS Code, Jupyter Lab, por exemplo). Assim, os pré-requisitos são:

- Baixar e instalar a versão mais recente do Quarto disponível neste link;
- Baixar e instalar a versão mais recente do `R` clicando aqui. Caso já tenha, verifique se o seu `R` está atualizado;
- Baixar e instalar a versão mais recente do RStudio;
- Instalar os seguintes pacotes:

```
#| label: code_01
#| echo: true
#| eval: false
```

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

```
pacotes <- c("tidyverse", "quarto", "rmarkdown", "dados
  ", "reactable")
install.packages(pacotes)
```

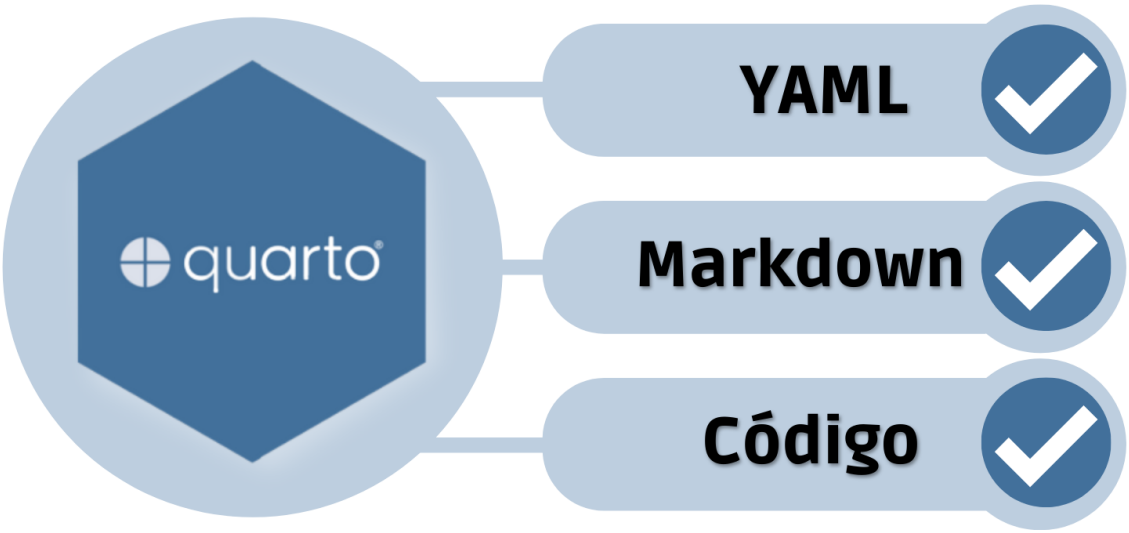
Vale ressaltar que cada sistema operacional (SO) possui especificidades de instalação, portanto, é importante verificar as configurações do SO durante as instalações.

Orientações sobre a utilização do RStudio com Quarto estão disponíveis na página do Quarto.org. Outra opção é utilizar o *Posit Cloud*, que possibilita a utilização da IDE do RStudio na nuvem através do site da Posit. O passo a passo está descrito também em Farias (2023a).

2.3.1 Estrutura

Para o desenvolvimento do *Quarto Book*, é importante compreender a estrutura que o compõe, a qual pode ser resumida em três partes principais, conforme ilustrado na Figura 2.1:

Figura 2.1: Estrutura.



Fonte: A autora.

- YAML (*Yet Another Markup Language*): local onde são inseridos os metadados sobre o documento (formatação, data, título, autor, entre outros). Nele, é possível configurar a aparência do documento, bem como especificar quais serão os formatos de saída (HTML, pdf, Word, etc);
- Texto: usa *Markdown* como sua sintaxe de documento principal;
- Códigos: versão multilíngue, portanto, nos *chunks* (blocos de códigos) podemos inserir códigos em `R`, Python, Julia e outros.

Partindo do entendimento da estrutura comum entre os documentos desenvolvidos em Quarto, parte-se para a criação do *Quarto book*. As configurações do livro estão contidas em um arquivo separado que estará localizado no **diretório do projeto**, que é denominado `_quarto.yml`.

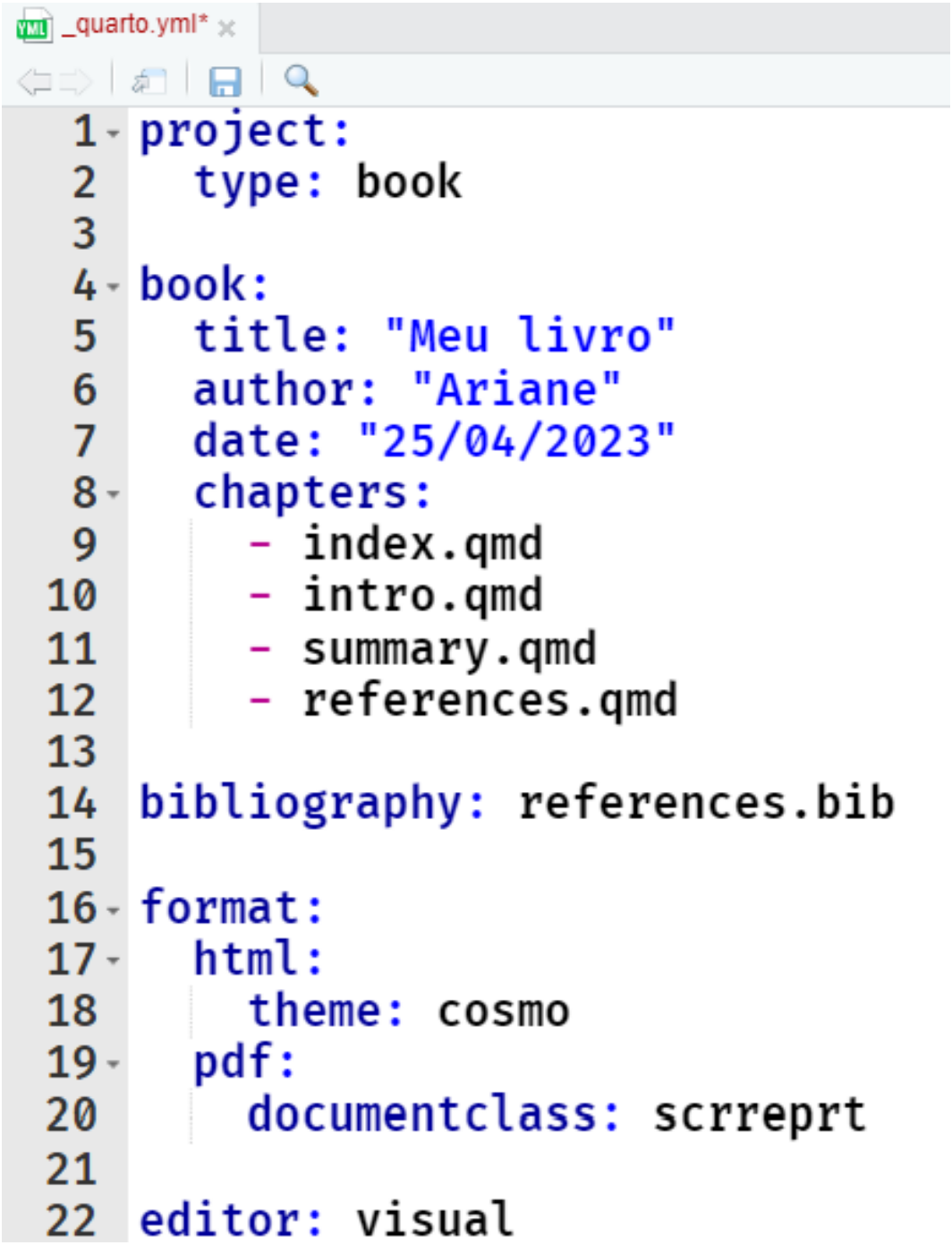
Este arquivo contém a configuração inicial do livro e é onde são inseridas configurações sobre o documento (formatação, data, título, autor, entre outros). Para exemplificar, a Figura 2.2 mostra algumas configurações do `_quarto.yml`. Alguns pontos cabem destaque nas configurações do *Quarto book*:

(1) a estrutura do YAML é baseada em indentação, portanto, é importante que as configurações estejam corretamente indentadas, caso contrário, aparecerá alguma mensagem de erro durante a renderização do livro;

(2) o Quarto *book* pode ser configurado com diferentes formatos de saída, entretanto, caso o arquivo contenha alguma funcionalidade específica de um dos formatos, possivelmente não irá funcionar adequadamente em outro formato (como, por exemplo, a utilização de gráficos interativos no formato HTML, que não funcionarão em PDF).

Cabe destacar que o Quarto é baseado no Pandoc, assim, na fase de composição do conteúdo no livro, a sintaxe básica é **markdown**. Conforme apresentado na documentação do Quarto (2023b),

Figura 2.2: Tela do `_quarto.yml`.



Fonte: A autora.

"o markdown do Pandoc é uma versão estendida e ligeiramente revisada da sintaxe Markdown de Gruber (2023)".

Portanto, para pessoas que devolvem documentos com Markdown/R Mark-

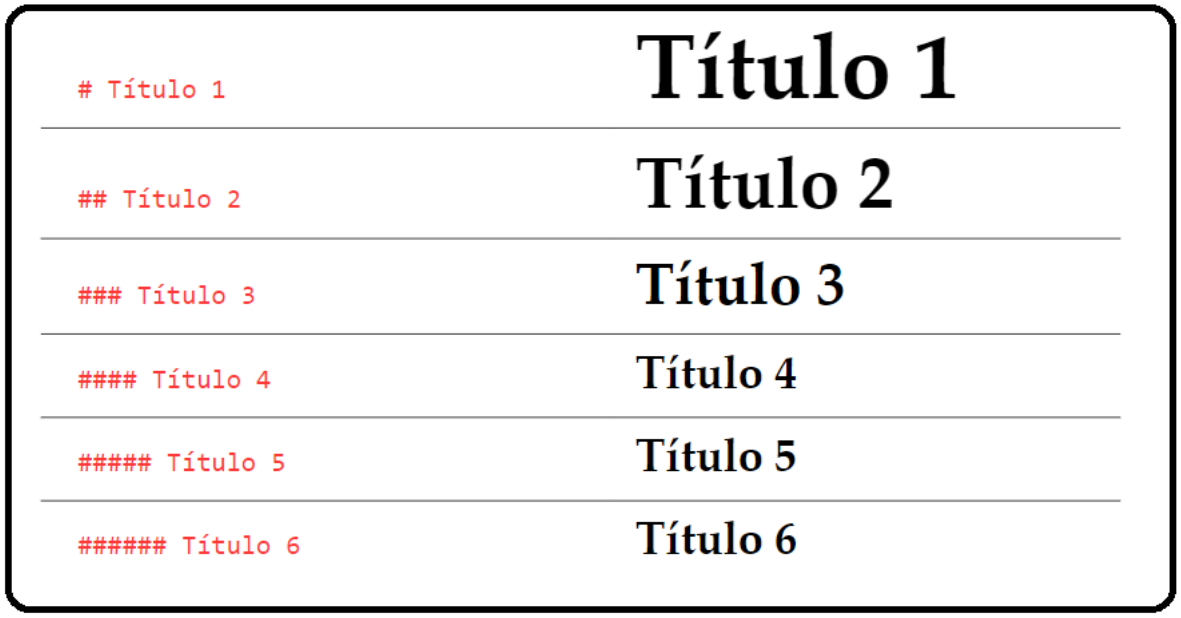
APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

down, a sintaxe será a mesma em algumas estruturas. Na Seção 2.3.1.1 será abordado algumas sintaxes básicas para personalização do texto.

2.3.1.1 Markdown

Nesta seção serão apresentadas algumas das possíveis configurações para personalização do texto. A primeira consiste em criar títulos/subtítulos (ou seções e subseções), que são estruturadas conforme a quantidade de caractere #, seguido do texto desejado, conforme apresentado na Figura 2.3. Ressalta-se que pode criar até seis níveis distintos com as #. Observe que, à medida que aumenta a quantidade de #, menor será o nível do título/subtítulos.

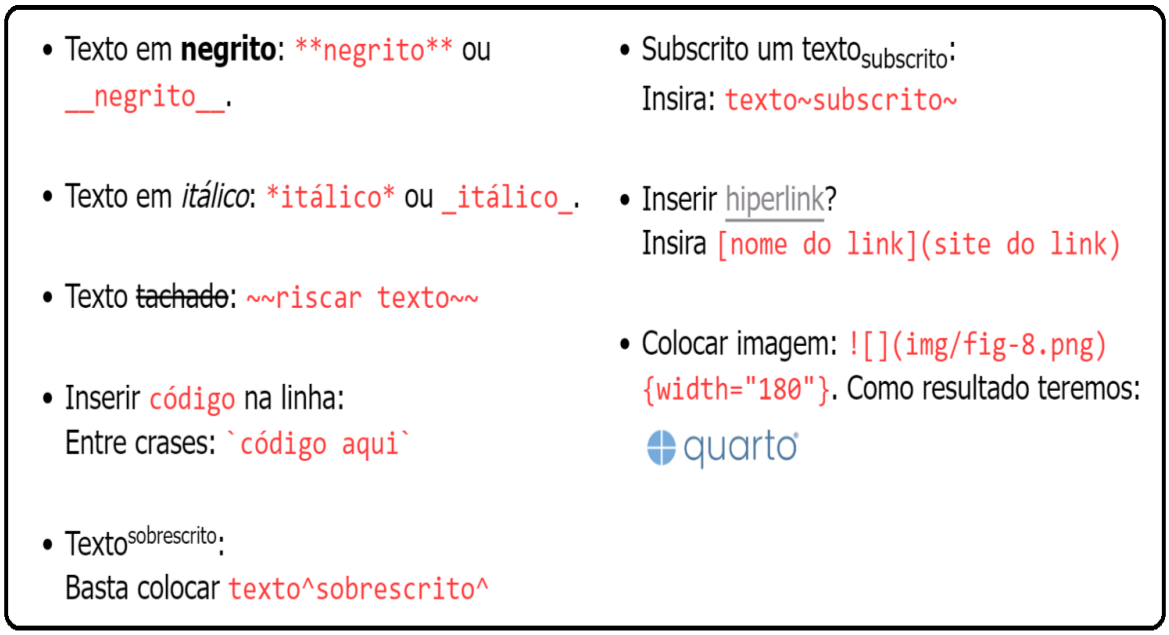
Figura 2.3: Títulos/Subtítulos ou Seções/Subseções.



Fonte: A autora.

Outras configurações básicas para o texto estão disponíveis no esquemático da Figura 2.4. Para a personalização dos textos, existem diversas possibilidades, com detalhes disponíveis na documentação do (QUARTO, 2023b).

Figura 2.4: Estruturas para o texto.



Fonte: A autora.

Figura 2.5: Nomenclaturas e referências.

Prefixo	Descrição
#def-	Definição
#exm-	Exemplo
#exr-	Exercício
#fig-	Figura
#tbl-	Tabela
#eq-	Equações

Fonte: (FARIAS, 2023a).

Além disso, outra praticidade na construção de um livro com Quarto é a inserção de referências cruzadas, que facilitam a navegação dos leitores em seu documento, possibilitando a criação de referências numeradas automaticamente para `exemplos`, `exercícios`, `definições`, `corolários` e outros. A Figura 2.5 apresenta algumas estruturas, que podem ser criadas com o caractere `#` no prefixo.

Assim, por exemplo, caso queira queira criar um exercício, basta inserir o prefixo `#exr-`, conforme o código:

```
::: {#exr-exercicio01}
Enunciado do meu exercício aqui. Enunciado do meu exerc
  ício aqui.
:::
::: callout-note
## Solução
Solução do meu exercício aqui. Solução do meu exercício
   aqui.
:::
```

Como resultado do código acima no Quarto *book* será:

Exercício 3.1. Enunciado do meu exercício aqui. Enunciado do meu exercício aqui.

> **Solução**
>
> Solução do meu exercício aqui. Solução do meu exercício aqui.

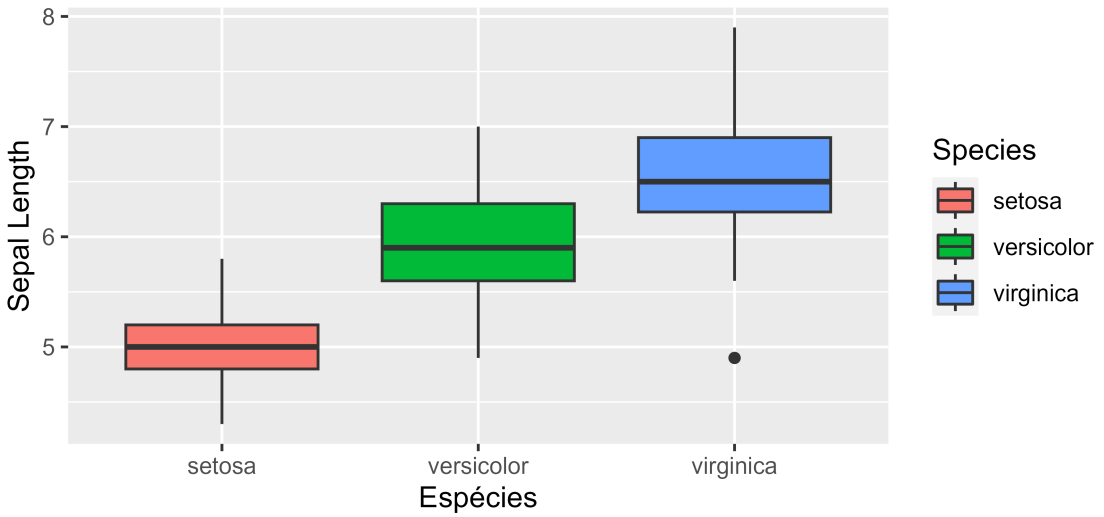
2.3.1.2 Código

As saídas para os blocos de códigos (*chunks*) podem ser personalizadas de diversas formas. É possível modificar a largura e altura de gráficos, por exemplo, bem como ocultar a apresentação dos códigos em seu documento, deixando apenas os resultados. Para tanto, os argumentos podem ser inseridos precedidos de `#|`,

conforme o exemplo a seguir:

```
#| label: code_02
#| fig-width: 4
#| fig-height: 2
#| warning: false
#| fig-align: "center"
#| echo: false
library(ggplot2)
plot <- ggplot(iris,
  aes(x = Species,
      y =  Sepal.Length)) +
  geom_boxplot(aes(fill = Species)) +
  xlab("Espécies") +
  ylab("Sepal Length") +
  theme_grey();plot
```

Figura 2.6: Resultado exemplificativo.



Fonte: A autora.

Note que foi definida uma largura (*fig-width*), altura (*fig-height*), alinhamento (*fig-align*) e evitou-se que avisos gerados pelo código apareçam no documento utilizando `warning: false`.

Diversas configurações podem ser utilizadas nos *chunks* e na Figura 2.7 apresenta-se algumas possibilidades:

Figura 2.7: Configurações de blocos de código.

Configuração	Descrição
`#| fig-width: 5`	Largura padrão para figuras geradas por gráficos em R (ou Matplotlib);
`#| fig-height: 3`	Altura padrão para figuras geradas por gráficos em R (ou Matplotlib);
`#| fig-align: "center"`	Alinhamento horizontal da figura (pode ser `left`, `right` ou `center`);
`#| message: false`	Se `false`, omite as mensagens do código;
`#| warning: false`	Se `false`, omite os avisos do código;
`#| echo: false`	Se `false`, omite o código e mostra somente a saída;
`#| eval: false`	Se `false`, mostra somente o código do chunk (sem rodar o trecho do código);
`#| out.width: "90%"`	Para especificar a largura das saídas;
`#| fig-cap: "Minha imagem"`	Adicione legenda na figura.

Fonte: (QUARTO, 2023a).

2.3.1.3 Quarto Pub

O Quarto Pub é um serviço gratuito de publicação de conteúdo criado com o Quarto. Os documentos criados são **visíveis publicamente** e fáceis para publicar. Para a publicação, alguns passos devem ser seguidos:

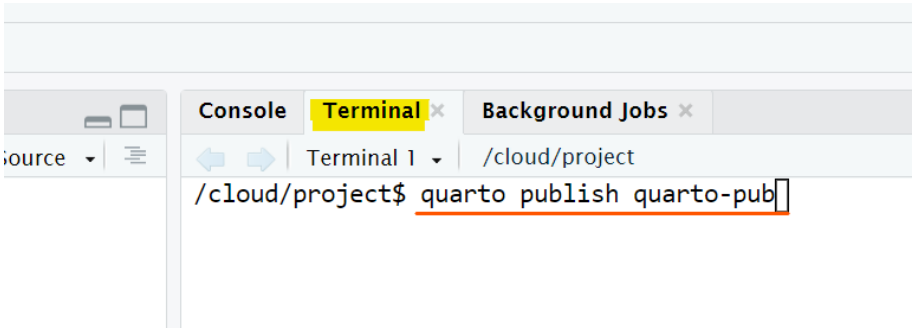
Passo 1) Criar uma conta gratuita no Quarto Pub;

Passo 2) Executar o comando no **Terminal**:

```
quarto publish quarto-pub
```

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 2.8: Configuração no terminal.



Fonte: A autora.

Passo 3) Aguardar e seguir as instruções que aparecerão na tela:

```
$ quarto publish quarto-pub
? Authorize (Y/n) >
> In order to publish to Quarto Pub you need to
  authorize your account. Please be sure you are
  logged into the correct Quarto Pub account in
  your default web browser, then press Enter or
  'Y' to authorize.
```

Passo 4) Após autorização e autenticação na conta, volte ao RStudio para confirmar que deseja publicar;

Passo 5) Aguardar renderizar e implantar. Em seguida, uma janela do seu navegador será aberta e o seu conteúdo estará pronto para visualização.

Os detalhes de publicação no Quarto Pub podem ser visualizados através do link. Em face às etapas de desenvolvimento abordadas nesta subseção, destaca-se a possibilidade de publicação não somente do *Quarto Book*, mas também de *blogs* e demais documentos desenvolvidos com Quarto.

2.4 RESULTADOS E DISCUSSÃO

Nesta seção apresenta-se os resultados obtidos no desenvolvimento de um modelo prático de um livro desenvolvido com *Quarto Book*. Todos os códigos e pastas relacionadas ao modelo apresentado podem ser acessados no repositório disponível em (FARIAS, 2023b). Na Figura 2.9 é possível visualizar a capa do livro:

Figura 2.9: Capa do modelo.

Fonte: A autora.

Perceba que metadados e outras informações são dispostas na capa, sendo possível a inserção de autores, imagens, índice entre outras. Tais configurações são inseridas no arquivo `_quarto.yml`, que contém:

```
project:
  type: book
  output-dir: docs
```

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

```
bibliography: references.bib # Referências
lang: pt # Idioma: português (pt)

book:
  title: "Contando histórias com Quarto" # Título
  subtitle: "Quarto Book" # Subtítulo do livro
  author:
    - name: "Ariane Hayana" # Nome Autor(a) 1
      orcid: "0000-0003-1571-8739" # ORCID Autor(a) 1
    - name: "Autor 2" # Nome Autor(a) 2
      orcid: "0000-0000-0000-0000" # ORCID Autor(a) 2
    - name: "Autor 3" # Nome Autor(a) 3
      orcid: "0000-0000-0000-0000" # ORCID Autor(a) 3

  date: "04/28/2023" # Inserir no formato mm/dd/yyyy
  favicon: logo.png # Icone na aba do navegador
  sidebar:
    style: docked # Estilo: 'docked' ou 'floating'
  search: true
  repo-url: https://github.com/a-hayana/ser2023 # Url
    do repositório

  sharing: [twitter, facebook, linkedin] # Redes
  reader-mode: true #oculta a barra lateral e o sumário
  page-footer:
    left: "SER | Seminário Internacional de Estatística
       com `R`" # Rodapé esquerdo
    right: "Por: Ariane Hayana" # Rodapé direito
```

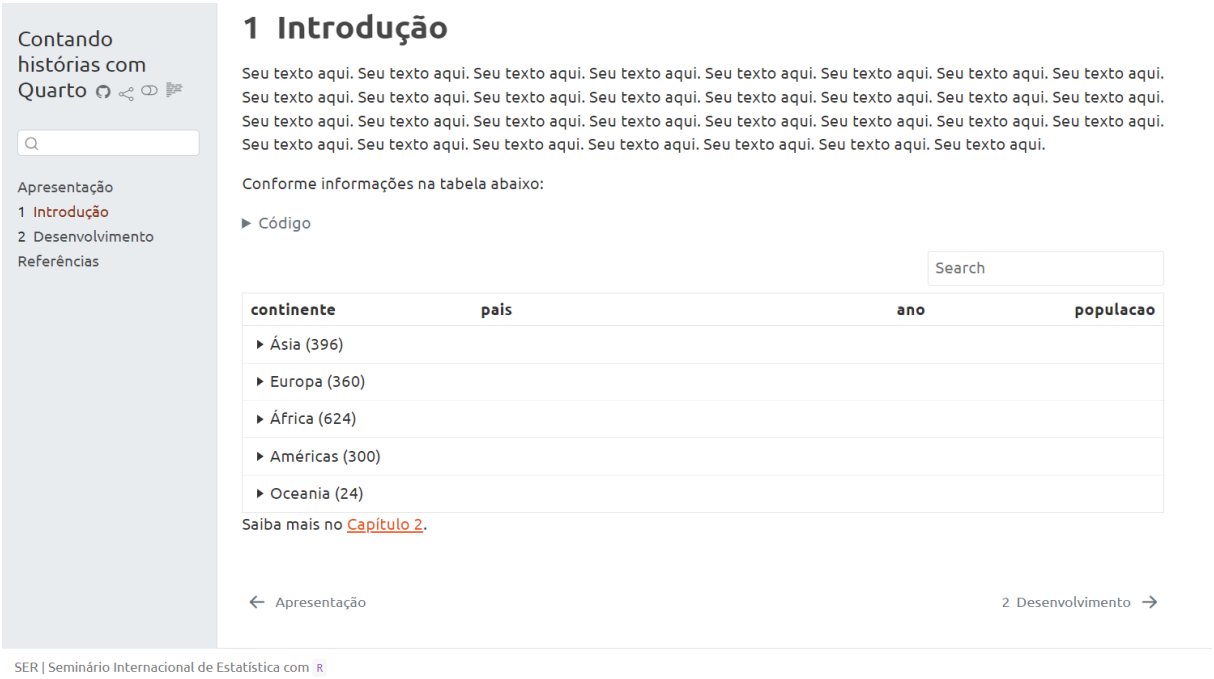
```
  chapters:
    - index.qmd
    - introducao.qmd
    - desenvolvimento.qmd
    - referencias.qmd
format:
  html:
    theme:
      light: [united,theme.scss]
      dark: superhero
    cover-image: logo.png
editor: visual
```

Note que o formato do livro foi definido em `HTML`, porém, outras configurações são possíveis, tais como `.pdf` e `.epub`, por exemplo. Após definidas as configurações desejadas, partiu-se para etapa de criação dos capítulos do livro, que ficam acessíveis na barra lateral esquerda da Figura 2.9. Em cada capítulo foram inseridas os seguintes conteúdos:

- **Introdução** - Neste capítulo, foram dispostos no modelo de criação:

Veja que neste capítulo foi possível inserir o texto, bem como referenciar outros capítulos, além de poder introduzir uma tabela interativa criada com o pacote `reactable` (LIN, 2022) e com dados oriundos do pacote `dados` (QUIROGA et al., 2022). Para tornar o texto mais limpo, os códigos da tabela foram ocultados utilizando no *chunk* `code-fold:true`. É possível navegar entre os capítulos ao final da página.

Figura 2.10: Capítulo *Introdução*.



Fonte: A autora.

- **Desenvolvimento** - No desenvolvimento, além da parte textual, os conteúdos foram divididos em subseções (conforme apresentado na Seção 2.3.1.1). Outra aplicação interessante é a utilização de abas para apresentar conteúdos diversos, sem sair do capítulo. Este trecho foi desenvolvido com a utilização do `panel-tabset`, conforme o trecho do código a seguir:

```
::: panel-tabset
## Aba 1
![](img/img1.png){fig-align="center"}
## Aba 2
![](img/img2.png){fig-align="center"}
:::
```

Note que para cada aba é definido um título que aparecerá na parte superior e dará acesso para cada conteúdo.

Figura 2.11: Exemplo de *Referências* no *Quarto Book*.

Fonte: A autora.

Também temos a possibilidade de inserir equações, criar exemplos e exercícios. No exemplo do modelo temos um exercício e solução, conforme apresentado:

Exercício 4.1. Suponha que as notas de uma turma de 5 alunos foram: 6, 7, 8, 9 e 10. Calcule a média.

Solução

Para calcular a média das notas, podemos usar a seguinte fórmula:
$\text{Média} = \frac{\text{soma das notas}}{\text{número de alunos}}$
Para calcular a média, podemos seguir os seguintes passos:
$\text{Soma} = 6 + 7 + 8 + 9 + 10 = 40$
$\text{Média} = \frac{\text{soma}}{\text{número de alunos}} = \frac{40}{5} = 8$
Portanto, a média das notas da turma é 8.

Outro ponto relevante é a praticidade de inserção de referências e citações no *Quarto Book*, cujas referências são citadas e referenciadas em um arquivo `.bib`, onde serão preenchidas as informações de autores(as), sendo possível também a citação de pacotes, livros, artigos, entre outros. Tais elementos podem ser dispostos conforme a normatização desejada e aparecerão no capítulo *Referências* padronizados conforme mostra a Figura 2.11.

Finalizadas todas as etapas de criação, ao final parte-se para a publicação do material criado no Quarto Pub. Seguindo as etapas sugeridas na Seção 2.3.1.3, ao final, o livro estará disponível para acesso em qualquer navegador. No exemplo

deste livro modelo, o mesmo está acessível no endereço https://ariane.quarto.pub/contando-historias-com-quarto/.

2.5 CONCLUSÃO

Neste capítulo apresentou-se orientações e definições de como desenvolver um *Quarto Book*. Foram abordados conceitos iniciais, explorando desde a instalação do Quarto até a etapa de publicação do livro no Quarto Pub. Dentre as principais contribuições, destaca-se a possibilidade de automatizar conteúdos, além de trazer interatividade e compartilhamento de informações através da publicação do conteúdo criado no Quarto Pub. Viu-se neste capítulo algumas das diversas possibilidades de criação e personalização de um livro desenvolvido com Quarto, e assim como qualquer outra ferramenta, é importante destacar que o ingrediente principal para o aprendizado do Quarto é a *curiosidade* em explorar a rica documentação disponível no Quarto.org.

2.6 REFERÊNCIAS

FARIAS, Ariane Hayana Thomé de. **E aí, vamos falar de Quarto?** Tutorial: Relatório em Quarto. [S.l.]: R-Ladies São Paulo, jul. 2023. Disponível em: https://rladies-sp.org/posts/2023-02-tutorial-quarto/. Acessado em: 23 jul. 2023.

_____. **Repositório do projeto SER 2023**. [S.l.: s.n.], 2023. Disponível em: https://github.com/a-hayana/ser2023. Acessado em: 16 set. 2023.

GRUBER, John. **Markdown: Syntax**. [S.l.: s.n.], jul. 2023. Disponível em: https://daringfireball.net/projects/markdown/syntax. Acessado em: 23 jul. 2023.

LIN, Greg. reactable: Interactive Data Tables Based on 'React Table', 2022.

QUARTO. **Code Cells: Knitr**. [S.l.: s.n.], jul. 2023. Disponível em: https://quarto.org/docs/reference/cells/cells-knitr.html. Acessado em: 23 jul. 2023.

_____. **Markdown Basics**. [S.l.: s.n.], jul. 2023. Disponível em: https://quarto.org/docs/authoring/markdown-basics.html. Acessado em: 23 jul. 2023.

QUIROGA, Riva et al. dados: Translate Datasets to Portuguese, 2022.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Capítulo 3

O QUARTO COMO FERRAMENTA DE PKM PARA PESQUISA CIENTÍFICA

Autor: Marcus Antonio Cardoso Ramalho e Ariel Levy
Universidade Federal Fluminense, UFF-RJ
e-mail: marcusantonio@id.uff.br e alevy@id.uff.br

Neste capítulo o leitor será apresentado a conceitos relacionados a gestão do conhecimento pessoal, programação letrada e desenvolvimento prático de pesquisa com o sistema QUARTO. Para isso será apresentado o modelo *seek*»sense»*share* de Harold Jarche e como ele pode ser aplicado em cada uma das etapas do desenvolvimento de uma pesquisa científica feita usando o QUARTO. Além disso também serão apresentados métodos, técnicas e fluxos para o uso da ferramenta dentro desse contexto. Espera-se que ao final do capítulo o leitor tenha uma visão geral sobre o tema e possa explorar a documentação oficial para se aprofundar nos temas que mais lhe interessarem.

3.1 INTRODUÇÃO

O termo gestão do conhecimento é amplamente conhecido e vem sendo estudado e aplicado em grandes organizações a muito tempo. Já a gestão do conhecimento pessoal, ou *personal knowledge management* (PKM), é um termo mais

recente pois as discussões sobre o tema só ganharam força nas últimas duas décadas.

Um conceito amplo de PKM é desenvolvido por (KASSIM; SHA'ARI; BAHARUDDIN, 2018) que traz uma abordagem de processo quando diz que a gestão do conhecimento pessoal é uma estratégia de expansão do conhecimento pessoal que envolve a transformação da informação desorganizada em conhecimento útil através da obtenção, avaliação, organização , colaboração, análise e apresentação do conhecimento. Neste ponto é importante destacar a importância da administração da informação, isso fica evidente quando verificamos que as ferramentas de PKM, sejam softwares ou modelos tem grande foco na administração da informação (VÖLKEL; HALLER, 2009).

Existem outras abordagens para o PKM, como a de (JARCHE, 2014), que apresenta algumas estruturas para entender os mecanismos por traz da prática. Para (JARCHE, 2014), o domínio sobre a gestão do conhecimento pode ser buscado através do modelo Seek»*sense*»*Share*, que aprofunda o conceito geral como o que é apresentado por (KASSIM; SHA'ARI; BAHARUDDIN, 2018) e outros autores.

Com o crescimento nas pesquisas sobre o PKM surgiram diversas ferramentas e frameworks que buscam de alguma forma viabilizar ou facilitar esse processo. Porém, para aplicar os conceitos de PKM não é necessário ter uma ferramenta específica, pode-se usar ou adaptar de acordo com disponibilidade de recursos.

Neste sentido as iniciativas de software livre ajudam a criar um ambiente perfeito para implementação das práticas de gestão do conhecimento pessoal que ao contrário da gestão do conhecimento tradicional aplicada por grandes empresas, prevê a disseminação do conhecimento de forma aberta.

Neste trabalho será explorado o QUARTO (ALLAIRE et al., 2022), um sistema de código aberto para publicação técnico-científica, que possui características que ajudam na viabilização de práticas de PKM. Para facilitar a compreensão serão explorados os conceitos de PKM segundo a metodologia *seek, sense, share*

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

de (JARCHE, 2014).

3.2 OBJETIVO

O objetivo deste capítulo é entender e explorar o QUARTO como uma ferramenta de PKM aliado à pesquisa científica, utilizando a metodologia *seek, sense, share*. Para isso serão apresentados os conceitos estudados por (JARCHE, 2014) e como eles podem ser aplicados em cada uma das etapa do desenvolvimento de uma pesquisa científica feita usando o QUARTO. Além disso, também serão apresentados métodos e técnicas para a aplicação do PKM em cada uma das etapas.

3.3 O QUARTO

Quem já usou R por algum tempo, provavelmente teve contato com o rmarkdown (XIE; ALLAIRE; GROLEMUND, 2019), um formato de documento baseado em markdown que permite a editoração de arquivos que podem intercalar linguagens de programação como R e Python com texto, algo semelhante ao conhecido Jupyter Notebook. Essa ideia de intercalar texto e código não é nova e foi explorada por (KNUTH, 1984), que criou o conceito de programação letrada, onde o programa é feito em um notebook que contém o código e a explicação ou documentação.

Antes da criação do rmarkdown, o conceito criado por Knuth foi aplicado usando noweb (JOHNSON; JOHNSON, 1997) e Sweave no R (FRIEDRICH LEISCH, 2002), que permitiam o uso de Latex e R dentro do mesmo documento com a intenção de facilitar a reprodução de resultados de pesquisas científicas. Assim, inspirado pelo conceito de programação letrada e pelo Sweave (XIE, 2014) introduziu o Knitr que usava os conceitos do Sweave e dos pacotes que foram criados para suprir suas limitações, como o cache de resultados e a possibilidade de usar outros formatos de saída além do Latex, como HTML e Markdown, além de viabilizar o uso de outras linguagens de programação além do R, culminando

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

posteriormente no formato que se popularizou, o rmarkdown.

A partir do desenvolvimento do rmarkdown, a Posit apresentou em 2022 o QUARTO (ALLAIRE et al., 2022), um sistema completo de editoração e publicação técnico-científica baseado em Pandoc que expandiu as possibilidades criadas pelo rmarkdown dentro do conceito de programação letrada. O quarto permite a criação de diversos tipos de documentos, de artigos e livros a sistemas de enciclopédia (Wikis) e outros tipos de sites como blogs. Seguindo a tradição de software livre do rmarkdown, o QUARTO é um sistema de código aberto e pode ser usado gratuitamente por qualquer pessoa ou organização. A característica mais marcante é possibilidade de editar os arquivos com a linguagem de marcação preferencial do usuário, é possível usar Latex, html, markdown, css e scss por exemplo.

Considerando a história que levou ao desenvolvimento do QUARTO é possível traçar um paralelo entre os paradigmas de programação letrada e os conceitos de PKM pois o objetivo final de ambos está associado a construção e disseminação do conhecimento. Além disso, o próprio processo que levou ao desenvolvimento das soluções para programação letrada dentro de um ambiente de desenvolvimento de software livre, representado pela linguagem de programação R, também pode ser associado ao conceito de PKM, pois software livre envolve além do próprio desenvolvedor, os usuário e comunidades práticas criadas em ambientes abertos como o GitHub, onde o QUARTO e o rmarkdown foram criados e são mantidos.

3.4 COMO USAR O QUARTO

Ao contrário do rmarkdown o QUARTO não está preso a uma linguagem ou ambiente de programação único, é possível usá-lo com qualquer editor de texto ou IDE e renderizar os documentos através do terminal do sistema operacional após instalar a sua interface de linha de comando. Porém, ao usar ambientes de programação como o VSCode, RStudio ou Jupyter é possível fazer a edição e renderização dos documentos através de interface visual e atalhos de teclado.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Além do R, os desenvolvedores oferecem oficialmente suporte para Python, Julia e Observable.

Outra característica é a facilidade para publicação dos documentos criados, qualquer usuário pode criar um blog, por exemplo, sem ter conhecimento prévio em html. Basta criar um projeto com esse formato predefinido e começar a escrever e publicar posteriormente em um servidor ou através dos serviços oferecidos pela POSIT, GitHub ou outras empresas que oferecem hospedagem grátis. Um exemplo de site criado com esse sistema pode ser visto em https://quarto.org.

Para instalar o sistema é necessário baixar o instalador do site oficial https://quarto.org/docs/get-started/, escolhendo o sistema operacional que será usado.

Também é possível realizar a instalação em ambiente Linux em distribuições como o Ubuntu, basta executar os comandos da Figura 3.1 no terminal (ALLAIRE et al., 2022) ou seguir as instruções atualizadas do site oficial https://docs.posit.co/resources/install-quarto/.

Figura 3.1: Instalação do QUARTO em ambiente Linux.

```
sudo curl -LO https://quarto.org/download/latest/quarto-linux-
   amd64.deb
sudo apt-get install gdebi-core
sudo gdebi quarto-linux-amd64.deb
```

Fonte: Os autores.

Após a instalação e criação de um arquivo com a extensão `.qmd`, o próximo passo em um documento básico é a criação de um cabeçalho YAML, para viabilizar a criação dos metadados do documento. O cabeçalho é delimitado por três traços e pode conter informações como título, autor, data, resumo, palavras chave, etc. Um exemplo de cabeçalho YAML pode ser visto na Figura 3.2.

Se o formato não for especificado usando o parâmetro `format` no cabeçalho, o documento será renderizado como html por padrão. Além disso, é possível explicitar também a engine usada que pode variar de acordo com o ambiente de

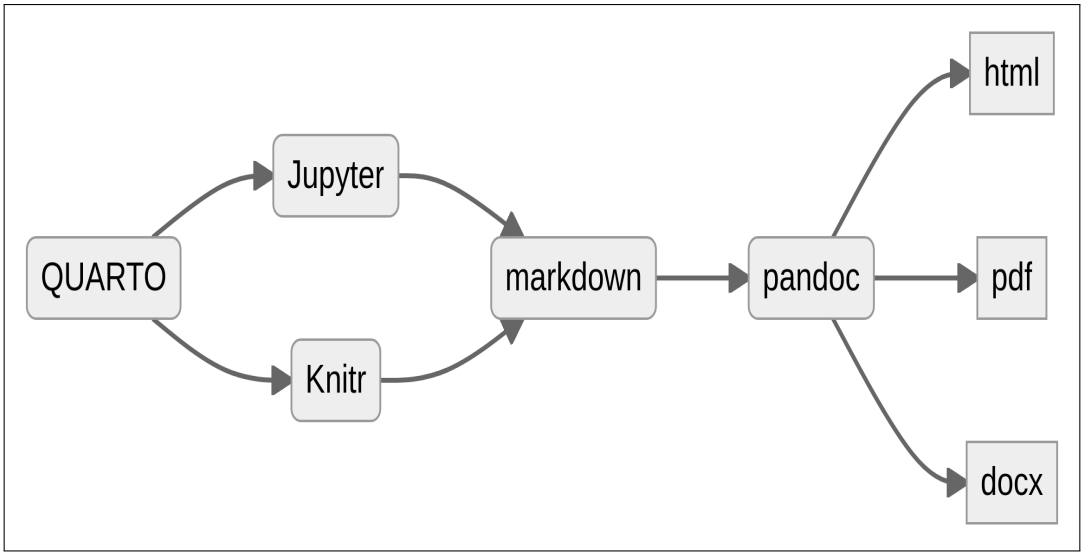
Figura 3.2: Exemplo de cabeçalho YAML.

```
---
title: O Quarto como ferramenta de PKM para pesquisa científica
author: Marcus Antonio Cardoso Ramalho
date: 2023-07-30
abstract: |
        Este é um exemplo de abstract.
keywords: [quarto, pkm, pesquisa científica]
format: pdf
bibliography: References.bib
csl: apa.csl
---
```

Fonte: Os autores.

programação escolhido. A Figura 3.3 mostra o fluxo de renderização padrão de um documento usando o QUARTO, onde é possível verificar que a primeira etapa consiste da conversão para o formato markdown, que é feita pelo Jupyter, no ambiente do Rstudio essa etapa é feita pelo knitr.

Figura 3.3: Fluxo de renderização de um documento usando o QUARTO.



Fonte: Os autores.

3.5 O PKM PARA HAROLD JARCHE

A ideia geral por trás de qualquer conceito de PKM é a de que o conhecimento pessoal é desenvolvido a partir de interações entre o indivíduo, as ferramentas disponíveis para organização da informação e a rede de pessoas que o indivíduo interage. Nesse sentido, o PKM pode ser visto como um processo de transformação da informação em conhecimento, onde o indivíduo busca, avalia, organiza, colabora, analisa e apresenta o conhecimento adquirido (KASSIM; SHA'ARI; BAHARUDDIN, 2018).

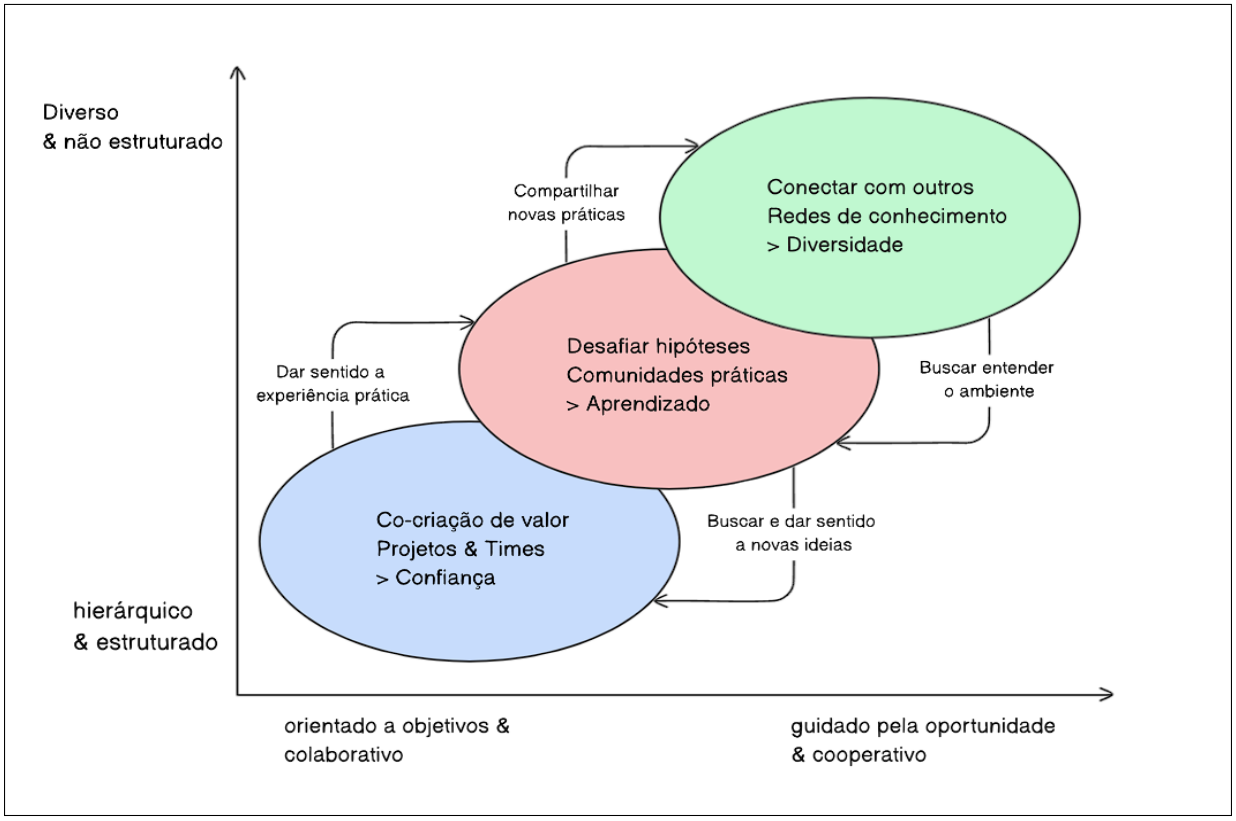
Existem diversos frameworks que sintetizam esse processo e em certo grau todos eles acabam convergindo para o mesmo conceito básico. Harold Jarche tem sido uma referência sobre conhecimento pessoal nas últimas décadas, ele desenvolveu um framework para o processo de desenvolvimento do conhecimento pessoal chamado PKM (Personal Knowledge Mastery) (JARCHE, 2014). O framework de Jarche é baseado em três etapas: *Seek»Sense»Share*, que podem ser traduzidas como Buscar»Entender»Compartilhar.

Neste *framework* existe a ênfase de que o conhecimento é construído através da prática individual, porém, deve ser compartilhado para que possa ser validado e enriquecido através da interação com outras pessoas. Neste sentido (JARCHE, 2014) ressalta o papel das comunidades de prática como um ambiente propício para o desenvolvimento do conhecimento pessoal.

Podemos interpretar a Figura 3.4 apresentada por (JARCHE, 2014) como um ciclo que parte de um nível diverso e não estruturado de informação (*seek*) que é guiado pela curiosidade e colaboração para um nível mais estruturado de conhecimento (*sense*) que é validado e enriquecido através da interação com outras pessoas e comunidades de prática, culminando na disseminação do conhecimento (*share*), construído de forma estruturada e orientada a objetivos específicos de forma colaborativa.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 3.4: Modelo *seek sense share*.



Fonte: (JARCHE, 2014).

Essa estrutura conversa com o desenvolvimento de uma pesquisa científica moderna que é baseada em um processo de construção de conhecimento que parte de uma revisão da literatura (*seek*) para a construção de um arcabouço teórico e desenvolvimentos de hipóteses, experimentos ou outros métodos que estejam dentro do escopo do método científico (*sense*) que então são validados ou não através da interação com outros pares através da comunicação da pesquisa em períodicos, congressos e outros meios de comunicação científica (*share*).

3.6 O QUARTO COMO FERRAMENTA DE PKM NA PESQUISA CIENTÍFICA

Por ter sido desenvolvido dentro de um ambiente de software livre, o QUARTO é uma ferramenta que se encaixa perfeitamente no conceito de PKM, pois permite

a criação de documentos que podem ser compartilhados e editados por qualquer pessoa. Além disso, seu principal objetivo é servir como um instrumento multi-plataforma para a publicação de documentos técnicos e científicos. Isso, aliado ao uso de plataformas como o GitHub ou GitLab, por exemplo, permite a criação de um ambiente de colaboração e disseminação do conhecimento, que pode ser usado para viabilizar o processo de PKM de forma individual ou coletiva.

Considerando o modelo desenvolvido por (JARCHE, 2014) podemos enquadrar essa ferramenta nas três fazes do ciclo do PKM (*seek»sense»share*), porém, cabe a cada pesquisador adaptar as especificidades de sua pesquisa as potencialidades e limitações do QUARTO.

3.6.1 seek e sense

Quando se inicia um processo de pesquisa, o cientista geralmente busca junto a outros pesquisadores ou na literatura as informações que irão embasar todo o trabalho, nessa etapa temos um acessório indispensável, o notebook. Seja em pesquisa quantitativa ou qualitativa, esse item está sempre presente. Através dessa ferramenta ideias de fontes diferentes podem se conectar criando redes de conhecimento onde é possível identificar padrões ou experimentar interações entre métodos e abordagens diferentes.

Nesse contexto, usando os conceitos de programação letrada (KNUTH, 1992) pode-se usar o quarto como um notebook dentro de uma wiki ou blog onde as informações podem ser organizadas e encontradas facilmente através do mecanismo de busca do próprio site. Dentro dessa proposta, cada post é um ou mais tópicos de pesquisa que podem ser consultados posteriormente com facilidade graças ao mecanismo de busca que vem integrado dentro de todo site QUARTO.

A grande vantagem do framework do QUARTO em relação a outras soluções como o Rmarkdow ou o Jupyter notebook é justamente a possibilidade de transitar entre diversas linguagens de programação e marcação sem sair do editor de texto ou IDE de escolha. Além disso, por ser uma iniciativa de código aberto, existe

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

a possibilidade de desenvolver ou obter extensões que atendam a especificidades não contempladas na instalação padrão.

Outro ponto a ser destacado é o de gerenciamento de referências bibliográficas, que é uma das principais dificuldades enfrentadas por pesquisadores inciantes. O QUARTO possui uma interface própria que também pode ser integrado com outros gerenciadores como o Zotero, porém, é possível realizar buscas pelo DOI e importar as referências diretamente do site da Crossref, DataCite ou PubMed. Esse recurso pode ser acessado de duas formas, através do menu de contexto do editor de texto (insert»citation) como mostrado na Figura 3.5 ou através de um atalho de teclado `ctrl+shift+f8`.

3.7 SHARE

É na viabilização do compartilhamento de conhecimento que o QUARTO mais se destaca, já que permite a publicação em mídias diversas, desde um blog simples até uma *wiki* ou um livro como este que você está lendo. Por ser baseado em markdown tem uma barreira de entrada relativamente baixa, o que habilita cientistas e pesquisadores de diversas áreas a publicar seus trabalhos de forma simples e rápida.

A documentação oficial traz ainda a possibilidade de usar ou criar *templates* de artigos de periódicos como o IEEE, ACM, Elsevier, Springer, etc. Assim, é possível criar *templates* personalizados para qualquer tipo de documento, como livros e relatórios, algo que é potencializado pela integração com Pandoc e Latex.

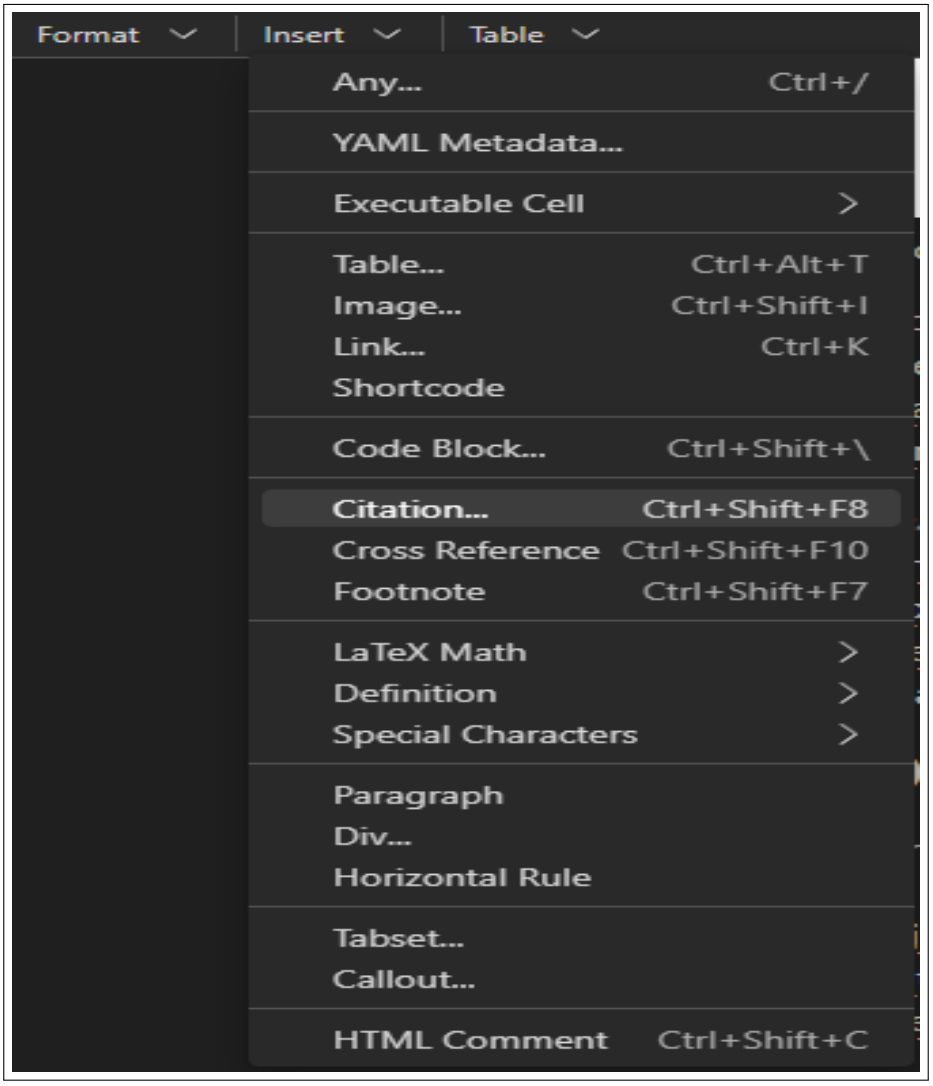
3.8 APLICAÇÃO

3.8.1 Criando um documento

Para ilustrar a aplicação do QUARTO como ferramenta de PKM na pesquisa científica, será apresentado um exemplo de uso do QUARTO para a criação de um projeto ou um documento. Para tanto deve-se abrir o terminal e digitar o comando

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 3.5: Menu para acessar a janela de citações.



Fonte: Os autores.

`quarto create` seguido do nome do projeto, como mostrado na Figura 3.6.

Para simplificar o primeiro contato com a ferramenta vamos criar um documento simples apresentando as principais funcionalidades usadas na criação de um texto acadêmico. O tipo do documento pode ser definido em dois níveis, o primeiro é o formato do arquivo de saída, que pode ser *html, pdf, docx, beamer ou revealjs* por exemplo. No caso de arquivos *beamer* e *revealjs* o resultado será renderizado para uma apresentação de slides. O segundo nível é o formato do documento, que pode ser um artigo, livro, relatório, etc. Para criar um documento

Figura 3.6: Criando um projeto ou arquivo QUARTO.

```
#para criar um projeto com quarto CLI:
quarto create blog "nome do projeto"
#para criar um arquivo quarto em uma pasta com linux
#ou mac pela linha de comando:
touch "nome do arquivo"\texttt{.qmd}
#para criar um arquivo quarto em uma pasta
#com windows pela linha de comando:
echo. > "nome do arquivo"\texttt{.qmd}
```

Fonte: Os autores.

simples, vamos usar o formato de artigo e o formato `html`, para isso basta criar um arquivo com a extensão `.qmd` e adicionar o cabeçalho YAML como mostrado na Figura 3.2. Após configurar o cabeçalho YAML, basta salvar o arquivo e renderizar o documento usando o comando `quarto render nome_do_arquivo` no terminal. Será gerado um html com o nome do arquivo criado.

O usuário iniciante pode se deparar com algumas limitações relacionadas a formatação ao usar o QUARTO, porém, a maioria delas pode ser resolvida com o uso de Latex, html ou css direto no corpo do documento. Também é possível usar pacotes Latex no YAML para resolver problemas de formatação ou criar templates como mostra a Figura 3.7 que usa o pacote `fancyhdr` para criar um cabeçalho e rodapé personalizado usando o argumento `include-in-header` no YAML.

Apesar de ser útil para quem já tem familiaridade com Latex essas configurações também podem ser feitas usando html, css ou argumentos do Pandoc no YAML. Isso demonstra a flexibilidade do QUARTO para lidar com diferentes formatos de documentos e linguagens de programação.

Um ponto que deve ser observado ao usar pacotes Latex em documentos QUARTO e o fato de que qualquer um que for usar o template ou reproduzir o documento em um ambiente diferente do original deve levar em consideração a necessidade de instalação das dependências, por isso, em casos de documentos complexos é recomendável usar pacotes que já estejam disponíveis na instala-

Figura 3.7: Exemplo de uso de pacotes Latex no YAML.

```
---
title: |
     \begin{center}
     \includegraphics[angle=0,keepaspectratio,width=3cm]{UFF.png}
     \end{center}
     \Large
     Universidade Federal Fluminense\
     Programa de Pós-Graduação em Administração\
     \vspace{4cm}
subtitle: \Large Título do artigo
documentclass: article
papersize: a4
format:
  pdf:
    include-in-header:
      - text: |
          \usepackage{fancyhdr}
          \usepackage{hyperref}
          \usepackage{multido}
          \pagestyle{fancy}
          \fancyhf{}
          \renewcommand{\headrulewidth}{0pt}
          \fancyfoot[R]{\thepage}
          \pagenumbering{arabic}
    latex_engine: xelatex
    fontsize: 12pt
    lineheight: 1.5
    linestretch: 1.5
    geometry: "left=2.54cm, top=2.54cm, right=2.54cm, bottom=2.54
       cm"
    keep-tex: false
crossref:
  fig-title: Figura
  fig-prefix: figura
  tbl-title: Tabela
  tbl-prefix: tabela
bibliography: references.bib
csl: apa.csl
editor: visual
---
```

Fonte: Os autores.

ção padrão do Latex ou que sejam amplamente usados e estejam disponíveis em repositórios como o CTAN.

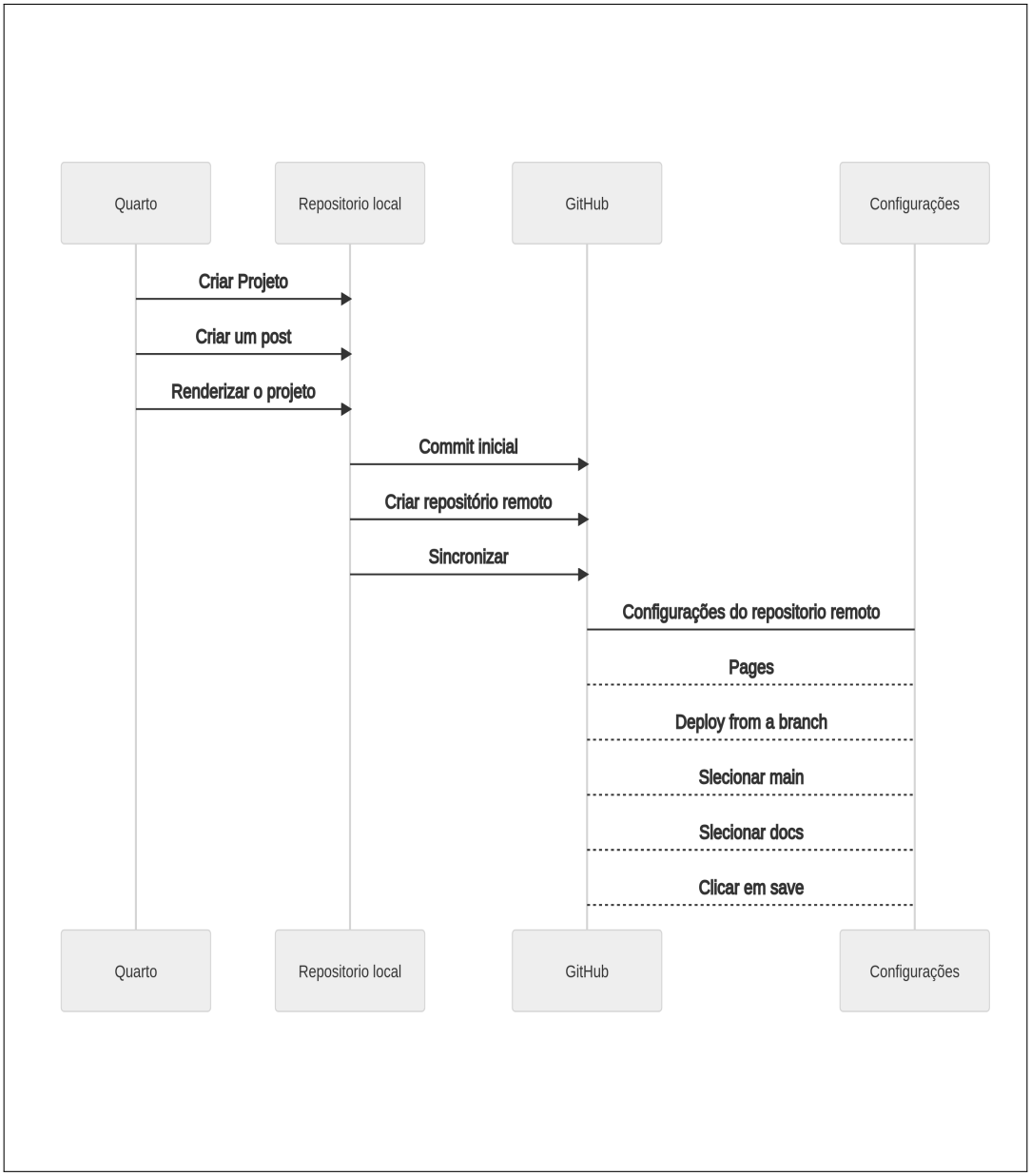
Além da personalização usando recursos de Latex, html e css também é possível usar *templates* dos principais *jounals* e editoras, para isso basta usar o comando `quarto use template quarto-journals/nome_do_template` que criará um artigo novo em branco. Para adicionar um *template* a um artigo em andamento o comando CLI é `quarto add quarto-journals/nome_do _template`. A lista completa de *templates* disponíveis pode ser encontrada na documentação oficial https://quarto.org/docs/journals/.

Além dos *templates* de periódicos também é possível usar *templates* para blogs, o que facilita a customização e permite que o usuário crie sites profissionais sem precisar ter conhecimento aprofundado em frameworks web ou conhecimento avançado em desenvolvimento front-end. Este recurso associado a possibilidade de publicação em plataformas como o GitHub Pages ou Netlify viabiliza que qualquer pesquisador com acesso a internet possa publicar seus trabalhos de forma simples, rápida, gratuita e independente.

A Figura 3.8 mostra o passo a passo para criação e publicação de um blog com quarto e GitHub Pages. Após criar o projeto o sistema inicializa um repositório git e cria uma estrutura de pastas e arquivos pré-definida. Ao editar ou criar e renderizar um arquivo QMD dentro da estrutura da pasta Posts, o QUARTO cria um arquivo html dentro da pasta _site, porém, no caso específico do GitHub Pages é necessário mudar o nome da pasta para docs, o que permitirá que o GitHub Pages carregue o site automaticamente. O problema da pasta docs pode ser resolvido ao adicionar no arquivo yaml do blog o códio mostrado na Figura 3.9.

Após a realização das etapas sugeridas, o GitHub irá carregar o site automaticamente e o blog estará disponível no endereço `https://seu_usuario.github.io/nome_do_repositorio/`. Para adicionar um novo post basta criar um arquivo QMD dentro da pasta Posts, renderizar e sincronizar o projeto com o repositório remoto.

Figura 3.8: Fluxo sugerido de criação e publicação de um blog com QUARTO e GitHub Pages.



Fonte: Os autores.

3.9 CONCLUSÃO

A intenção deste capítulo foi apresentar o leitor os conceitos básicos de gestão do conhecimento pessoal através do modelo *seek»sense»share* de (JARCHE, 2014) e como eles podem ser aplicados em etapas do desenvolvimento de uma pesquisa científica feita usando o QUARTO. Além disso também foram apresentados

Figura 3.9: Configuração da pasta raiz para renderização no GitHub Pages.

```
project:
  type: website
  output-dir: docs
```

Fonte: Os autores.

métodos e técnicas para usar o poder do QUARTO no que se refere principalmente as etapas de registro e disseminação do conhecimento através de *notebooks*, artigos e blogs.

Ademais, sugere-se que o leitor explore a documentação para aprender sobre temas não explorados neste capítulo na documentação oficial do QUARTO que pode ser encontrada no site https://quarto.org/ e também na documentação do Pandoc https://pandoc.org/MANUAL.html.

Por fim, deve-se destacar que o QUARTO é uma ferramenta de código aberto e que está em constante desenvolvimento, por isso, é importante verificar constantemente a documentação oficial para se manter atualizado sobre as novas funcionalidades e possibilidades de uso. Além disso, deve-se ter em mente que assim como qualquer ferramenta, o QUARTO é apenas um meio para viabilizar a prática da pesquisa. Portanto, este texto não apresentou uma "bala de prata" que pode resolver todos os problemas de um pesquisador, mas sim, alguns dos métodos mais recentes ao se tratar de desenvolvimento prático de ambientes de pesquisa científica amigáveis a linguagens como R e Python dentro do contexto de programação letrada.

3.10 REFERÊNCIAS

ALLAIRE, J. J. et al. **Quarto**. [S.l.: s.n.], 2022. Disponível em: https://doi.org/10.5281/zenodo.5960048.

FRIEDRICH LEISCH. Sweave, Part I: Mixing R and LATEX. **R News - The Newsletter of the R Project**, v. 2/3, 2002. ISSN 1609-3631.

JARCHE, Harold. **The Seek > Sense > Share Framework**. [S.l.: s.n.], 2014. Disponível em: https://jarche.com/2014/02/the-seek-sense-share-framework/. Accessed: 25-07-2023.

JOHNSON, Andrew L; JOHNSON, Brad C. Literate Programming Using Noweb, 1997.

KASSIM, N. A.; SHA'ARI, I.; BAHARUDDIN, K. Conceptualizing Personal Knowledge Management Enabler and Personal Knowledge Management Capability. **International Journal of Academic Research in Progressive Education and Development**, v. 7, n. 1, 2018. DOI: 10.6007/ijarped/v7-i1/3853.

KNUTH, D. E. Literate Programming. **The Computer Journal**, v. 27, n. 2, p. 97–111, 1 fev. 1984. DOI: 10.1093/comjnl/27.2.97.

_____. **Literate Programming**. [S.l.: s.n.], 1992.

VÖLKEL, Max; HALLER, Heiko. Conceptual data structures for personal knowledge management. **Online Information Review**, v. 33, n. 2, p. 298–315, 2009. DOI: 10.1108/14684520910951221.

XIE, Y. **Dynamic Documents with R and Knitr**. [S.l.]: CRC Press, Taylor & Francis, 2014.

XIE, Yihui; ALLAIRE, J. J.; GROLEMUND, Garrett. **R Markdown: the definitive guide**. Boca Raton: CRC Press, Taylor e Francis Group, 2019. ISBN 978-0-429-78296-1.

Capítulo 4

SERIES DE TIEMPO CON R

Autor: Manuel Febrero-Bande
Departamento de Estadística, Análisis Matemático y Optimización
Universidade de Santiago de Compostela
e-mail: manuel.febrero@usc.es

Este documento pretende ser una guía introductoria al análisis clásico de las series temporales en R que, en general, está incluido en los paquetes que se instalan por defecto en una instalación básica. No se pretende aquí hacer una documentación exhaustiva de todas las posibilidades que hay disponibles en R sino más bien proporcionar una serie de recomendaciones y advertencias para empezar a analizar series macroeconómicas y financieras al nivel más básico. Por supuesto, esto es sólo el punto de partida y una profundización en la documentación que proporcionan los paquetes de R es necesaria para adaptar al interés de cada caso particular.
Palabras-Clave: R; Series temporales; Paquetes de R.

4.1 INTRODUCCIÓN

Simplificando mucho, una serie temporal (que denotaremos por $\{X_t\}_{t\in T}$) es una secuencia ordenada de variables aleatorias que se suponen dependientes o correladas respecto al índice t (usualmente el tiempo). Matemáticamente hablando,

una serie temporal es la realización de un proceso estocástico $X(t, \omega)$ del que estamos observando sólo una trayectoria (fijado ω) y por tanto, deberemos suponer ciertas condiciones de homogeneidad a lo largo del índice para poder hacer inferencias/predicciones acerca del proceso generador de la serie temporal.

Atendiendo a las características del índice, podemos clasificar las series de tiempo de la siguiente manera:

- Continuas: La serie temporal puede ser observada en cualquier instante de tiempo del intervalo T
- Discretas: La serie temporal sólo puede ser observada en ciertos instantes prefijados o cada cierto intervalo de tiempo $\{t_1, \ldots, t_n\} \in T$
  - Serie de Tiempo Regular: El intervalo entre índices consecutivos es constante, i.e. $|t_{j+1} - t_j| = \delta$.
  - Serie de Tiempo Irregular: El intervalo entre índices consecutivos no es constante e incluso el proceso de aparición de observaciones podría ser interesante desde el punto vista estadístico.

En lo que sigue nos centraremos en las series de tiempo más clásicas que se corresponden con series de tiempo regulares aunque en R existen herramientas para tratar los otros tipos de series.

Los objetivos usuales del análisis de las series de tiempo serán:

- Entender el mecanismo generador de observaciones.
- El control óptimo de un sistema dinámico.
- Predicción de valores futuros.

Es importante tener en cuenta características o condiciones de las series temporales que pueden influir en el tipo de análisis o en la elección de la herramienta

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

apropiada. También, para solventar alguna de las siguientes consideraciones, puede ser necesario filtrar la información recibida o transformarla para conseguir una señal más homogénea.

- *Tipo de la escala de tiempo.* La escala de tiempo de los índices de la serie temporal puede ir desde segundos (o incluso fracciones de segundo) hasta intervalos de varios años (o milenios). Así podemos hablar según la escala de tiempo de series minutales, horarias, diarias, semanales, mensuales, cuatrimestrales, anuales, etc. Las series macroeconómicas clásicas suelen tener escalas superiores a la mensual mientras que la modelización de series financieras suele estar en escalas diarias o incluso en frecuencias *intradia* que son conocidas como series de alta frecuencia.

- *Agregación.* El valor observado a intervalos regulares puede provenir de una agregación de un proceso a lo largo del intervalo como podría ser, por ejemplo, la producción industrial mensual de un bien o ser una observación instantánea observada a intervalos regulares, como podría ser el precio al cierre de mercado de un activo financiero. Ambas situaciones pueden diferir en como se estructura la dependencia temporal entre observaciones.

- *Problemas de calendario.* Algunas series de tiempo se ven afectadas por el calendario que es una convención para repartir el ciclo anual (órbita de la Tierra alrededor del Sol) en días (ciclo de rotación terrestre), semanas y meses. Como consecuencia, tenemos meses de distinta longitud o fiestas movibles en el calendario y ambas situaciones podrían afectar a la serie temporal. El ejemplo más paradigmático es el de la Semana Santa que según los años puede caer en Marzo o en Abril y, en términos trimestrales, puede asignarse al primer o segundo trimestre según lo anterior.

- *Cambios en el valor del dinero.* En aquellas series temporales que reflejen valores monetarios puede ser necesaria una deflación respecto a un índice de precios para evitar tendencias sistemáticas debidas a la inflación.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

- *Estacionalidad.* En muchas series de tiempo se observan ciclos o estacionalidades debidas a la propia naturaleza de la serie de tiempo. Por ejemplo, es esperable un ciclo diario si trabajamos con temperaturas horarias o anuales si hablamos de producciones industriales.

Por último, a continuación se listan los paquetes de R más usuales por tópico de aplicación. La lista no pretende ser exhaustiva sino más bien el punto de comienzo para investigar sobre alguno de los apartados más básicos del análisis en las series de tiempo. Un buen punto de partida es la TaskViews sobre series de tiempo disponible en CRAN (`https://cran.r-project.org/web/views/TimeSeries.html`). En secciones posteriores se verán con más detalle alguno de los paquetes aquí descritos.

- Tratamiento de Fecha/Hora
  - `base`, `zoo` ((ZEILEIS, 2023)), `chron` ((HORNIK, 2023)), `lubridate` ((SPINU, 2023)), `timeDate` ((BOSHNAKOV, 2023a))
- Importación de datos
  - `quantmod` ((ULRICH, 2023)), `fImport`
- Modelización ARIMA
  - Básico: `stats`, `forecast` ((HYNDMAN, 2023))
  - Multivariante: `stats`, `vars`, `MTS`
  - Otras opciones: `FinTS`, `StructTS`, `TSA`
- Modelización GARCH
  - `rugarch` ((GALANOS, 2022)), `rmgarch`, `tseries`
- Tests de raíz unitaria
  - `urca`, `fUnitRoot`

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

- ARIMA fraccional
  - `fracdiff`

4.2 FECHA Y HORA

La primera tarea que un analista de series de tiempo debe abordar es como almacenar y gestionar la información sobre el índice t que usualmente será la fecha y hora de la obtención del valor. Por supuesto, los objetos básicos para la representación de fechas y horas están en el paquete `base` que, suelen ser suficientes para la mayoría de aplicaciones. El uso de otros paquetes para la gestión de fechas/horas es recomendable cuando se quieran usar las extensiones que estos nuevos paquetes proponen.

4.2.1 Objetos básicos de Fecha/Hora

En todos los lenguajes de programación la fecha/hora se suele guardar como el número de días transcurridos desde una fecha de referencia (habitualmente 15/10/1582, 14/9/1752, 01/01/1970). El objeto internamente es un número real donde la parte entera corresponde al intervalo de días y la parte decimal sirve para guardar unidades de rango inferior (horas, minutos y segundos).

- `Date`: Días desde 01/01/1970.
- `POSIXct`: Segundos desde 01/01/1970. Permite una precisión mayor.
- `POSIXlt`: Fecha/hora en formato lista con componentes: `sec:0-61, min:0-59, hour:0-23, mday:1-31, mon:0-11, year:  since 1900, yday:0-365` .
- `Sys.getlocale()`, `Sys.setlocale()`: Comandos para consultar/cambiar la configuración local. Es útil cuando se quiere leer/formatear fechas que no están en el idioma propio del equipo.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

- Formato de escritura/lectura de fechas (véase la ayuda `strptime` para más detalles).
  - Día: Día de la semana: `%a` (abreviado),`%A` (completo). Día del mes: `%d` (01–31). Día del año: `%j` (001–366)
  - Mes: Nombre del mes: `%b` (abreviado), `%B` (completo). Número de mes: `%m` (01–12).
  - Año: `%y` (corto), `%Y` (largo)
  - Hora: Formato 24 hr. `%H` (00–23). Formato 12 hor. `%I %p` (01–12, AM/PM)
  - Minutos: `%M` (00–59).
  - Segundos: `%S` (00–61)
  - Número de día de la semana: `%w`: (0–6, domingo=0)
  - Número de semana: `%U,%W`: (00–53) (según la semana empiece en domingo=1 o lunes=1)

Un ejemplo sencillo de lectura/escritura de fechas se muestra a continuación:

```
Sys.getlocale("LC_TIME")  # Idioma local para tiempo

> [1] "Spanish_Spain.utf8"

mybirth = "27/03/1967 13:45:00"
myb = strptime(mybirth, "%d/%m/%Y %H:%M:%S")  #lectura
format(myb, "%A, %d-%B-%Y, Day:%j - Week:%W")  #escritura

> [1] "lunes, 27-marzo-1967, Day:086 - Week:13"

Sys.setlocale("LC_TIME", "pt_BR.UTF-8")  # cambio a brasileiro

> [1] "pt_BR.UTF-8"

format(myb, "%A, %d-%B-%Y, Day:%j - Week:%W")  #escritura

> [1] "segunda-feira, 27-março-1967, Day:086 - Week:13"
```

```
invisible(Sys.setlocale("LC_TIME", "es_ES.UTF-8"))  # cambio a español
x = Sys.Date()  # Fecha del sistema
difftime(x, myb)

> Time difference of 20560.47 days

dates = c("2012/05-01", "2012/06-15", "2012/07-01")  # Fechas en formato no estándar
datesread = as.Date(dates, format = "%Y/%m-%d")  # Lectura
datesread  # Escritura

> [1] "2012-05-01" "2012-06-15" "2012-07-01"
```

Una de las opciones más usadas con datos de tiempo es la posibilidad de crear secuencias regulares de fechas o calcular diferencias entre dos fechas determinadas. En el código siguiente se muestra un ejemplo de estas dos tareas usando en el primer caso el comando `seq` y en el segundo, el comando `difftime`.

```
myb1 = as.Date(mybirth, format = "%d/%m/%Y", tz = "CET")
myb2 = as.POSIXlt(myb, format = "%d/%m/%Y %H:%M", tz = "CET")
today = as.POSIXlt(Sys.Date(), tz = "CET")
difftime(today, myb2, units = "weeks")

> Time difference of 2937.198 weeks

x1 = seq(myb1, as.Date(today), by = "1 year")  #Secuencia de valores
x2 = seq(myb2, today, by = "1 week")  # Secuencia por semanas
head(x1, 4)

> [1] "1967-03-27" "1968-03-27" "1969-03-27" "1970-03-27"

head(x2, 4)

> [1] "1967-03-27 13:45:00 CET" "1967-04-03 13:45:00 CET"
> [3] "1967-04-10 13:45:00 CET" "1967-04-17 13:45:00 CET"
```

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

4.2.2 Otros paquetes

El paquete `lubridate` ((SPINU, 2023)) destaca por añadir funciones simples para manipular fechas, duraciones e intervalos. No establece nuevas clases para la fecha/hora sino que usa las existentes predefinidas en R. En el código siguiente se muestran ejemplos de los comandos más distintivos del paquete asociados a redondeos de fechas, la creación de intervalos temporales y duraciones. Véase específicamente el uso de `%within%` para determinar si una fecha está en un intervalo temporal concreto.

```
library(lubridate)
c(round_date(myb, "quarters"), floor_date(myb, "quarters"), ceiling_date(myb,
    "quarters"))

> [1] "1967-04-01 CET" "1967-01-01 CET" "1967-04-01 CET"

paste0(myb1, " fue ", wday(myb2, label = TRUE, abbr = FALSE),
    ". Dia del año:", qday(myb1))

> [1] "1967-03-27 fue lunes. Dia del año:86"

xint = interval(myb2, today)
as.duration(xint)

> [1] "1776417300s (~56.29 years)"

(myb1 + 65 * 365.25) %within% xint

> [1] FALSE
```

Otro paquete para manipulación de fechas/horas es `timeDate` ((BOSHNAKOV, 2023a)). Este paquete forma parte del conjunto `Rmetrics` (`Rmetrics.org`) que integra varios paquetes dedicados a distintos tópicos del análisis financiero (predicción, valoración de activos, optimización de carteras,...). El paquete define la clase `timeDate` que es del tipo `S4` e incluye tres slots: `@Data (POSIXct)`,

`@format` y `@FinCenter`. Entre las funciones que incorpora el paquete para manipulación de fechas y tiempos se destacan las siguientes que no son habituales en otras distribuciones:

- Información: `isWeekday, isWeekend, isBizday, isHoliday`
- Alineamiento: `time{First|Last}Dayin{Month|Quarter}`, `timeNdayOnOr{After|Before}`
- Selección: `start, end, length, window`
- Reordenación: `cut, sort, sample, unique, rev`
- Tipo de frecuencia: `isDaily, isMonthly, isQuarterly, isRegular`
- Muchas funciones para calcular días festivos: `EasterSunday(2013)`, `Easter(2012:2016)`, `Ascension(2013)`, véase `?holidayDate`.

```
library(timeDate)
Dates <- c("1989-09-28", "2001-01-15")
Times <- c("23:12:55", "10:34:02")
tDates = timeDate(paste(Dates, Times), zone = "GMT")
tDates@Data

> [1] "1989-09-28 23:12:55 GMT" "2001-01-15 10:34:02 GMT"

tSeq = timeSequence(from = tDates[1], to = tDates[2], by = "2years")
# by= sec, min, hour, day, week, month, year (--'5mins'--)
dayOfWeek(tSeq[1:3])

> 1989-09-28 23:12:55 1991-09-28 23:12:55 1993-09-28 23:12:55
>               "Thu"               "Sat"               "Tue"

# Vacaciones ?holidayDate
c(EasterSunday(2023), Ascension(2023))

> GMT
> [1] [2023-04-09] [2023-05-18]

timeLastDayInQuarter("2023-05-13")

> GMT
> [1] [2023-06-30]
```

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

4.3 SERIES DE TIEMPO

Antes de iniciar el análisis de una serie de tiempo, es importante ver como podemos importarlas en una sesión de R. Importar una serie de tiempo significa guardar sus valores numéricos en un vector o una matriz si la serie es multivariante e importar sus marcas de tiempo en un objeto Fecha/Hora. Por supuesto, todas las opciones de leer ficheros pueden usarse aquí si las series de tiempo vienen en un formato de archivo tipo texto (usando, por ejemplo, `read.table`) o para el que exista un paquete capaz leer el formato (tipo hoja de cálculo o similar). En el primer caso, la información de la fecha se leerá en formato `character` y se transformará con las funciones vistas en la sección anterior. En el segundo caso, se debe tener cuidado con el formato de la fecha que almacena el programa original y como se lee desde R. Por ejemplo, el formato Excel `.xlsx` puede leerse con varios paquetes entre los que están `xlsx` y `openxlsx`. Ambas librerías tienen una función con el nombre `read.xlsx` que permite leer este tipo de ficheros. Si intentamos leer un fichero con una columna tipo Fecha simple usando `openxlsx` debemos usar el parámetro `detectDates=TRUE` para asegurarnos de que la columna correspondiente del `data.frame` es de la clase `Date`. Esto no es necesario con el paquete `xlsx` que es capaz de detectar bien el formato de la columna sin indicarle ningún parámetro incluso si la columna tiene un formato largo (`dd-mm-yyyy HH:MM:SS`) usando una clase `POSIXct`. Por el contrario, el paquete `openxlsx` importa fecha/horas con formato largo como el número de días desde 01/01/1900 (el valor por defecto en Excel). En este punto, parece que `xlsx` siempre es mejor que `openxlsx` pero el primero es una API de Excel programada en Java y por tanto su funcionamiento depende de esta componente y de si la arquitectura de la máquina donde se ejecute la tiene instalada. `openxlsx` no tiene esta limitación y se puede ejecutar en cualquier arquitectura. Para solventar la dificultad con la importación de fechas, el paquete `openxlsx` dispone de las funciones `convertToDate` y `convertToDateTime` que ayudan con este tipo de

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

conversiones.

4.3.1 Importando directamente de Internet

R dispone de varias funciones que permiten leer directamente la información de páginas web y en el caso de algunos proveedores pueden proporcionar directamente los datos construyendo la dirección url apropiada donde se especifique el valor y los periodos que se desean. Esta aproximación es simple pero requiere de una cierta investigación ya que los proveedores cambian el formato de la dirección url de cuando en cuando.

A continuación se muestra un ejemplo de importación directa desde Yahoo del valor del BitCoin en dólares americanos. La dirección se construye con el prefijo `pre<-"https://query1. finance.yahoo.com/v7/finance/download/"` al que se le concatena la parte con los parámetros de la consulta: símbolo, inicio, final e intervalo `url<-paste0(pre,"BTC-USD?period1=aaa & period2=bbb & interval=1d".`

- símbolo: El símbolo debe ir `URLencode` por si algún carácter es incompatible con la dirección url.
- inicio, final: Fechas de inicio y final en segundos desde 1/1/1970.
- intervalo: Frecuencia de la medida.

  El código de manera simplificada sería de la siguiente forma:

```
symbol = "BTC-USD"  # Bitcoin vs USD
prelim = "https://query1.finance.yahoo.com/v7/finance/download/"
ini = as.Date("01/01/2009", "%d/%m/%Y")
nini = as.numeric(ini) * 60 * 60 * 24  # in seconds
fin = Sys.Date() - 1
nfin = as.numeric(fin) * 60 * 60 * 24  #in seconds
url = paste(prelim, URLencode(symbol, reserved = TRUE), "?period1=",
    nini, "&period2=", nfin, "&interval=1d", sep = "")
serie <- read.table(url, header = TRUE, sep = ",")
serie <- serie[order(serie[, "Date"]), ]
```

Una vez leída la serie con intervalo diario, se puede usar la función `aggregate` para cambiar la resolución temporal.

```
serie$ym = strftime(serie$Date, format = "%Y-%m")
seriem = aggregate(serie$Close, by = list(Date = serie$ym), FUN = mean)
colnames(seriem) = c("Date", "mClose")
tail(seriem)

>        Date   mClose
> 102 2023-02 23304.54
> 103 2023-03 25116.90
> 104 2023-04 28857.57
> 105 2023-05 27499.31
> 106 2023-06 27763.20
> 107 2023-07 30491.80
```

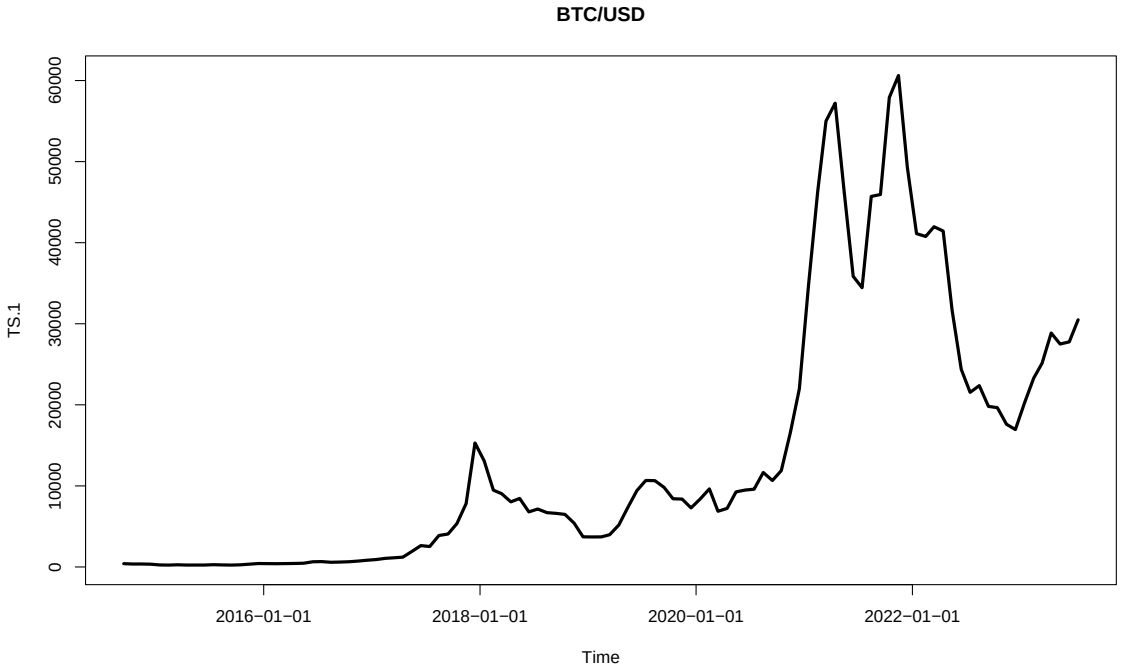
```
library(timeSeries)
plot(timeSeries(seriem[, "mClose"], paste0(seriem$Date, "-15")),
    main = "BTC/USD", lwd = 3)
```

4.3.2 Paquetes para importar datos

El paquete `quantmod` permite extraer información financiera de distintas fuentes: Yahoo, OANDA, FRED (Federal Reserve Economic Data) o mercados de cambio o de metales. El paquete también incluye Análisis Técnico (véase `?TA`) que es un conjunto de herramientas gráficas muy populares entre los analistas financieros.

```
library(quantmod)
getSymbols("GOOG", src = "yahoo")  # Se crea objeto GOOG en el entorno de trabajo
```

Figura 4.1: Promedio mensual de la cotización del BitCoin.



Fonte: O autor.

```
> [1] "GOOG"

# \tgetQuote('GOOG')
getSplits("GOOG")

>             GOOG.spl
> 2014-03-27 0.4995005
> 2015-04-27 0.9972620
> 2022-07-18 0.0500000

getMetals("XAU", from = Sys.Date() - 5 * 365)  # Gold objecto XAUUSD

> character(0)

getFX("EUR/USD", from = Sys.Date() - 4 * 365, to = Sys.Date() -
    1)  # EUR/USD últimos 4 años

> character(0)
```

De algunas fuentes no es posible obtener el período completo solicitado si este

excede los límites que el proveedor tiene implementados. El siguiente ejemplo de código plantea como se podría obtener la serie completa si el período máximo fuese de 5 años.

```
ini = as.Date("01/01/2008", "%d/%m/%Y")
fin = Sys.Date()
r = c(seq.Date(ini, fin, by = "5 year"), fin)
getSymbols("GOOG", from = r[1], to = r[2] - 1)

> [1] "GOOG"

goog = GOOG
for (i in 2:(length(r) - 1)) {
    getSymbols("GOOG", from = r[i], to = r[i + 1] - 1)
    goog = rbind(goog, GOOG)
}
# Serie mensual
googm = as.data.frame(aggregate(goog, by = list(strftime(index(goog),
    format = "%Y-%m")), FUN = mean))
colnames(googm) = c("mOpen", "mHigh", "mLow", "mClose", "mVolume",
    "mAdjusted")
```

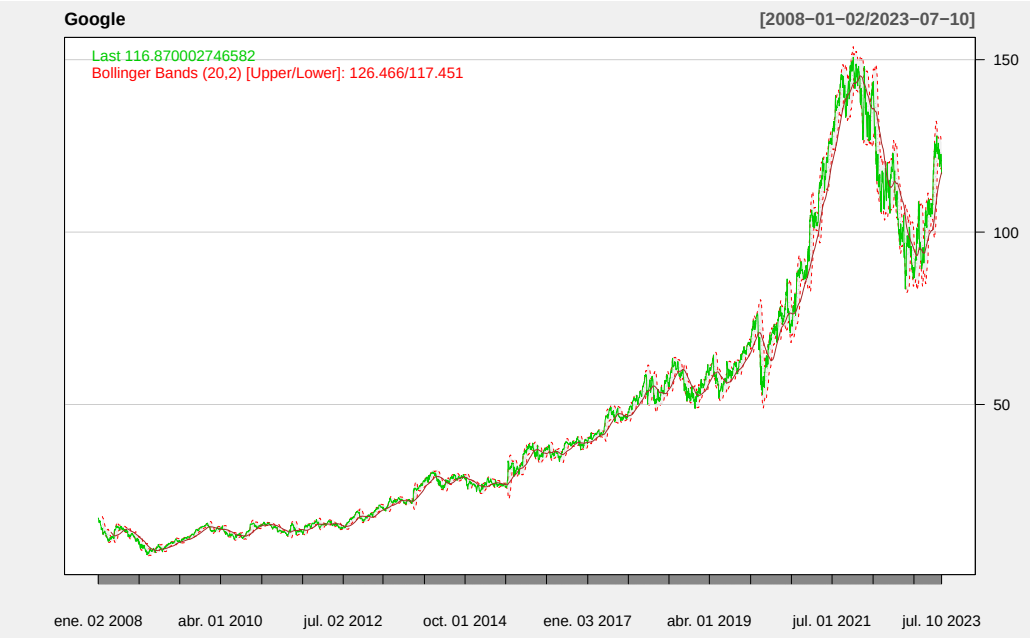
```
chartSeries(goog$GOOG.Adjusted, name = "Google", TA = c(addBBands(),
    addSMA(n = 60)), theme = chartTheme("white"))
```

4.3.3 Clases para series de tiempo

La clase más simple para almacenar una serie de tiempo regular es `ts`. Este formato es suficiente para las series macroeconómicas usuales (frecuencia mensual o mayor).

```
googm.ts = ts(data = googm[, "mAdjusted"], frequency = 12, start = c(2008,
    1))
# Datos mensuales desde enero 2008.
```

Figura 4.2: Cierre diario de GOOG con bandas de Bollinger y ajuste por medias móviles.



Fonte: O autor.

Los datos se proporcionan en un vector o matriz y la información temporal se construye sistemáticamente a partir del comienzo de la serie (`start`) y de su frecuencia (`frequency`).

Utilidades asociadas:

- `tsp(x)`: Inicio, final y frecuencia.
- `time(x)`: Fecha/Hora.
- `cycle(x)`: Posición en el ciclo.
- `frequency(x)`: Número de elementos por ciclo.
- `window(x,start=c(2012,3),end=c(2014,4))`: Subintervalo de la serie.
- `cbind.ts, ts.union, ts.intersect`: Funciones para unir o intersecar series de tiempo.
- `embed(x,k)`: Matriz de retardos dim. k: $\{X_t, X_{t-1}, \ldots, X_{t-k+1}\}$.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

- `filter(x,filter=c(1,1,1)/3)`: Filtro (media móvil de X_{t-1}, X_t, X_{t+1}).
- `filter(x,c(1,1,1), method="recursive")`: Filtro recursivo. $Y_t = X_t + f_1 X_{t-1} + \ldots + f_k X_{t-k}$.
- Otras herramientas descriptivas: `lag.plot, monthplot, stl, HoltWinters`

La clase `xts` (clase heredada del paquete `zoo`) proporciona una representación un poco más compleja de series de tiempo regulares así como comandos exploratorios y de conversión.

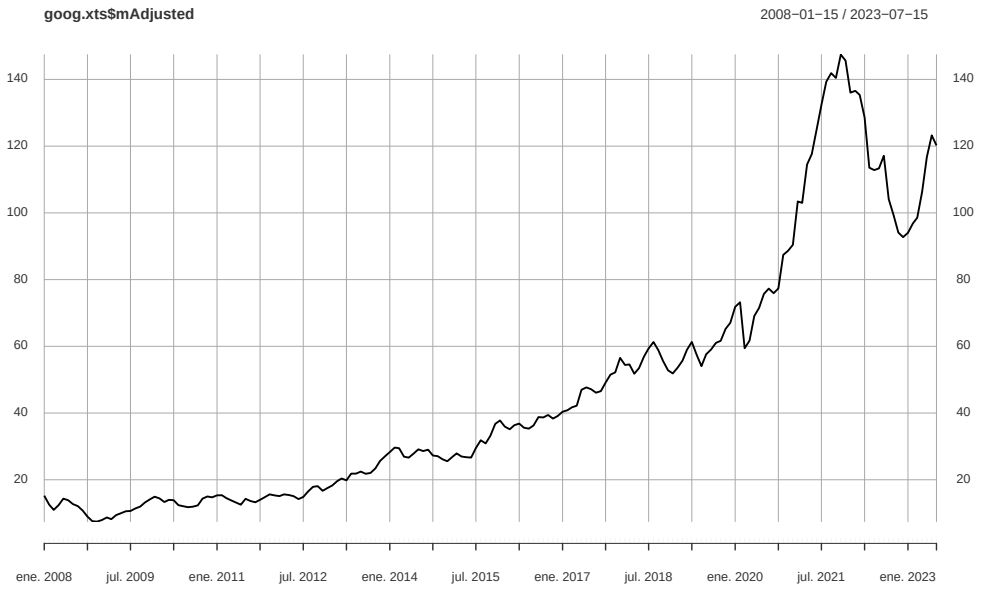
- Utilidades: `start`, `end`, `align`, `endpoints`, `split`, `subset`, `.index{day|mday|hour|wday|week|ymon|...}`
- Herramientas exploratorias: `period.apply`, `apply.{daily|weekly|monthly|quarterly|yearly}`
- Conversión: `to.period`, `to.{minutes,hourly,daily|weekly|monthly|quarterly|yearly}`

```
library(xts)
goog.xts = xts(googm, order.by = strptime(paste0(rownames(googm),
    "-15"), "%Y-%m-%d"))
plot(goog.xts$mAdjusted)
head(round(to.quarterly(goog.xts), 2), 3)  # Cambio a trimestres

>          goog.xts.Open goog.xts.High goog.xts.Low
> 2008 Q1         15.34         15.58        10.77
> 2008 Q2         12.37         14.55        12.19
> 2008 Q3         12.74         12.91        10.58
>          goog.xts.Close goog.xts.Volume goog.xts.Adjusted
> 2008 Q1          10.97       949699653             10.97
> 2008 Q2          13.86       640410630             13.86
> 2008 Q3          10.76       577867842             10.76

head(round(apply.quarterly(goog.xts[, "mAdjusted"], max), 2),
    3)  #Máximo por trimestres
```

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

```
>            mAdjusted
> 2008-03-15     15.24
> 2008-06-15     14.34
> 2008-09-15     12.70
```

Figura 4.3: Promedio mensual cierre de GOOG (formato xts).



Fonte: O autor.

La clase más compleja de series de tiempo es la `timeSeries` ((BOSHNAKOV, 2023b)) que, sobre todo, proporciona soporte ampliado para series de tiempo con frecuencia inferior a la diaria. Esta clase es la usada por los paquetes del conjunto `Rmetrics` como por ejemplo `fBasics` o `fArma`. La información del tiempo es guardada en formato `timeDate`. Esta clase es de tipo S4 y por tanto, el objeto dispone de *slots* a los que se accede con `@`.

- Slots: `positions`, `.Data`, `format`, `unit`, `title`, `FinCenter`
- Utilidades: `durations, returns, spreads, scale`
- Análisis exploratorio: `colCum{maxs|mins|prods|returns|sums}`, `colStats`, `rowStats`, `rollStats`, `ranks`, `runlengths`

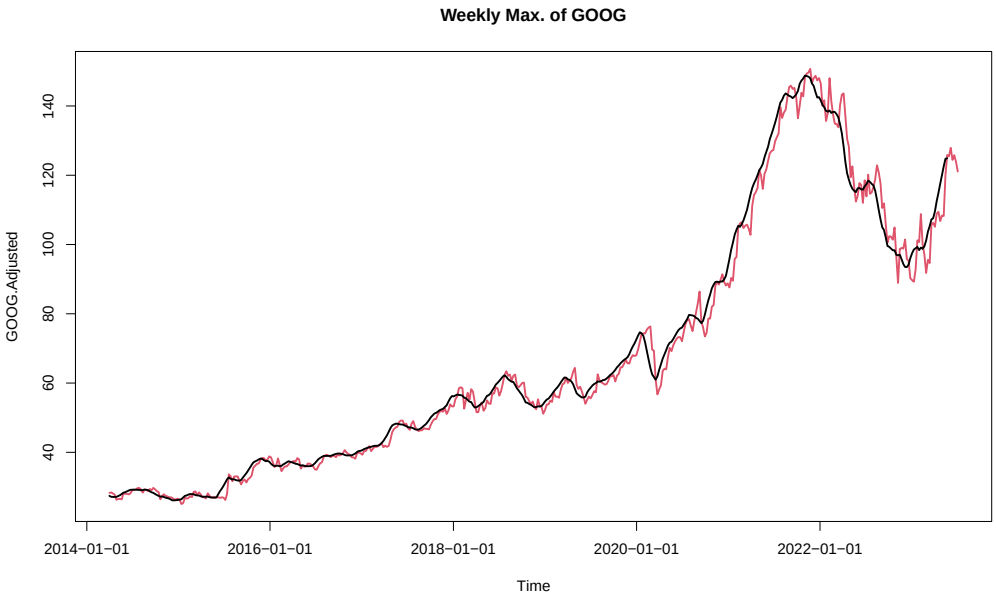
APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

```
library(timeSeries)
googc = timeSeries(goog$GOOG.Adjusted, charvec = index(goog))
head(googc@.Data, 3)

>      GOOG.Adjusted
> [1,]      17.06578
> [2,]      17.06927
> [3,]      16.36366

head(returns(googc), 3)

> GMT
>             GOOG.Adjusted
> 2008-01-03  0.0002042846
> 2008-01-04 -0.0422164175
> 2008-01-07 -0.0118662305

googd = window(googc, "2014-04-01", end(googc))  #dato diario desde 2014/04
by = timeSequence(start(googd), end(googd), by = "week")
googw = aggregate(googd, by, FUN = max)  # Max. de datos semanales
head(series(googw), 3)

>             GOOG.Adjusted
> 2014-04-01      28.28036
> 2014-04-08      28.40900
> 2014-04-15      28.12977

plot(googw, lwd = 2, col = 2, main = "Weekly Max. of GOOG")
lines(rollStats(googw, 7, FUN = mean), lwd = 2)  # Media móvil de 7 semanas
```

```
goog.tS = timeSeries(goog[, 1:5], index(goog))
plot(tail(goog.tS[, "GOOG.Open"], 520), plot.type = "single",
    main = "Precio apertura GOOG últimos dos años.")  #Últimos 2 años
```

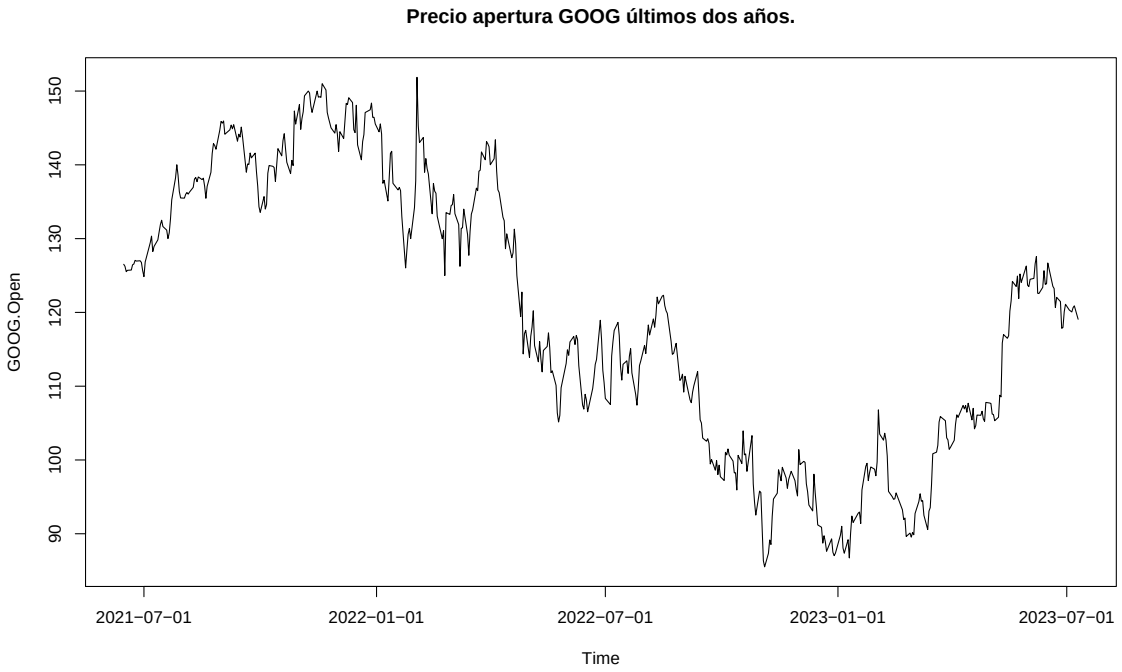
APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 4.4: Gráfico de Google. Máximos semanales y media móvil de 7 semanas.



Fonte: O autor.

Figura 4.5: Últimos dos años del precio de apertura de GOOG. Datos diarios.



Fonte: O autor.

4.4 MODELIZACIÓN CLÁSICA DE UNA SERIE DE TIEMPO

Una vez importados en R los datos, se trata ahora de ver como ajustar un modelo matemático para poder hacer inferencias sobre el proceso generador de la serie temporal y, por ejemplo, predecir valores futuros. Lo primero que hay que tener en cuenta es que para poder hacer inferencias sólo con una trayectoria de la serie temporal, debemos suponer ciertas hipótesis de regularidad de la serie a lo largo del tiempo. Estas hipótesis se suelen resumir en pedir que la serie temporal sea estacionaria en sentido débil o de segundo orden (las medidas estadísticas hasta orden 2 son estacionarias). Un texto clásico donde consultar los pasos a seguir es (WEI, 2006) o (TSAY, 2010) para el entorno más financiero. (CRYER; CHAN, 2008) o (SHUMWAY; STOFFER, 2017) son referencias con abundantes ejemplos en R.

Específicamente, diremos que una serie Z_t es estacionaria de segundo orden si:

$$\begin{aligned} \mathbb{E}(Z_t) &= \mu \quad \forall t \\ \operatorname{Cov}(Z_t, Z_{t-j}) &= \mathbb{E}((Z_t-\mu)(Z_{t-j}-\mu)) = \gamma_j \quad \forall t, j \\ \rho_j &= \frac{\operatorname{Cov}(Z_t, Z_{t-j})}{\sqrt{\operatorname{Var}(Z_t)\operatorname{Var}(Z_{t-j})}} = \frac{\gamma_j}{\gamma_0} \end{aligned}$$

Estas condiciones imponen que la media y varianza sean constantes y que la covarianza sólo dependa del retardo entre dos instantes que, claramente, no es una condición que cumplan de partida las series de tiempo obtenidas del mundo real (véase, por ejemplo, la serie de Google en apartados anteriores). Por eso, el análisis clásico parte de la idea de que una serie de tiempo X_t se descompone en:

$$X_t = T_t + S_t + Z_t$$

donde T_t es la tendencia a largo plazo y que supondremos determinística, S_t es la componente estacional (también determinística) y Z_t es el residuo, la parte

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

irregular o el efecto aleatorio que tendrá carácter estocástico y es la parte de la serie temporal que supondremos

Por tanto, el primer objetivo en un análisis clásico es identificar (o eliminar) las componentes determinísticas (T_t y S_t) y posiblemente transformar la serie original para obtener de forma limpia la serie Z_t con las propiedades mencionadas antes. Los posibles pasos previos que deben seguirse para obtener Z_t se resumen en los apartados siguientes:

- Ajustar tendencia con un modelo de regresión de forma conocida:

$$\hat{T}_t = \alpha_0 + \alpha_1 t + \ldots + \alpha_k t^k$$

$$\hat{T}_t = \nu_0 + \sum_{j=1}^{m} \left(\alpha_j \cos \omega_j t + \beta_j \operatorname{sen} \omega_j t\right), \omega_j = 2\pi/p_j$$

- Ajustar tendencia por medias móviles:

$$\hat{T}_t = \sum_{j=-m}^{m} \alpha_j X_{t-j}$$

  Para evitar interacción con la parte estacional, m debe ser múltiplo del ciclo de la serie.

- Aplicar diferencias: Este es un procedimiento para librarse de la tendencia más que para identificarla. Es bien conocido que el operador $\nabla^d X_t = \nabla^{d-1}(X_t - X_{t-1})$ puede eliminar tendencias polinómicas hasta grado d.

- Estacionalidad: Para identificar la estacionalidad podemos, como en el caso de T_t, plantear modelos de regresión usando como covariable el momento del ciclo.

- Estabilización de la varianza: Si la serie X_t no es estacionaria por la varianza, podemos convertirla a una serie estacionaria si la varianza es función de la

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

media, i.e.

$$\text{Var}\,(X_t) = f(\mu_t) \implies T(X_t) = \int \frac{1}{\sqrt{f(\mu_t)}} d\mu_t$$

Para identificar esta situación lo habitual es calcular las medias y varianzas por períodos (con ciclo completos) y buscar si existe una relación entre ambas. Si esto es posible, podemos aplicar una transformación Box–Cox que estabilice la varianza. La transformación más habitual cuando la varianza es una función cuadrática de la media es la logarítmica.

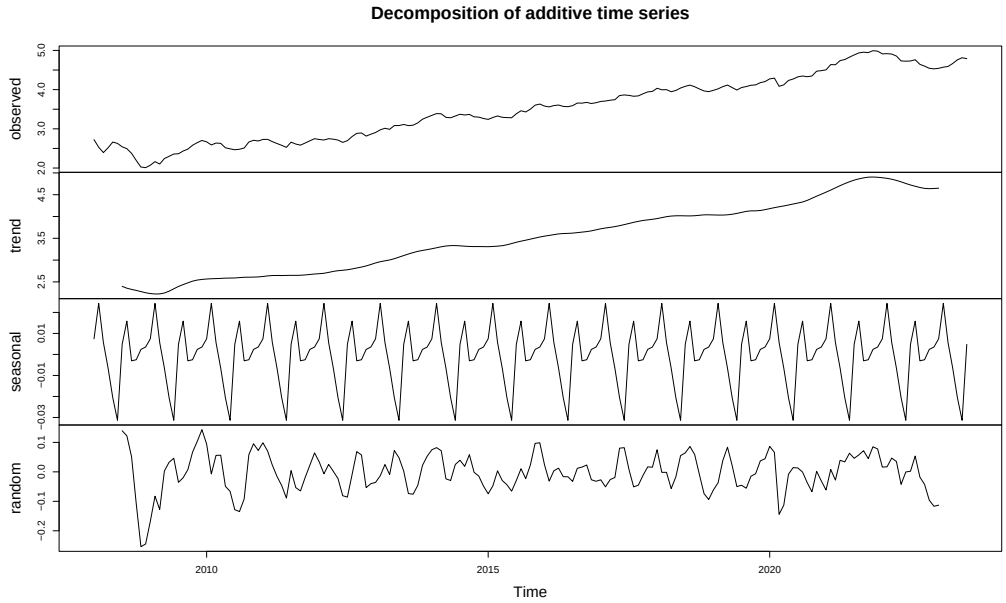
El comando `decompose` ofrece una herramienta gráfica que identifica cada una de las componentes clásicas de una serie temporal.

```
lgoog.ts = ts(log(googm[, "mAdjusted"]), frequency = 12, start = c(2008,
    1))
plot(decompose(lgoog.ts))
```

La estimación de tendencia se puede hacer manualmente usando las herramientas de regresión disponibles en R. En el ejemplo que se muestra a continuación, el uso del paquete `nlme` nos permite estimar el modelo de regresión con algún tipo de dependencia temporal (`corAR1`). Además en el ejemplo se ha incluido una variable binaria para determinar si hay un cambio en la media antes y después de una fecha determinada. La estimación de la tendencia por medias móviles se ha realizado en el siguiente ejemplo usando el filtro de Spencer de 15 puntos que es una regla clásica para este fin.

```
library(nlme)
t = index(lgoog.ts)
dummy = ifelse(time(lgoog.ts) < 2014.25, 0, 1)
x.gls = gls(lgoog.ts ~ poly(t, 3) + dummy, correlation = corAR1())
summary(x.gls)
```

Figura 4.6: Descomposición clásica de la serie log($GOOG$).



Fonte: O autor.

```
....
> Correlation Structure: AR(1)
>  Formula: ~1
>  Parameter estimate(s):
>       Phi
> 0.9999926
>
> Coefficients:
>                 Value Std.Error   t-value p-value
> (Intercept)  3.577902 16.023185  0.223295  0.8236
> poly(t, 3)1 11.566639  3.650886  3.168173  0.0018
> poly(t, 3)2  1.448743  1.496320  0.968204  0.3342
> poly(t, 3)3 -1.974702  0.987046 -2.000619  0.0469
> dummy       -0.108877  0.061816 -1.761302  0.0799
....

 #lgoog.ts=lgoog.ts-dummy*z.gls$coefficients[['dummy']]
Sp.15 = filter(lgoog.ts, c(-3, -6, -5, 3, 21, 46, 67, 74, 67,
    46, 21, 3, -5, -6, -3)/320)  #Spencer's 15 point
```

Para determinar la estacionalidad, se aplica un modelo de regresión a los

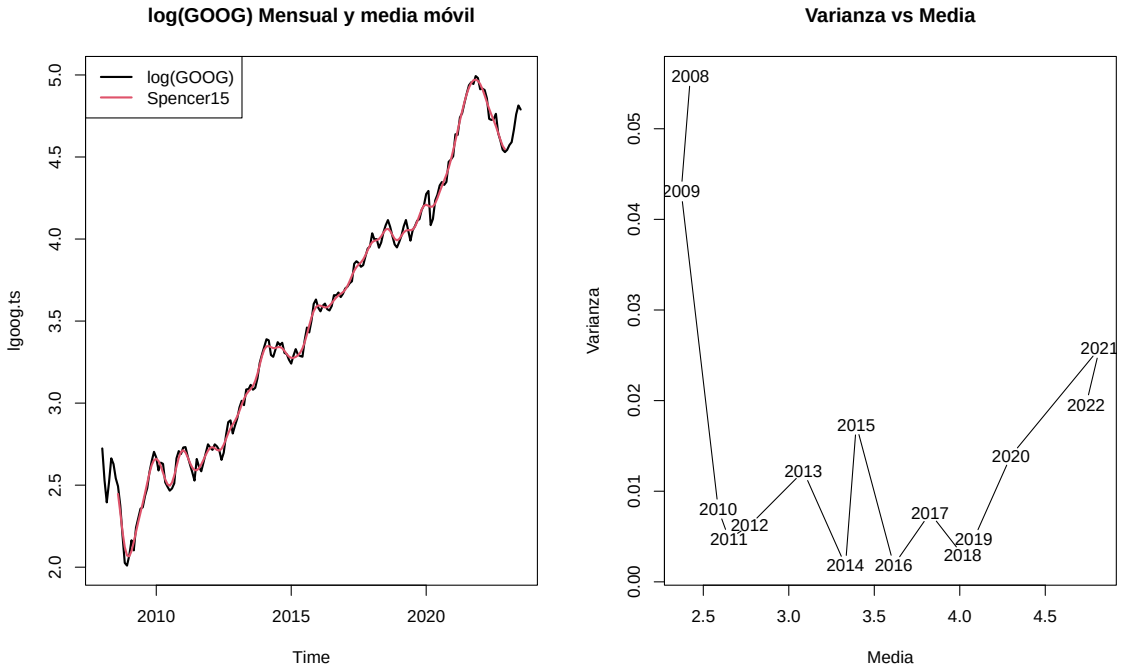
residuos del modelo de tendencia donde no se estima el parámetro ordenada en el origen y se incluye como covariable el momento del ciclo de la serie. El resultado muestra una estacionalidad muy débil ya que sólo dos puntos del ciclo (septiembre y octubre) son significativos y no por mucho.

```
x.seas = lm(residuals(x.gls) ~ -1 + as.factor(cycle(lgoog.ts)))
m = aggregate(lgoog.ts, nfrequency = 1, FUN = mean)
v = aggregate(lgoog.ts, nfrequency = 1, FUN = var)

....
> Coefficients:
>                              Estimate Std. Error t value
> as.factor(cycle(lgoog.ts))1   0.006723   0.046626   0.144
> as.factor(cycle(lgoog.ts))2   0.005021   0.046626   0.108
> as.factor(cycle(lgoog.ts))3  -0.017703   0.046626  -0.380
> as.factor(cycle(lgoog.ts))4  -0.008031   0.046626  -0.172
> as.factor(cycle(lgoog.ts))5  -0.003672   0.046626  -0.079
> as.factor(cycle(lgoog.ts))6  -0.009923   0.046626  -0.213
> as.factor(cycle(lgoog.ts))7   0.007629   0.046626   0.164
> as.factor(cycle(lgoog.ts))8   0.017777   0.048155   0.369
> as.factor(cycle(lgoog.ts))9  -0.001592   0.048155  -0.033
> as.factor(cycle(lgoog.ts))10 -0.001343   0.048155  -0.028
> as.factor(cycle(lgoog.ts))11  0.002945   0.048155   0.061
> as.factor(cycle(lgoog.ts))12  0.003499   0.048155   0.073
....
```

En el trozo de código anterior se calculan la media y varianza por ciclo anual. El dibujo de estas cantidades (véase figura 4.7) puede dar una idea de si es necesaria una transformación.

```
plot(lgoog.ts, lwd = 2, main = "log(GOOG) Mensual y media móvil")
lines(Sp.15, col = 2, lwd = 2)
plot(m, v, xlab = "Media", ylab = "Varianza", type = "p", main = "Varianza vs Media")
```

Figura 4.7: Serie log($GOOG$) y una estimación por media móvil (izda). Gráfico de media y varianza por años (dcha) de la serie transformada por logaritmos.



Fonte: O autor.

4.4.1 Modelización ARMA

La modelización ARMA consiste en representar la parte estacionaria de una serie como función lineal de su pasado (parte AR), como función lineal de un proceso de ruido blanco (parte MA) o de ambas. Un proceso de ruido blanco es una serie temporal ϵ_t que cumple que:

$$\mathbb{E}(\epsilon_t) = 0,\ \mathrm{Var}(\epsilon_t) = \sigma^2,\ \text{ y } \rho_j = 0,\ \forall j \neq 0$$

Un proceso de ruido blanco es una secuencia i.i.d. y no tiene dependencia temporal. Por tanto, encontrar un modelo ARMA que deje como residuo final un proceso de ruido blanco es decir que la tarea del modelizador se ha acabado porque no hay más patrones temporales que encontrar. Si llamamos B al operador *backward* i.e. $B^j(Z_t) = Z_{t-j}$ podemos escribir los siguientes modelos:

4.4.1.1 Modelo AR(p)

Un modelo AR(p) escribe el valor actual de una serie como función lineal de valores de su pasado hasta el retardo p más un proceso de ruido blanco. Por tanto, un modelo AR(p) se escribe:

$$Z_t = c + \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + \epsilon_t$$

En términos del operador *backward* podemos escribir el modelo como un polinomio de este operador con la ventaja de que ciertas propiedades del polinomio se corresponden con propiedades relevantes de la serie temporal. A este polinomio se le llama polinomio característico de la parte AR.

$$\Phi_p(B)Z_t = c + \epsilon_t \text{ con } \Phi_p(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$$

En particular, si las raíces del polinomio $\Phi_p(B)$ caen fuera del círculo unidad podemos afirmar que la serie es estacionaria. El parámetro c está relacionado con la media de la serie, de manera que, $c = \frac{1}{\Phi(B)\mu}$ en una serie estacionaria. Por tanto, sin pérdida de generalidad, podríamos considerar sólo series con media cero. La ventaja de usar un modelo AR(p) radica en que la predicción de valores futuros se puede construir recursivamente sólo con los datos de la serie (que son observables). Así la predicción $\hat{Z}_{n+h}$ se calcula usando exactamente la ecuación AR con los parámetros estimados, esto es,

$$\hat{Z}_{n+h} = \hat{c} + \hat{\phi}_1 \hat{Z}_{n+h-1} + \ldots + \hat{\phi}_p \hat{Z}_{n+h-p}$$

donde $\hat{Z}_{n+h-j} = Z_{n+h-j}$ si $n+h-j \leq n$ y por tanto puede ser calculada fácilmente de forma iterativa en secuencia $\hat{Z}_{n+1}, \ldots, \hat{Z}_{n+h}$.

4.4.1.2 Modelo MA(q)

Un modelo MA(q) escribe el valor actual de la serie como función lineal de los valores actuales y pasados (hasta retardo q) de un proceso de ruido blanco.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Por tanto, un modelo MA(q) se escribe:

$$Z_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \ldots + \theta_q\epsilon_{t-q}$$

De nuevo, en términos del operador *backward* el modelo se escribe como un polinomio de este operador (polinomio característico de la parte MA) de la siguiente forma:

$$Z_t = \mu + \Theta_q(B)\epsilon_t \text{ con } \Theta_q(B) = 1 + \theta_1 B + \ldots + \theta_q B^q$$

Un modelo MA(q) siempre es estacionario (por definición) y de hecho tener una representación MA es una caracterización de la estacionariedad, esto es, una serie temporal es estacionaria si y solo si se puede escribir como una representación MA sumable (quizás infinita). La utilidad de esta representación radica en la facilidad para calcular las varianzas de predicción de valores futuros. Se puede calcular fácilmente que

$$\text{Var}\left(\hat{Z}_{n+h}\right) = \sigma_\epsilon^2 \sum_{j=0}^{\min(h,q)} \theta_j^2$$

4.4.1.3 Modelo ARMA(p,q)

Un modelo ARMA(p,q) mezcla los dos tipos de modelos anteriores para obtener una escritura más sintética del modelo temporal:

$$Z_t = c + \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} + \ldots + \theta_p\epsilon_{t-q}$$

que, en términos de los polinomios AR y MA, se resume en:

$$\Phi_p(B)Z_t = c + \Theta_q(B)\epsilon_t$$

Como antes, un modelo ARMA(p,q) será estacionario si su polinomio AR es estacionario. La propiedad de que las raíces estén fuera del círculo unidad tiene

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

que ver con que al escribir el modelo ARMA como

$$Z_t = \frac{\Theta_q(B)}{\Phi_p(B)}\epsilon_t$$

el polinomio resultante en la parte MA tenga una representación sumable. Si alguna de las raíces fuera, en módulo, mayor que uno tendríamos un modelo AR explosivo que no se puede modelizar con las herramientas aquí descritas. Si alguna de las raíces tuviese módulo exactamente igual a uno, la solución sería aplicar una diferencia.

La condición equivalente a la estacionariedad con el polinomio MA $\Theta_q(B)$ (esto es que sus raíces estén fuera del círculo unidad) se conoce como *invertibilidad*. En este caso, la falta de invertibilidad puede corregirse cuando hay raíces mayores que la unidad ya que es posible encontrar una formulación equivalente invertible. Aquí no es posible corregir la falta de invertibilidad cuando hay raíces del polinomio MA en módulo exactamente igual a uno.

Las series temporales estacionarias e invertibles permiten cambiar de una representación ARMA a la AR o a la MA indistintamente lo que nos permite tanto calcular predicciones y varianzas de predicción de forma más sencilla. Por tanto, para un modelo ARMA estacionario e invertible, podemos reescribirlo de forma equivalente en un representación AR o MA (posiblemente de orden infinito): AR(∞)$\equiv$ARMA(p,q)$\equiv$MA(∞).

4.4.2 Identificación del modelo ARMA

Para identificar los órdenes de un modelo ARMA, se usan dos funciones: la función de autocorrelación simple (ACF) y la función de autocorrelación parcial (PACF). La idea es, básicamente, comparar estas funciones estimadas de la muestra $\{Z_t\}_{t=1}^n$ con las que debieran ser si el modelo ARMA(p,q) es cierto. La ACF, como se definió antes, es $\rho(j) = \text{Corr}\,(Z_t, Z_{t+j})$ que se va a estimar con $\hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_0}$ donde $\hat{\gamma}_j$ es la estimación de autocovarianza. La PACF es la correlación entre

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

(Z_t, Z_{t+j}) dada la información que está entre ambos instantes $(Z_{t+1}, \ldots, Z_{t+j-1})$ y se puede calcular resolviendo el siguiente sistema:

$$P_j = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{j-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{j-3} & \rho_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{j-1} & \rho_{j-2} & \rho_{j-3} & \cdots & \rho_1 & \rho_j \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{j-2} & \rho_{j-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{j-3} & \rho_{j-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{j-1} & \rho_{j-2} & \rho_{j-3} & \cdots & \rho_1 & 1 \end{vmatrix}}$$

Las estimaciones muestrales de las cantidades anteriores tienen medias y varianzas, en general, complicadas y sólo en ciertos casos se pueden interpretar. Por ejemplo, la varianza de la estimación de la ACF a retardo j se aproxima por la expresión:

$$\text{Var}\left(\hat{\rho}_j\right) \approx \frac{1}{n} \sum_{i=-\infty}^{\infty} \left(\rho_i^2 + \rho_{i+j}\rho_{i-j} - 4\rho_j\rho_i\rho_{i-j} + 2\rho_j^2\rho_i^2\right)$$

que usa un sumatorio infinito. Si suponemos que $\rho_j = 0, |j| > m$, entonces la anterior expresión se simplifica a $\text{Var}\left(\hat{\rho}_j\right) \approx \frac{1}{n}\left(1 + 2\rho_1^2 + \ldots + 2\rho_m^2\right)$ que es conocida como la aproximación de Bartlett. Si aún suponemos más y $\rho_j = 0, j \neq 0$ (proceso de ruido blanco) entonces $\text{Var}\left(\hat{\rho}_j\right) \approx \frac{1}{n}$. El caso de la PACF es mucho más difícil y sólo se conocen sus propiedades para un proceso de ruido blanco donde la varianza es la misma que en el caso de la ACF, i.e. $\text{Var}\left(\hat{P}_j\right) \approx \frac{1}{n}$

Volviendo a los modelos, a continuación se describen como son teóricamente la ACF y la PACF para los distintos tipos de modelos.

AR(p) – ACF: $\rho_k = \sum_{i=1}^{m} G_i^k \sum_{j=1}^{d_i-1} A_{ij}k^j$ con G_i^{-1} las raíces de $\Phi_p(B)$ y A_{ij} coeficientes distintos de cero. Decaimiento rápido de forma exponencial y/o sinusoidal.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

– PACF: $P_k \neq 0, k \leq p$, and $P_k = 0, k > p$

MA(q)
– ACF: $\rho_k \neq 0, k \leq q$, y $\rho_k = 0, k > q$
– PACF: Decaimiento rápido de forma exponencial y/o sinusoidal.

ARMA(p,q)
– ACF: Primeros q valores distintos de cero y luego se produce un decaimiento rápido de forma exponencial y/o sinusoidal.
– PACF: Primeros p valores distintos de cero seguido de un decaimiento rápido de forma exponencial y/o sinusoidal.

Hablaremos de un modelo ARIMA(p,d,q) cuando sea necesario diferenciar d veces para obtener un modelo ARMA(p,q). Por tanto, un modelo ARIMA(p,d,q) se escribirá de la siguiente forma resumida:

$$\Phi_p(B)(1-B)^d Z_t = c + \Theta_q(B)\epsilon_t$$

La extensión a modelos con componente estacional es sencilla. Para su identificación se usan las mismas herramientas (ACF y PACF) pero fijándonos en los retardos múltiplos del ciclo: $s, 2s, 3s, ...$ En términos de la escritura del modelo, hablaremos ahora del modelo ARIMA$(p,d,q) \times$ ARIMA$_s(P,D,Q)$ conocido como ARIMA estacional que se escribe como:

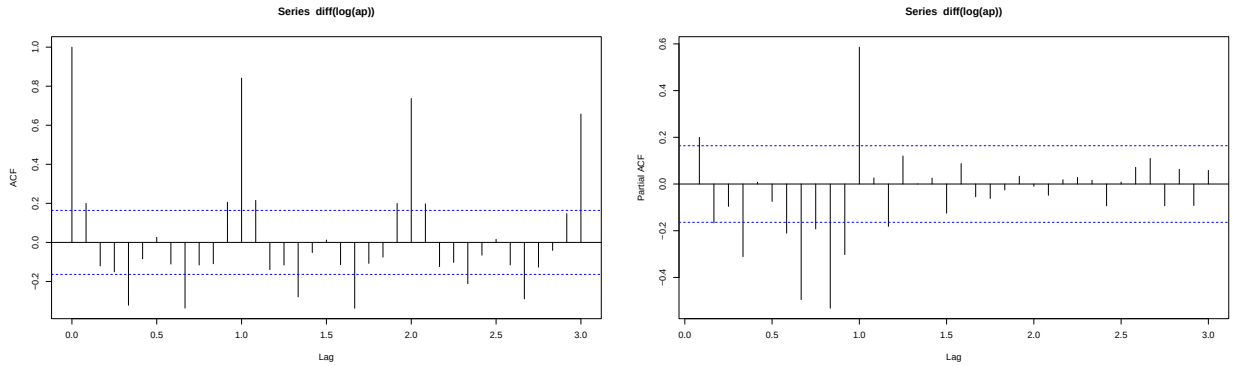
$$\Phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D z_t = c + \Theta_q(B)\Theta_Q(B^s)\epsilon_t$$

El modelo ARIMA estacional gana tres nuevos parámetros P, D y Q que representan el orden de la componente estacional. Un ejemplo de este tipo de modelos se muestra a continuación con una serie clásica de número de pasajeros de avión.

Como se puede ver en las funciones ACF y PACF de esta serie (véase figura 4.8), hay una marcada dependencia en los múltiplos del ciclo (12, 24 y 36) que en el gráfico se marcan como 1, 2 y 3.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 4.8: ACF y PACF de la serie *AirPassengers* transformada por logaritmos.



Fonte: O autor.

4.4.3 Identificación y estimación del modelo

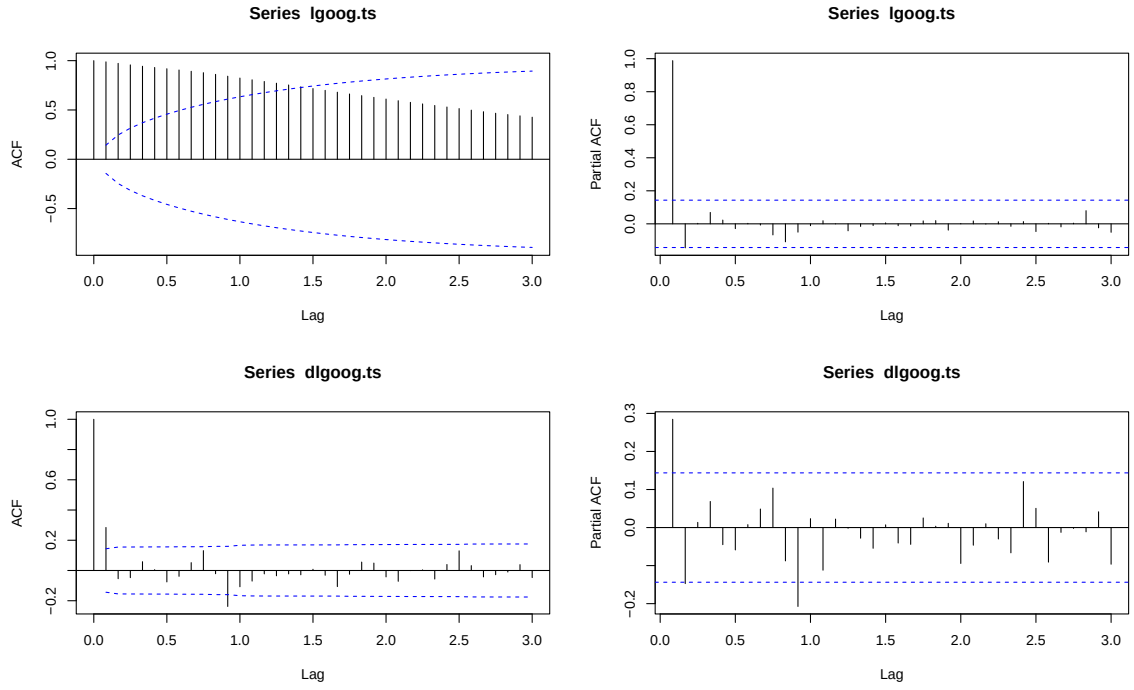
El objeto de este apartado es afrontar la modelización de una serie de tiempo paso a paso. Por supuesto, lo primero es siempre dibujar la serie para ver si se aprecian a simple vista ciertas características. A partir de la figura 4.7 ya se aprecia que la media no es constante y que la transformación por logaritmo ha estabilizado la varianza cambiante. Es por esto, que en el gráfico siguiente figura 4.9, se calcula la ACF (primera columna) y la PACF (segunda columna) para la serie $\log(GOOG)$ y su diferencia.

```
dlgoog.ts = diff(lgoog.ts)
acf(lgoog.ts, ci.type = "ma", lag.max = 36)
pacf(lgoog.ts, lag.max = 36)
acf(dlgoog.ts, ci.type = "ma", lag.max = 36)
pacf(dlgoog.ts, lag.max = 36)
```

En la primera fila se observa que las barras de la ACF decaen muy lentamente hacia cero y la PACF tiene la barra del primer retardo muy cercano a 1. Esto sugiere que la serie necesita ser diferenciada para conseguir estacionariedad. En la segunda fila se calcula la ACF y PACF de la serie diferenciada donde ya sólo se ve significación de $\hat{\rho}_1$ (la primera barra es siempre $\rho_0 = 1$) y de $\hat{P}_1$. También se sale del intervalo de confianza la barra del retardo 11. El intervalo de confianza en la ACF es calculado con la aproximación de Bartlett (`ci.type="ma"`). Si este

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 4.9: ACF y PACF de la serie log(*GOOG*) (primera fila) y de su diferencia (segunda fila).



Fonte: O autor.

argumento no se especifica se calculan los intervalos de confianza bajo el supuesto de que tenemos ruido blanco (`ci.type="white"`). Para la PACF se usa esta segunda opción que es la única disponible.

De las gráficas anteriores, y viendo que sólo una barra de la PACF se sale, podemos empezar pensando en un modelo AR simple (AR(1) o AR(2)). Para este caso, el comando `ar` permite estimar un AR(p) eligiendo de forma óptima el orden hasta un margen determinado.

```
res.ar = ar(dlgoog.ts, order.max = 8, method = "mle")   #meth=c('yw','burg','ols')
res.ar

> Call:
> ar(x = dlgoog.ts, order.max = 8, method = "mle")
> Coefficients:
>       1        2
>  0.3408  -0.1509
> Order selected 2  sigma^2 estimated as  0.003445
```

El procedimiento selecciona un AR(2) como el mejor. Ahora debemos comprobar las hipótesis estructurales. Si la modelización es buena, los residuos debieran comportarse como una secuencia de proceso de ruido blanco. Por otro lado, las raíces del polinomio característico de la parte AR debieran estar fuera del círculo unidad. Para lo primero vamos a emplear el comando `Box.test` que realiza un contraste de Ljung–Box para determinar si los primeros retardos (`lag`) pueden suponerse simultáneamente nulos. Para lo segundo calculamos las raíces del polinomio con el comando `polyroot`.

```
Box.test(res.ar$resid, lag = 24, "Ljung-Box")

>  Box-Ljung test
> data:  res.ar$resid
> X-squared = 22.32, df = 24, p-value = 0.5602

abs(polyroot(c(1, -res.ar$ar)))

> [1] 2.57458 2.57458
```

Ambas comprobaciones resultan satisfactorias y por tanto podríamos suponer que un modelo AR(2) es razonable. El comando `ar` tiene más bien una función exploratoria preliminar. Si queremos estimar un modelo ARIMA general el comando apropiado es `arima` que permite especificar el orden de modelo de forma simple. `arima` calcula, por defecto, los estimadores por máxima verosimilitud usando internamente una representación en espacio de estados (filtro de Kalman) aunque con el parámetro `method` se puede especificar alguna alternativa. A continuación, se muestra el ejemplo de modelizar un ARI(2) sobre la serie sin diferenciar. Al resultado de `arima` se le puede aplicar el comando `tsdiag` que proporciona diagnósticos gráficos de las hipótesis. El más interesante de estos diagnósticos (véase figura 4.10) es el gráfico del p-valor del test de Ljung–Box en función del re-

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

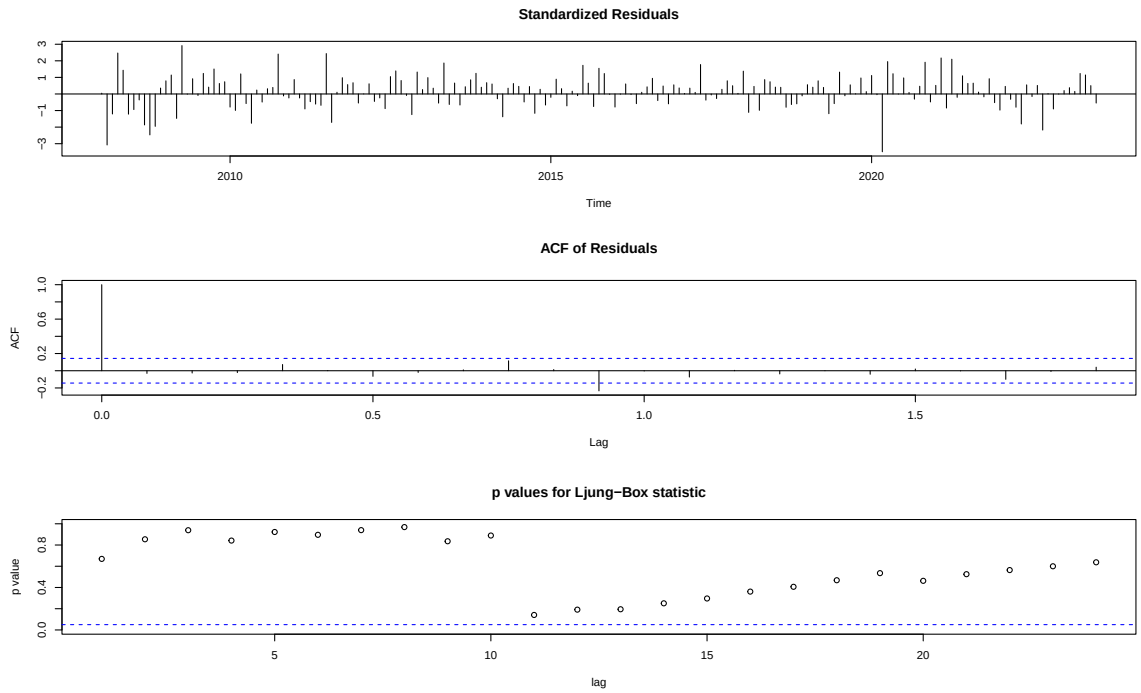
tardo (tercero) porque de ese gráfico y el anterior podemos determinar un modelo alternativo si apreciamos alguna deficiencia en la estimación.

```
res.arma = arima(lgoog.ts, order = c(2, 1, 0))
res.arma

>
> Call:
> arima(x = lgoog.ts, order = c(2, 1, 0))
>
> Coefficients:
>          ar1      ar2
>       0.3625  -0.1266
> s.e.  0.0740   0.0749
>
> sigma^2 estimated as 0.003518:  log likelihood = 261.45,  aic = -516.89

Box.test(res.arma$resid, lag = 24, "Ljung-Box")

>
>  Box-Ljung test
>
> data:  res.arma$resid
> X-squared = 21.026, df = 24, p-value = 0.6372

tsdiag(res.arma, gof.lag = 24)
```

También debemos analizar los parámetros obtenidos y su posible significación. En el ejemplo anterior, el valor del parámetro `ar2` estaría dentro de la zona de aceptación de la hipótesis $H_0 : \phi_2 = 0$ $(\pm 1.96\sigma(\hat{\phi}_2))$ bajo normalidad lo que podría sugerir que con un AR(1) es suficiente. Efectivamente, si se estima un AR(1) en el anterior ejemplo se pasan todos los diagnósticos de dependencia pero el modelo resultante tiene un criterio de Akaite (AIC) peor. Además, debe recordarse que la estimación de la varianza de los estimadores de los parámetros se hace a partir de la Hessiana y por tanto es una aproximación a la varianza verdadera. En caso de que necesitemos ajustar la parte estacional, esto se hace

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 4.10: Gráficos de diagnóstico para ARIMA(2,1,0)



Fonte: O autor.

añadiendo el parámetro `seasonal` en la forma en la que se ve el ejemplo siguiente donde se modeliza un $\text{ARIMA}(2,1,0) \times \text{ARIMA}_{12}(0,0,1)$.

```
res.arma2 = arima(lgoog.ts, order = c(2, 1, 0), seasonal = list(order = c(0,
    0, 1), period = 12))  #Peor AIC

>
> Call:
> arima(x = lgoog.ts, order = c(2, 1, 0), seasonal = list(order = c(0, 0, 1),
>     period = 12))
>
> Coefficients:
>          ar1      ar2    sma1
>       0.3695  -0.1315  0.0273
> s.e.  0.0771   0.0764  0.0854
>
> sigma^2 estimated as 0.003516:  log likelihood = 261.5,  aic = -515
>
>  Box-Ljung test
>
```

```
> data:  res.arma2$resid
> X-squared = 21.055, df = 24, p-value = 0.6355
```

Este último modelo también es aceptable desde el punto de vista de verificar las hipótesis estructurales pero comparando ambos vemos que este segundo presenta un AIC peor que el anterior. Además, la desviación estándar del parámetro `sma1` (el MA estacional) es mayor que la estimación del parámetro lo que sugiere que podemos aceptar muy claramente la hipótesis de que este parámetro puede ser cero. De nuevo, en este ejemplo, el parámetro `ar2` está al borde de la significación.

El paquete `forecast` ((HYNDMAN, 2023)) añade funciones y gráficos de diagnóstico y modelizaciones alternativas que se toman prestadas de otros paquetes (por ejemplo, ARFIMA). Los dos procedimientos más populares de este paquete son `auto.arima` y `forecast`. El primero selecciona el modelo ARIMA óptimo entre un rango amplio cumpliendo los diagnósticos habituales. El segundo se usa para proporcionar predicciones futuras y se puede aplicar a una amplia variedad de modelos incluidos los ARIMA. Las funciones de este paquete se pueden categorizar en:

- Herramientas exploratorias
  - `tsdisplay`: Gráfico integrado con la serie de tiempo y las funciones ACF y PACF (véase ejemplo en figura 4.11)
  - `seasonplot`: Gráfico para investigar si hay componente estacional.
  - `tsoutliers`: Identificación de outliers.
- Otros modelos
  - `tbats`: Modelo tbat (Suavización exponencial del modelo de espacio de estados con transformaciones Box-Cox, errores ARMA y componentes de estacionales y de tendencia)

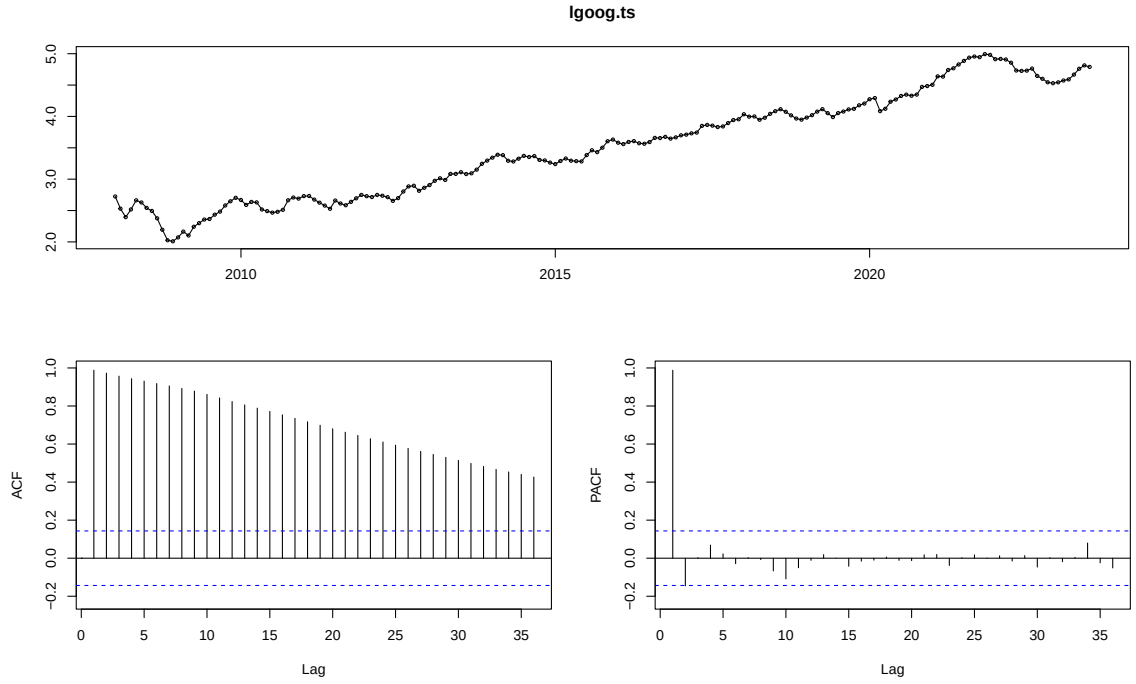
– `HoltWinters`: Predicción con modelo Holt-Winters.

– `StructTS`: Series temporales estructurales.

Un ejemplo de la aplicación de `auto.arima` se ve a continuación. En general, este comando tiene tendencia a seleccionar modelos más complejos de lo necesario. Comparando el modelo seleccionado con los anteriores se ve que el AIC es peor y que se estiman parámetros que podrían ser no significativos (`sar1`).

```
library(forecast)
tsdisplay(lgoog.ts)
aufit = auto.arima(lgoog.ts, d = 1, max.p = 12, max.q = 12, max.order = 15,
    max.d = 1)
summary(aufit)

> Series: lgoog.ts
> ARIMA(1,1,2)(1,0,0)[12] with drift
>
> Coefficients:
>           ar1     ma1     ma2     sar1   drift
>       -0.9385  1.2834  0.3407  -0.0296  0.0109
> s.e.   0.0910  0.1143  0.0765   0.0864  0.0057
>
> sigma^2 = 0.003541:  log likelihood = 263.36
> AIC=-514.71   AICc=-514.24   BIC=-495.36
>
> Training set error measures:
>                         ME       RMSE        MAE        MPE
> Training set 0.0001711061 0.05854602 0.04461566 -0.0308836
>                MAPE      MASE        ACF1
> Training set 1.4056 0.1987154 -0.02277447
>
>  Box-Ljung test
>
> data:  aufit$residuals
> X-squared = 17.021, df = 12, p-value = 0.1488
```

Figura 4.11: Salida de la función `tsdisplay`



Fonte: O autor.

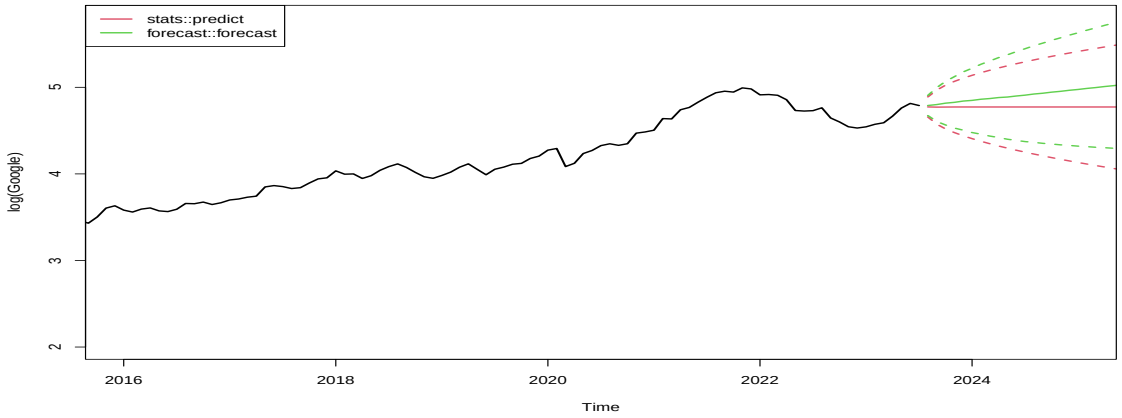
4.4.4 Predicción

Una vez que se ha seleccionado un modelo, se ha estimado y se ha diagnosticado como válido, la última etapa del camino es realizar predicciones. Para este fin, se usa el método `predict` para objetos que provienen del comando `arima`, o `forecast` para objetos conocidos por el paquete del mismo nombre. En el primer caso, se devuelven las predicciones en la componente `$pred` y la desviación estándar de predicción en la componente `$se`. Para el caso de `forecast`, la media de predicción se devuelve en `$mean` y los límites del intervalo de confianza en las componentes `$lower` y `$upper` (por defecto a niveles del 80% y 95%). El siguiente código muestra como construir a mano los gráficos con los intervalos de confianza aunque en el caso del paquete `forecast` se puede usar directamente el método `plot`.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

```
pr = predict(res.arma, n.ahead = 24)
pr2 = forecast(aufit, h = 24)
A = pr[["se"]] %*% matrix(c(-1.96, 1.96), nrow = 1)
pr.ts = cbind(pr[["pred"]], pr[["pred"]] + A[, 1], pr[["pred"]] +
    A[, 2])
pr2.ts = cbind(pr2$mean, pr2$lower[, "95%"], pr2$upper[, "95%"])
```

Figura 4.12: Predicciones a $h = 24$ con el modelo ARIMA(2,1,0) (stats) y el ARIMA(1,1,2)x(1,0,0) (forecast)



Fonte: O autor.

Otra opción para elaborar predicciones es realizar muchas simulaciones de las predicciones a futuro y a partir de estas calcular las características deseadas de la distribución de las predicciones. Estas simulaciones se pueden generar a partir de distribuciones fijas (por ejemplo, gaussiana) o bien a partir de técnicas de remuestreo como el Bootstrap. Este proceso se hace a partir del comando `arima.sim` que permite simular series ARIMA con órdenes y parámetros prefijados. Para simular una trayectoria se usa simplemente

```
x = arima.sim(model = list(order = c(2, 1, 0), ar = c(0.75, -0.25)),
    n = 200, rand.gen = rnorm, sd = 0.3)
```

El siguiente código usa la función `arima.simforecast` (de elaboración propia) que sólo es una reformulación de `arima.sim` para generar varias simulaciones futuras a un horizonte determinado con el mismo modelo y con el mismo comienzo (el final de la serie temporal). El resultado puede verse gráficamente en la figura 4.13.

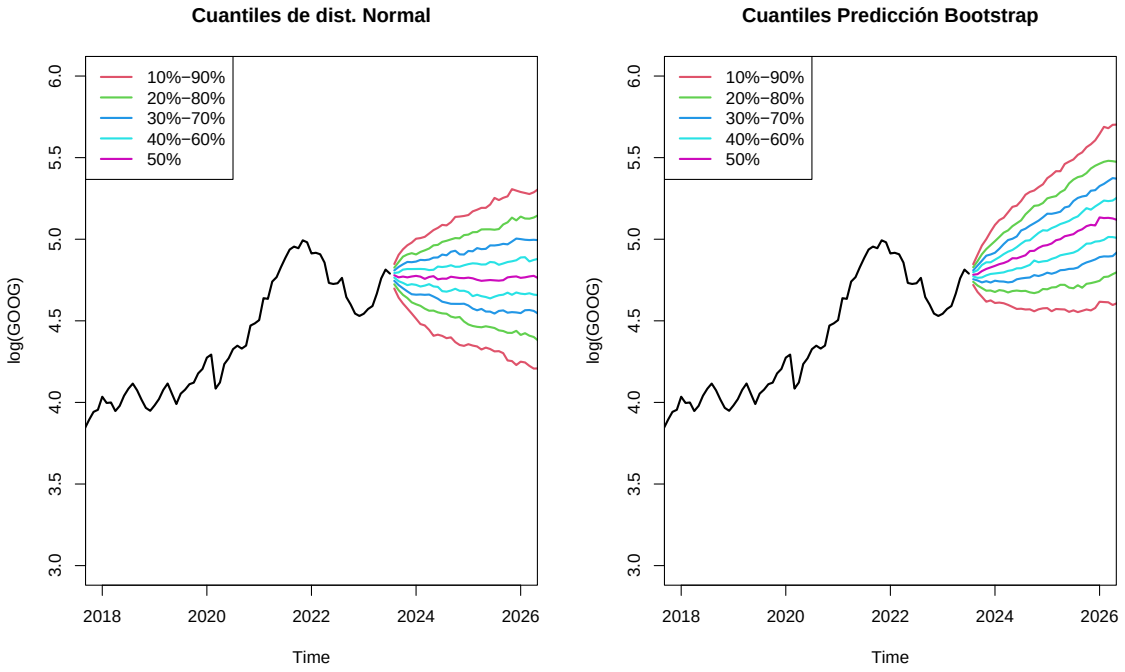
```
source("arima.simforecast.R")
# Función para generar residuos por Bootstrap
myboot = function(n, vec, prob = NULL) {
    sample(vec, size = n, replace = TRUE, prob = prob)
}
# Sim. de predicciones usando una distribución gaussiana
xsim = arima.simforecast(res.arma, n.ahead = 36, std = TRUE,
    nrep = 500)
# Sim. de predicciones usando Bootstrap (std=FALSE)
xsim2 = arima.simforecast(res.arma, n.ahead = 36, rand.gen = myboot,
    std = FALSE, vec = residuals(res.arma), nrep = 500)
# Calcular los cuantiles
quan.ts = ts(t(apply(xsim, 1, quantile, prob = seq(0.1, 0.9,
    by = 0.1))), start = start(xsim), frequency = frequency(xsim))
quan2.ts = ts(t(apply(xsim2, 1, quantile, prob = seq(0.1, 0.9,
    by = 0.1))), start = start(xsim2), frequency = frequency(xsim2))
```

4.4.5 Otros paquetes

Muchos otros paquetes están disponibles en el CRAN de R para la modelización ARIMA de series de tiempo. Véase, para más información, la *Task View* dedicada a series temporales en R https://cran.r-project.org/web/views/TimeSeries.html. Lo siguiente no pretende ser una lista exhaustiva de estos paquetes sino más bien una recomendación personal de alguno de ellos con sus funciones más destacadas.

`TSA` : Extiende alguna herramienta clásica e incluye ciertos tests y nuevos modelos como el TAR.

Figura 4.13: Cuantiles de predicción obtenidos a partir de la aproximación normal (izda) y por Bootstrap (dcha).



Fonte: O autor.

- `BoxCox.ar`: Transformación óptima de una serie.
- `eacf`: Función ACF extendida.
- `detectAO, detectIO`: Detección de outliers
- `arima, arimax`: Modelos ARIMA y de transferencia.
- `tar, predict.tar`: Threshold AR Models.
- Tests: `runs`: Independencia, `McLeod.Li.test`: Garch, `Keenan.test`: No linealidad, `Tsay.test`: AR cuadrático, `tlrt`: Threshold AR

`FinTS` : Paquete asociado con el libro de Tsay con muchos ejemplos y atajos a herramientas de otros paquetes.

`fArma` : Módulo de `Rmetrics` para la modelización ARIMA. No hay actualización para versiones de R superiores a la 4.0.0.

- `PerformanceAnalytics`: Funciones para la elaboración de tablas y gráficos exploratorios.

4.5 VOLATILIDAD CONDICIONAL

Los modelos de volatilidad (varianza) condicional surgen en el mundo financiero para modelar los retornos que son series de tiempo derivadas de los precios de activos financieros (P_t). Los retornos se definen como:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \text{ (discreto)}, \; r_t = \log\left(P_t/P_{t-1}\right) \text{ (continuo)}$$

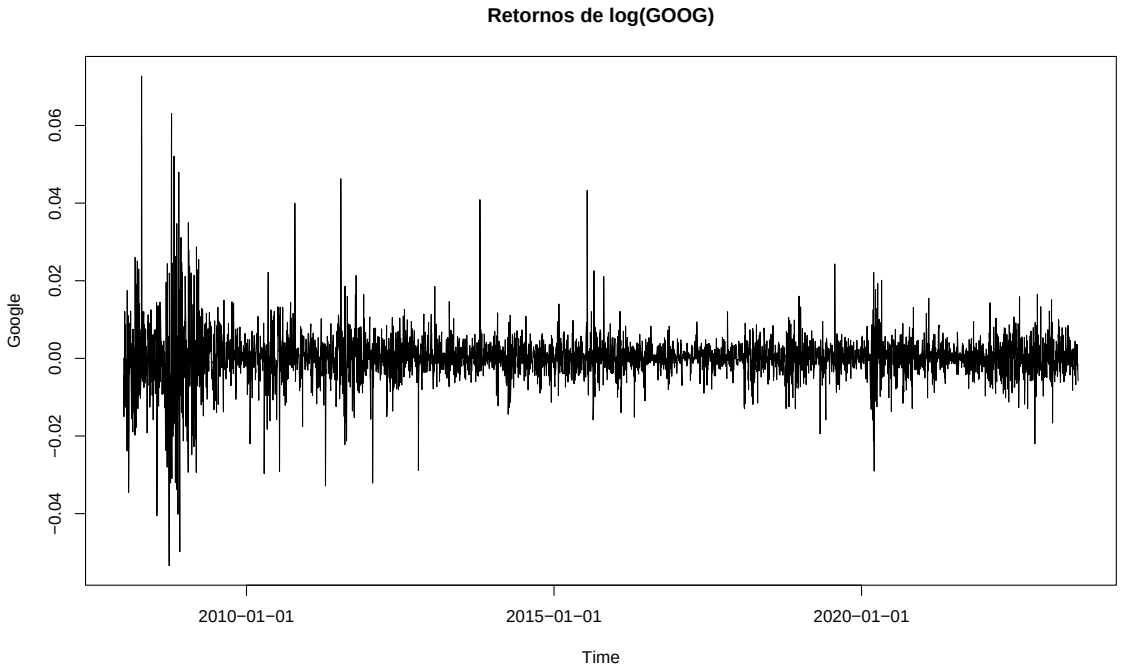
Sea en su versión discreta o continua, las series de retornos tienen una serie de características específicas muy marcadas que se listan a continuación:

- Media constante con estructura dinámica reducida (a lo sumo ARMA(1,1)).
- Alta kurtosis.
- Agrupamiento de volatilidad (véase figura 4.14).
- Volatilidad persistente (memoria larga).
- Efecto apalancamiento (impacto asimétrico de buenas y malas noticias).
- Cambios en el régimen de la volatilidad debido a factores estacionales, externos o inesperados.
- Volatilidad influenciada por la volatilidad de otros retornos/precios.

```
library(rugarch)
goor = returns(log(googc))  # Retornos a partir de datos diarios
```

Asimetría (log(GOOG))=0.504, Kurtosis(log(GOOG))=17.284

Figura 4.14: Retornos de log($GOOG$). Se aprecia un clúster de volatilidad muy marcado antes de 2010 y otro de menos intensidad al final de la serie.



Fonte: O autor.

4.5.1 GARCH Models

Los modelos GARCH (del inglés, modelo autorregresivo generalizado de heterocedasticidad condicional) es la manera clásica de enfrentarse a los modelos de volatilidad condicional. Estos modelos se definen a partir de series de ruido blanco donde la volatilidad (varianza) condicional es constante a largo plazo pero presenta cierta estructura local.

Siendo preciso, una serie $\{\epsilon_t\}$ se dice que es ARCH(p), si

$$\begin{aligned} \epsilon_t &= w_t \sigma_t \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 \end{aligned}$$

donde $\{w_t\}$ es un ruido blanco estándar (varianza unidad) y $\sigma_t^2 = \text{Var}\,(\epsilon_t|\mathcal{F}_{t-1})$ donde $\mathcal{F}_{t-1}$ es la σ-álgebra generada por $\{\epsilon_{t-1}, \epsilon_{t-2}, \ldots\}$. Por su propia definición,

siempre ocurrirá que la serie tiene media cero

$$\mathbb{E}\left(\epsilon_t\right) = \mathbb{E}\left(\mathbb{E}\left(\epsilon_t|\mathcal{F}_{t-1}\right)\right) = \mathbb{E}\left(\sigma_t\mathbb{E}\left(w_t|\mathcal{F}_{t-1}\right)\right) = 0$$

y la varianza viene dada por:

$$\text{Var}\left(\epsilon_t\right) = \mathbb{E}\left(\epsilon_t^2\right) = \mathbb{E}\left(\mathbb{E}\left(\epsilon_t^2|\mathcal{F}_{t-1}\right)\right) = \alpha_0 + \sum_{i=1}^{p}\alpha_i\mathbb{E}\left(\epsilon_{t-i}^2\right)$$

$$\text{Var}\left(\epsilon_t\right) = \frac{\alpha_0}{1-\sum_{i=1}^{p}\alpha_i}$$

Análogamente, una serie $\{\epsilon_t\}$ se dice que es GARCH(p,q), si

$$\begin{aligned}\epsilon_t &= w_t\sigma_t \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p}\alpha_i\epsilon_{t-i}^2 + \sum_{j=1}^{q}\beta_j\sigma_{t-j}^2\end{aligned}$$

donde $\{w_t\}$ es un ruido blanco estándar y $\sigma_t^2 = \text{Var}\left(\epsilon_t|\mathcal{F}_{t-1}\right)$.

Procediendo como en el ARCH(p) se tiene que:

$$\text{Var}\left(\epsilon_t\right) = \alpha_0 + \sum_{i=1}^{p}\alpha_i\mathbb{E}\left(\epsilon_{t-i}^2\right) + \sum_{j=1}^{q}\beta_j\mathbb{E}\left(\epsilon_{t-j}^2\right)$$

$$\text{Var}\left(\epsilon_t\right) = \frac{\alpha_0}{1-\sum_{i=1}^{p}\alpha_i - \sum_{j=1}^{q}\beta_j}$$

Como en el caso de los modelos ARIMA, lo importante es la determinación de los órdenes de la parte GARCH. La buena noticia es que las herramientas son similares a las usadas con los modelos ARIMA debido a la siguiente propiedad:

$$\epsilon_t \sim \text{GARCH}(p,q) \implies \epsilon_t^2 \sim \text{ARMA}(\max(p,q),q)$$

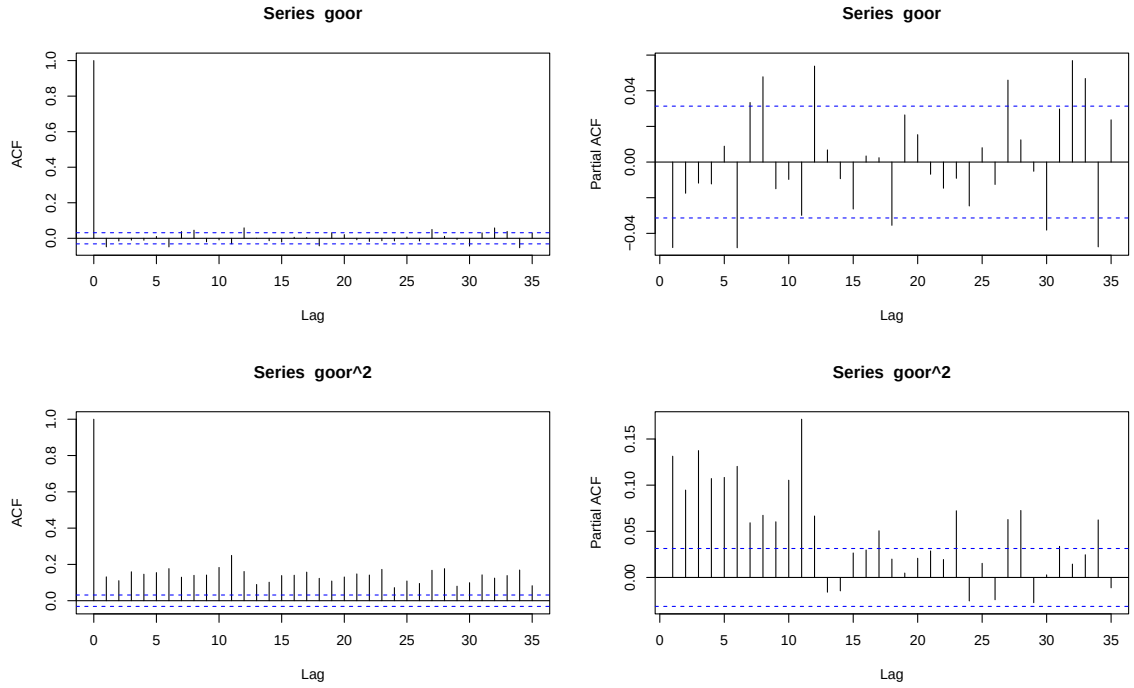
Esta propiedad nos dice que podemos usar las funciones ACF y PACF como antes para los modelos ARIMA pero ahora sobre la serie de los residuos al cuadrado

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

como vemos en el siguiente ejemplo:

```
acf(goor^2)
pacf(goor^2)
```

Figura 4.15: ACF y PACF de la serie de retornos (primera fila) y de la serie de retornos al cuadrado (segunda fila)



Fonte: O autor.

Como se ve en la figura 4.15, no hay prácticamente modelo en la parte dinámica (asociada a la media) que se ve en la primera fila pero si muchas barras se salen de los límites confidenciales en la segunda fila que es la parte asociada a la volatilidad.

4.5.2 Modelos ARIMA–GARCH

La modelización más completa de las series financieras se realiza ajustando la media (parte dinámica) con modelos ARIMA y la varianza (volatilidad) con

modelos GARCH. Esto es precisamente lo que hace el paquete `rugarch` ((GALANOS, 2022)).

4.5.2.1 Dinámica

Como en el modelo ARIMA

$$\begin{aligned}\Phi_p(B)(1-B)^d(Z_t-\mu_t) &= \Theta_q(B)\epsilon_t,\\ \text{con } \mu_t &= \mu+\sum_{i=1}^{m_1}\delta_i X_{i,t}+\sum_{i=m_1+1}^{m}\delta_i X_{i,t}\sigma_t+\xi\sigma_t^k\end{aligned}$$

donde $\{x_i\}$ son regresores externos y $\{\epsilon_t\}$ es un proceso de ruido estándar siguiendo alguna distribución de las habituales (Normal, GED, Student, ...). También se pueden considerar modelos de memoria larga ($0<d<1$) en esta formulación.

4.5.2.2 Volatilidad

Como en el modelo GARCH

$$\epsilon_t=\omega_t\sigma_t$$

donde se pueden considerar varias estructuras para σ_t:

- SGARCH: $\sigma_t^2=\underbrace{\left(\omega+\sum_{j=1}^{m}\varsigma_j\nu_{j,t}\right)}_{\omega_t}+\sum_{j=1}^{p}\beta_j\sigma_{t-j}^2+\sum_{j=1}^{q}\alpha_j\epsilon_{t-j}^2$ donde $\{\nu_j\}$ son regresores externos.
- EGARCH: $\log\sigma_t^2=\omega_t+\sum_{j=1}^{p}\beta_j\log\sigma_{t-j}^2+\sum_{j=1}^{q}\left(\alpha_j\frac{|\epsilon_{t-j}|+\gamma_i\epsilon_{t-j}}{\sigma_{t-i}}\right)$
- GJR-GARCH: $\sigma_t^2=\omega_t+\sum_{j=1}^{p}\beta_j\sigma_{t-j}^2+\sum_{j=1}^{q}\left(\alpha_j\epsilon_{t-j}^2+\gamma_j I_{(\epsilon_{t-j}<0)}\epsilon_{t-j}^2\right)$
- APGARCH: $\sigma_t^\lambda=\omega_t+\sum_{j=1}^{p}\beta_j\sigma_{t-j}^\lambda+\sum_{j=1}^{q}\alpha_j\left(|\epsilon_{t-j}|-\gamma_j\epsilon_{t-j}\right)^\lambda$
- fGARCH:
  $\sigma_t^\lambda=\omega_t+\sum_{j=1}^{p}\beta_j\sigma_{t-j}^\lambda+\sum_{j=1}^{q}\alpha_j\sigma_{t-j}^\lambda\left(\left|\frac{\epsilon_{t-j}}{\sigma_{t-j}}-\eta_{2j}\right|-\eta_{1j}\left(\frac{\epsilon_{t-j}}{\sigma_{t-j}}-\eta_{2j}\right)\right)^\delta$

- Two component GARCH:

$$\begin{aligned}\sigma_t^\lambda &= q_t^\lambda + s_t^\lambda \\ q_t^\lambda &= \alpha_1|\epsilon_{t-1}|^\lambda + \beta_1 q_{t-1}^\lambda \sim \text{PGARCH}(1,1) \\ s_t^\lambda &= \alpha_0 + \alpha_2|\epsilon_{t-1}|^\lambda + \beta_2 s_{t-1}^\lambda \sim \text{PGARCH}(1,1)\end{aligned}$$

  En el modelo de dos componentes, puede establecerse que cada parte componente siga su modelo específico de los anteriores.

Para estimar el modelo se crea primero un objeto con las especificaciones apropiadas (`ugarchspec`) que se le pasa a `ugarchfit` para su estimación. La salida va a ser muy detallada cubriendo todos los aspectos que puede necesitar un usuario para comprobar si la modelización es aceptable.

```
spec2 = ugarchspec(variance.model = list(model = "eGARCH", garchOrder = c(1,
    1), external.regressors = cbind(monday)), mean.model = list(armaOrder = c(0,
    0)), distribution.model = "sstd")
fit2 = ugarchfit(spec2, data = goor)
```

Troceando la salida, la primera parte está dedicada a la estimación de los parámetros

```
>
> *---------------------------------*
> *          GARCH Model Fit        *
> *---------------------------------*
>
> Conditional Variance Dynamics
> -----------------------------------
> GARCH Model : eGARCH(1,1)
> Mean Model : ARFIMA(0,0,0)
> Distribution : sstd
```

```
>
> Optimal Parameters
> ------------------------------------
>         Estimate  Std. Error    t value Pr(>|t|)
> mu      0.000191    0.000055    3.48244 0.000497
> omega  -0.095570    0.012749   -7.49631 0.000000
> alpha1 -0.079564    0.009224   -8.62596 0.000000
> beta1   0.990652    0.000210 4709.62411 0.000000
> gamma1  0.124972    0.005869   21.29340 0.000000
> vxreg1 -0.028150    0.067159   -0.41915 0.675104
> skew    0.966962    0.020151   47.98481 0.000000
> shape   4.019238    0.259962   15.46086 0.000000
>
> Robust Standard Errors:
>         Estimate  Std. Error    t value Pr(>|t|)
> mu      0.000191    0.000053    3.61359 0.000302
> omega  -0.095570    0.013219   -7.22975 0.000000
> alpha1 -0.079564    0.009127   -8.71775 0.000000
> beta1   0.990652    0.000328 3019.83351 0.000000
> gamma1  0.124972    0.008354   14.95963 0.000000
> vxreg1 -0.028150    0.066428   -0.42376 0.671740
> skew    0.966962    0.022507   42.96357 0.000000
> shape   4.019238    0.284528   14.12601 0.000000
>
> LogLikelihood : 15535.67
....
```

Los dos bloques de parámetros se diferencian en el cálculo de los errores estándar (no robusto/robusto) pero, en ambos casos, todos los parámetros son significativos excepto `vxreg1` que es el asociado a la variable indicadora incluida como regresor externo. El parámetro `gamma1` asociado al efecto apalancamiento es significativo, lo que indica que la serie de retornos se comporta de forma diferente ante noticias buenas y malas en el mercado. Los dos últimos parámetros `skew` y `shape` corresponden a los parámetros de sesgo y grados de libertad de la distribución SSTD (Skew Student t–Distribution) elegida en este caso.

El siguiente tramo de resultados incluye tres tests para verificar que ni en la serie del residuo ni en la serie del residuo al cuadrado tenemos algún tipo de

dependencia.

```
....

>
> Information Criteria
> ------------------------------------
>
> Akaike       -7.9548
> Bayes        -7.9419
> Shibata      -7.9548
> Hannan-Quinn -7.9502
>
> Weighted Ljung-Box Test on Standardized Residuals
> ------------------------------------
>                         statistic p-value
> Lag[1]                     0.2726  0.6016
> Lag[2*(p+q)+(p+q)-1][2]    0.3119  0.7891
> Lag[4*(p+q)+(p+q)-1][5]    0.9134  0.8789
> d.o.f=0
> H0 : No serial correlation
>
> Weighted Ljung-Box Test on Standardized Squared Residuals
> ------------------------------------
>                         statistic p-value
> Lag[1]                     0.8186  0.3656
> Lag[2*(p+q)+(p+q)-1][5]    1.3929  0.7662
> Lag[4*(p+q)+(p+q)-1][9]    1.8137  0.9260
> d.o.f=2
>
> Weighted ARCH LM Tests
> ------------------------------------
>             Statistic Shape Scale P-Value
> ARCH Lag[3]   0.08883 0.500 2.000  0.7657
> ARCH Lag[5]   0.32700 1.440 1.667  0.9330
> ARCH Lag[7]   0.59424 2.315 1.543  0.9691
>
....
```

El objeto `fit` que se devuelve de `ugarchfit` tiene asociados muchos métodos

para obtener información precisa de sus componentes. A los usuales métodos `fitted`, `coef`, `residuals` comunes en otros procedimientos se unen:

- `vcov`: Matriz de varianzas entre los parámetros.
- `confint`: Intervalos de confianza para los parámetros.
- `sigma`: Valores de σ_t
- `uncvariance`: Varianza a largo plazo (incondicional).
- `uncmean`: Media incondicional.
- `reduce`: Reestima el modelo eliminando las variables no significativas.
- `quantile`: Cuantiles condicionales.
- `nyblom`: Calcula el test de Hansen–Nyblom (1990).
- `newsimpact`: Calcula la función de impacto.

El método `plot`, por defecto, muestra un menú interactivo con 12 posibles gráficos:

1. Serie con los límites de $\pm 2\hat{\sigma}_t$.
2. Serie con el límite 1% del VaR (Value at Risk).
3. Desviación estándar condicional (vs |retornos|)
4. ACF de las observaciones.
5. ACF de las observaciones al cuadrado.
6. ACF de la observaciones en valor absoluto.
7. Correlación cruzada.
8. Densidad estimada de los residuos estandarizados.

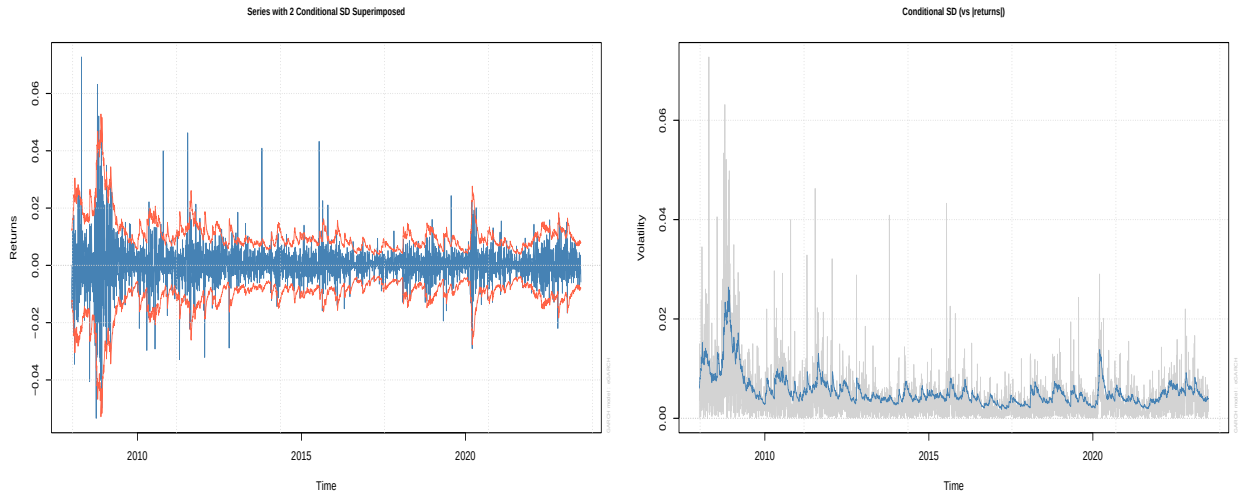
APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

9. QQ-Plot de los residuos estandarizados.
10. ACF de los residuos estandarizados.
11. ACF de los residuos estandarizados al cuadrado.
12. Curva de impacto de noticias.

Para dibujar sólo un gráfico individual debe usarse `plot(fit,which=a)` donde `a` es el número de orden de la anterior lista.

```
plot(fit2, which = 1)
plot(fit2, which = 3)
```

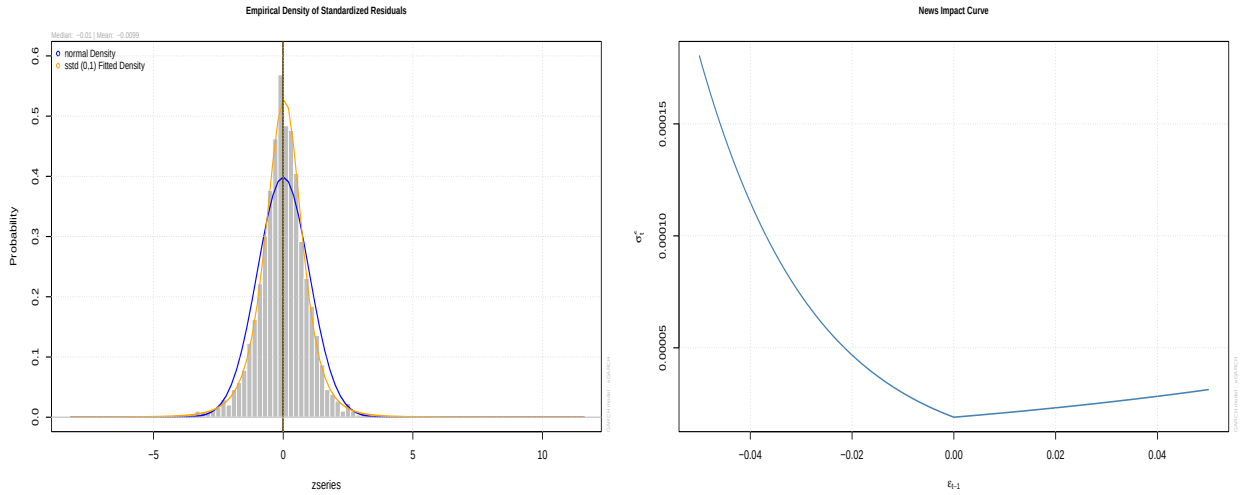
Figura 4.16: Gráfico de los retornos con el intervalo dado por la desviación condicional (izda). Gráfico de la desviación condicional vs valor absoluto de los retornos (dcha).



Fonte: O autor.

```
plot(fit2, which = 8)
plot(fit2, which = 12)
```

Figura 4.17: Gráfico de la densidad de los residuos estandarizados (izda). Función de impacto de noticias (dcha).



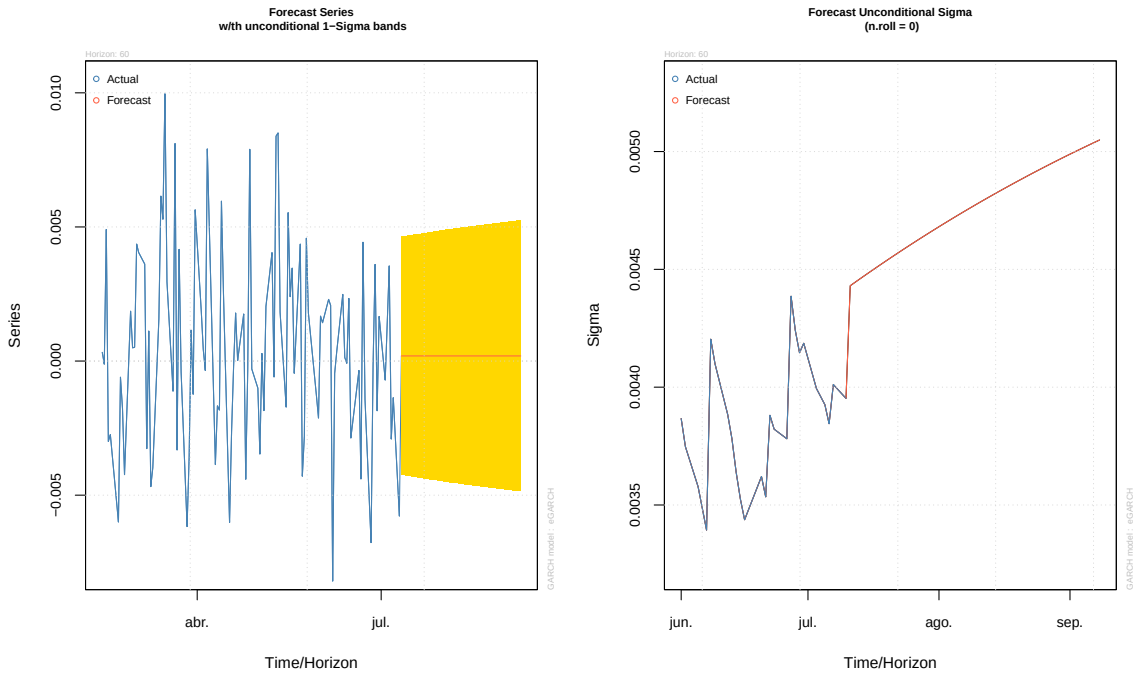
Fonte: O autor.

Finalmente para la predicción, el paquete `rugarch` dispone de dos opciones. La primera es calcular las predicciones usando `ugarchforecast`. Esta función usa las ecuaciones de predicción clásicas proporcionando valores futuros de la media y de la desviación condicional. El método `plot` aplicado a las predicciones presenta un menú interactivo con 4 opciones para dibujar la predicción de la media y la desviación condicional de las cuales las usuales son la 1 y la 3. Los valores pueden obtenerse directamente con `fitted` y `sigma` aplicados al objeto de predicción. Si el modelo tiene variables externas, deben proporcionarse para poder realizar la predicción. En el ejemplo siguiente se incluye el vector `nmonday` con la variable indicadora para las próximas 12 semanas (60 datos).

```
nmonday = rep(c(1, 0, 0, 0, 0), 12)
fit2.fore = ugarchforecast(fit2, n.ahead = 60, external.forecasts = list(mregfor = cbind(nmonday)))
par(mfrow = c(1, 2))
plot(fit2.fore, which = 1)
plot(fit2.fore, which = 3)
```

La segunda opción es proyectar simulaciones futuras con Bootstrap usando

Figura 4.18: Predicción de la serie con sus intervalos de confianza (izda) y de $\hat{\sigma}_t$ (dcha)



Tone: O autor.

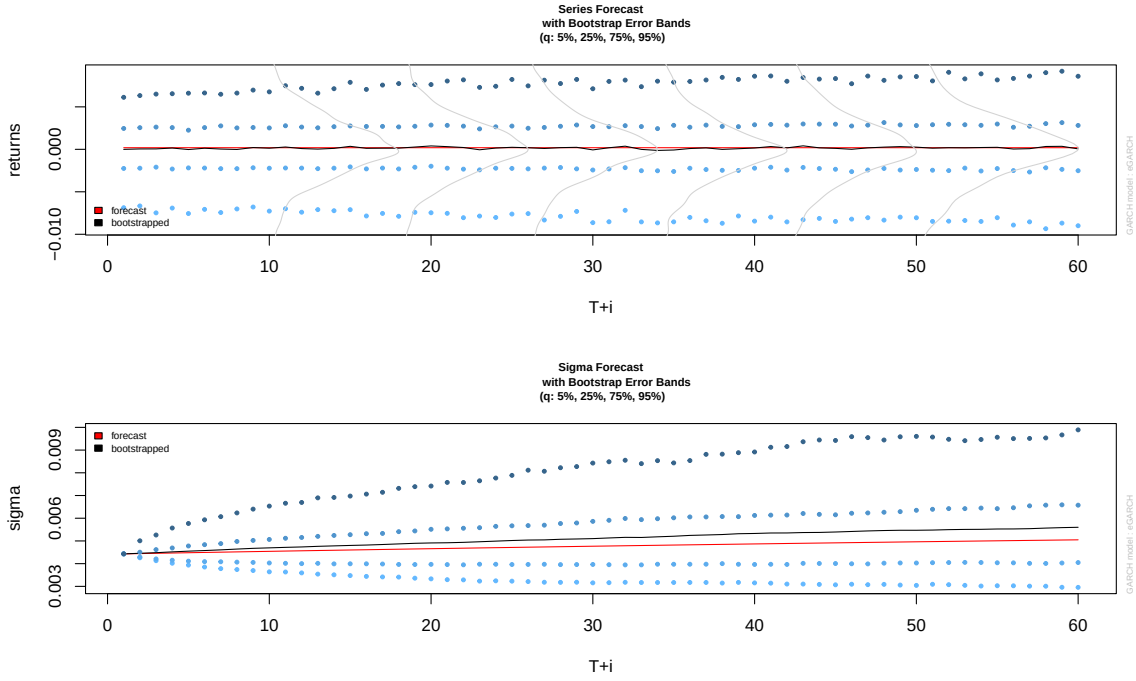
`ugarchboot` que dispone de muchas opciones. Un ejemplo simple aparece en el código siguiente. La salida de este procedimiento también dispone de un método `plot` que da como resultados los gráficos de la figura 4.19.

```
fit2.boot = ugarchboot(fit2, method = "Partial", sampling = "spd",
    n.ahead = 60, external.forecasts = list(mregfor = cbind(nmonday)),
    n.bootpred = 2000, mexsimdata = t(rbind(replicate(2000, nmonday))))
```

4.6 CONCLUSIÓN

El análisis de series de tiempo en R es un tópico accesible con muchas posibilidades. Este documento sólo pretende ser una guía rápida como ejemplo introductorio para un usuario que se quiera sumergir en la temática. Una revisión de la *Task Views* sobre series de tiempo (https://cran.r-project.org/web/

Figura 4.19: Gráficos de predicción obtenidos con Bootstrap.



Fonte: O autor.

views/TimeSeries.html) basta para darse cuenta de la cantidad ingente de documentación y opciones que hay disponibles. A modo de ejemplo, se incluye una lista a continuación de temáticas y paquetes asociados:

- **Importar datos**: `pdfetch`, `Quandl`, `alfred`, `BETS`, `readabs`
- **Nuevos clásicos**: `dygraphs`, `fable`, `feasts`, `ggTimeSeries`, `modeltime`, `TSstudio`
- **Regresión Dinámica**: `dyn`, `dynlm`, `tsDyn`, `dse`
- **Multivariante**: `vars`, `MTS`, `BigVAR`, `fsMTS`, `HDTSA`, `fMultivar`
- **Filtro de Kalman**: `astsa`, `KFAS`, `FKF`
- **ARFIMA**: `fracdiff`
- **Raíz Unitaria**: `urca`

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

- **GARCH**: `rugarch`, `rmgarch`, `fGarch`
- **Bayesian**: `bayesforecast`, `bayesGARCH`, `mbsts`, `bmgarch`

En esta lista se quedan fuera muchos otros temas que cuentan también con paquetes en R como Outliers, Predicción con métodos de Machine Learning, Análisis en el dominio de frecuencias o Modelos no lineales. Como se dijo al inicio, este documento trata de ser sólo el punto de partida para empezar a navegar por el universo de las series de tiempo que está lleno de peculiaridades y posibilidades. Llegar a dominar este universo es un viaje largo y no exento de dificultades pero, como en las grandes aventuras, el propio viaje es el premio aunque sea necesario armarse de paciencia y determinación.

4.7 REFERÊNCIAS

BOSHNAKOV, Georgi N. **timeDate: Rmetrics - Chronological and Calendar Objects (v.4022.108)**. [S.l.: s.n.], 2023. Disponível em: https://cran.r-project.org/web/packages/timeDate/index.html.

_____. **timeSeries: Financial Time Series Objects (Rmetrics) (v.4030.106)**. [S.l.: s.n.], 2023. Disponível em: https://cran.r-project.org/web/packages/timeSeries/index.html.

CRYER, Jonathan D.; CHAN, Kung-Sik. **Time Series Analysis. With Applications in R**. 2. ed. [S.l.]: Springer, 2008. ISBN 978-0-387-75958-6.

GALANOS, Alexios. **rugarch: Univariate GARCH Models (v.1.4-9)**. [S.l.: s.n.], 2022. Disponível em: https://cran.r-project.org/web/packages/rugarch/index.html.

HORNIK, Kurt. **chron: Provides chronological objects which can handle dates and times (v. 2.3-61).** [S.l.: s.n.], 2023. Disponível em: https://cran.r-project.org/web/packages/chron/index.html.

HYNDMAN, Rob. **forecast: Forecasting Functions for Time Series and Linear Models (v.8.21)**. [S.l.: s.n.], 2023. Disponível em: https://cran.r-project.org/web/packages/forecast/index.html.

SHUMWAY, Robert H.; STOFFER, David S. **Time series analysis and its applications**. Fourth. [S.l.]: Springer, Cham, 2017. p. xiii+562. (Springer Texts in Statistics). With R examples. ISBN 978-3-319-52451-1; 978-3-319-52452-8. DOI: 10.1007/978-3-319-52452-8.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

SPINU, Vitalie. **lubridate: Make Dealing with Dates a Little Easier (v. 1.9.2)**. [S.l.: s.n.], 2023. Disponível em: https://cran.r-project.org/web/packages/lubridate/index.html.

TSAY, Ruey S. **Analysis of financial time series**. Third. [S.l.]: John Wiley & Sons, Inc., Hoboken, NJ, 2010. p. xxiv+677. (Wiley Series in Probability and Statistics). ISBN 978-0-470-41435-4. DOI: 10.1002/9780470644560.

ULRICH, Joshua M. **quantmod: Quantitative Financial Modelling Framework (v.0.4.23)**. [S.l.: s.n.], 2023. Disponível em: https://cran.r-project.org/web/packages/quantmod/index.html.

WEI, William W. S. **Time series analysis**. Second. [S.l.]: Addison Wesley/Pearson, Boston, MA, 2006. p. xxii+614. Univariate and multivariate methods. ISBN 0-321-32216-9.

ZEILEIS, Achim. **zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations) (v.1.8-12)**. [S.l.: s.n.], 2023. Disponível em: https://cran.r-project.org/web/packages/zoo/index.html.

Capítulo 5

CURVAS E CORES EM R: MOVIMENTOS RÍGIDOS NO PLANO

Autor: João Paulo Martins dos Santos [1]
Academia da Força Aérea, Pirassununga, SP
e-mail: jp2@alumni.usp.br

Os processos de composição de movimentos rígidos de rotação, translação e homotetia serão aplicados às curvas planas para gerar figuras com simetrias radiais. Apesar da simplicidade das construções, as cores, utilizando modelo RGB de coloração, resultam em padrões intrigantes. Dois métodos de coloração distintos daqueles apresentados em (ALCOFORADO et al., 2023) são explorados: o método de coloração gradiente e o método de coloração por rotações sucessivas. As cores são adicionadas por meio do R básico, eliminando, dessa forma, instalações de pacotes adicionais, exceto o pacote **ggplot2** que será utilizado para a visualização. Os resultados expandem ainda mais o potencial de possibilidades para as construções utilizando o R como ferramenta base do processo, pois ilustra novas possibilidades de métodos de coloração; além disso, insere uma nova discussão, a dependência dos resultados com respeito a direção de crescimento ou decrescimento das razões de homotetias.

[1]Agradecimentos ao Comitê Organizador do VII SER por tornar essa obra possível

5.1 INTRODUÇÃO

Este capítulo foi baseado no **E-book** de (ALCOFORADO et al., 2023) denominado **Mandalas, curvas clássicas e visualização com R**, disponível no Portal de Livros Abertos da Universidade de São Paulo. O E-Book traz um detalhamento do processo de construção de mandalas utilizando curvas planas definidas por meio de expressões paramétricas explícitas e utiliza o **ggplot2** como recurso para visualização. Os desenvolvimentos partem das curvas simples tais como círculos e, por meio de movimentos de rotação, translação e homotetias resultam, eventualmente, em uma composição complexa e com um aspecto, no mínimo, intrigante. As cores foram inseridas por meio do modelo de cores RGB (*Red, Green, Blue*) ou métodos do tipo gradiente disponíveis no R básico. Neste texto, a contribuição será a apresentação de um método para colorações de rotações sucessivas; um método gradiente e uma discussão sobre a dependência dos resultados em relação às direções das razões de homotetias.

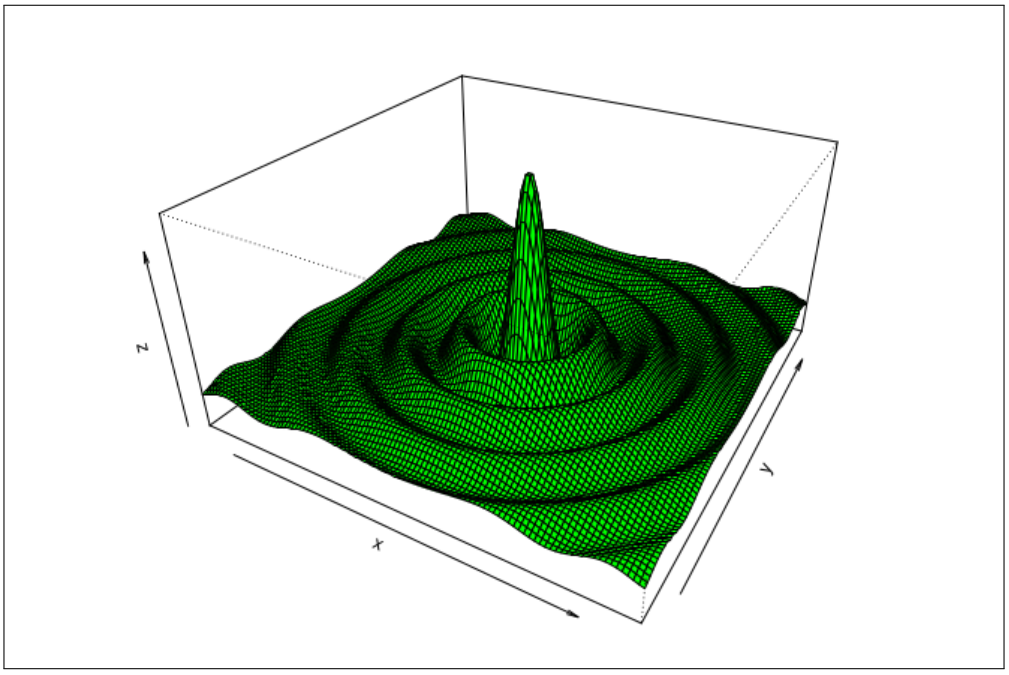
Os detalhes da metodologia de construção passo a passo e com códigos completos podem ser consultados na referência base. No entanto, a seção 5.3 faz uma revisão e detalhamento do processo de construção.

Dentre as motivações para as construções apresentadas estão a exploração da capacidade do R, de computação e visualização, das curvas paramétricas com expressões explícitas da Matemática. O potencial de visualização, aplicado a Matemática, pode ser observado na Figura 5.1, a qual ilustra uma função de duas variáveis reais a valores reais, $f : \mathbb{R} \longrightarrow \mathbb{R}$:

$$f(x,y) = \frac{\text{sen}\left(\sqrt{x^2+y^2}\right)}{\sqrt{x^2+y^2}}$$

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 5.1: Gráfico da função $f : \mathbb{R} \longrightarrow \mathbb{R}$ $f(x, y) = \frac{\text{sen}\left(\sqrt{x^2+y^2}\right)}{\sqrt{x^2+y^2}}$.



Fonte: O autor.

5.2 OBJETIVOS

De forma geral, os objetivos deste texto são a apresentação de dois métodos distintos para a coloração das figuras construídas por meio dos movimentos rígidos de rotação, translação e homotetias utilizando R/*RStudio*. De forma a fornecer um encadeamento do texto, os seguintes passos serão executados:

- Apresentar as curvas planas e respectivos códigos;
- Ilustrar os movimentos rígidos de rotação, translação e homotetia aplicados às curvas planas;
- Ilustrar o processo de composição de curvas planas e movimentos rígidos;
- Ilustrar a adição de cores utilizando os modelos RGB disponíveis no R por meio da função *colors*();

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

- Ilustrar a utilização de cores gradientes e cores para rotações sucessivas;
- Ilustrar outras possibilidades de coloração: dependência da direção da homotetia.

A complexidade das figuras e complexidade das cores resultam das interações entre os movimentos rígidos combinados com o processo de coloração.

5.3 APLICAÇÃO

A geração das figuras com simetria radial, denominadas Mandalas, pode ser resumida em um conjunto de elementos distintos apresentados a seguir. Este conjunto de passos não precisa, necessariamente, ser executado em sequência, mas é norteador do processo de construção. Caso o objetivo seja obter figuras com apenas uma cor, então os passos de coloração devem ser descartados. Se, por outro lado, o objetivo for explorar os diferentes padrões de cores e respectivas complexidades, então os métodos de coloração devem ser considerados.

1.) Escolher as curvas ou figuras geométricas. A escolha da curva ou figura geométrica não é um elemento restritivo ao processo de construção. Algumas possibilidades são apresentadas a seguir:

- curvas clássicas: lemniscata de Bernoulli, lemniscata de Gerono, deltoide, astroide, etc. Uma lista de curvas pode ser consultada em (O'CONNOR; ROBERTSON, 2022);
- figuras geométricas: triângulos, retângulos, polígonos regulares, etc..

Deve ser notado que mais que uma curva pode ser empregada (Ver Figura 5.3). Também, não é necessário que as curvas sejam definidas por expressões explícitas ou em formas paramétricas.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

2.) Aplicar transformações geométricas: as transformações geométricas consideradas neste material ficaram restritas às rotações, translações e homotetias (Ver Figura 5.4).

3.) Realizar a escolha do modelo de cores e escolha da paleta de cores. O R básico possui 657 cores disponíveis, as quais podem ser acessadas diretamente por meio da função *color*(). Referências de pacotes que podem ser adicionados podem ser consultados em (ALCOFORADO et al., 2023).

4.) Especificar o padrão de cores: quais cores em cada objeto ou objetos; qual a ordem de aplicação das cores, enfim, especificar todo o processo de coloração. A Figura 5.5 fornece o código e visualização detalhada do processo de coloração.

5.) Composição de uma ou mais figuras: as figuras apresentadas, em geral, são resultados de uma construção inicial composta com homotetias. A Figura 5.5 é ilustrativa.

No entanto, antes de iniciar o processo de construção é necessário instalar o **ggplot2** ((WICKHAM, 2016)) e o **tidyverse**, em específico o *tibble* ((MÜLLER; WICKHAM, 2022)), os dois únicos pacotes utilizados.

```
if(!require(ggplot2)){install.packages("ggplot2")}
library(ggplot2)
if(!require(tidyverse)){install.packages("tidyverse
  ")}
library(tidyverse)
```

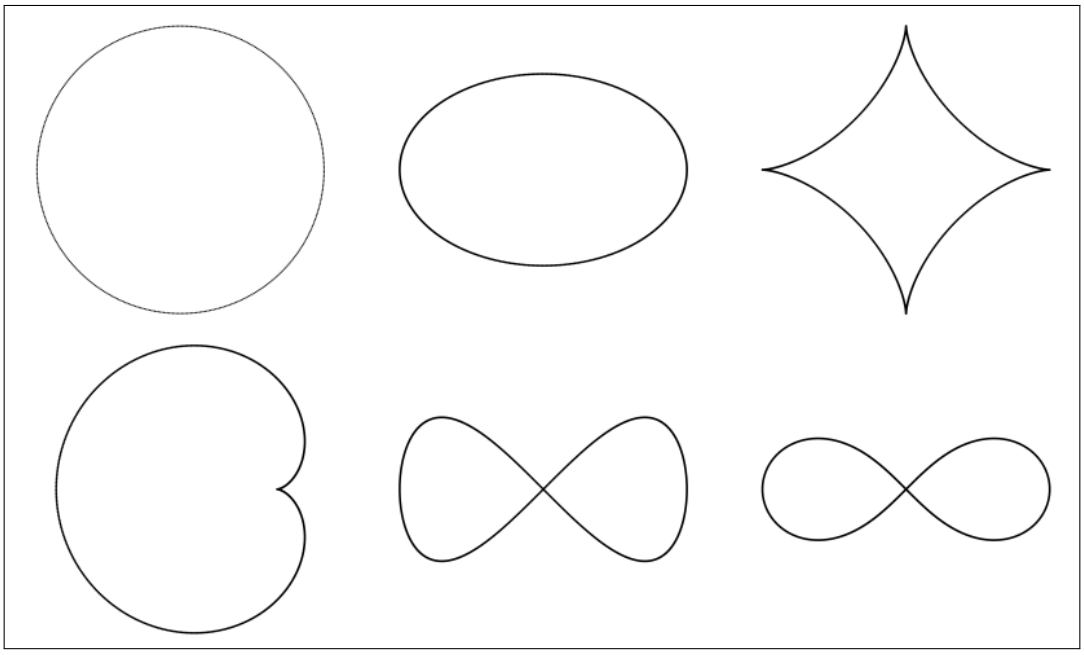
O código a seguir, disponível em (ALCOFORADO et al., 2023), mostra os detalhes da implementação da curva cardioide.

```
require(ggplot2)
```

```
n=700;  theta=seq(0,2*pi, length.out = n); raio=1
x=(1-cos(theta))*cos(theta)*raio
y=(1-cos(theta))*sin(theta)*raio
dt=tibble::tibble(x,y);  size=0.5
p = ggplot()+   coord_fixed()+theme_void()
p = p+geom_point(data=dt, aes(x=x, y=y), color='
   black',size=size)
p
```

Com base no código anterior, é necessário apenas modificar a expressão paramétrica para obter outras curvas; no caso de funções de uma variável ou figuras geométricas, as expressões paramétricas devem ser substituídas de forma conveniente. A Figura 5.2 ilustra circunferência, a elipse, o astroide, o cardioide, a lemniscata de Gerono e a lemniscata de Bernoulli.

Figura 5.2: Curvas planas: circunferência, elipse, astroide, cardioide, lemniscata de Gerono e lemniscata de Bernoulli.

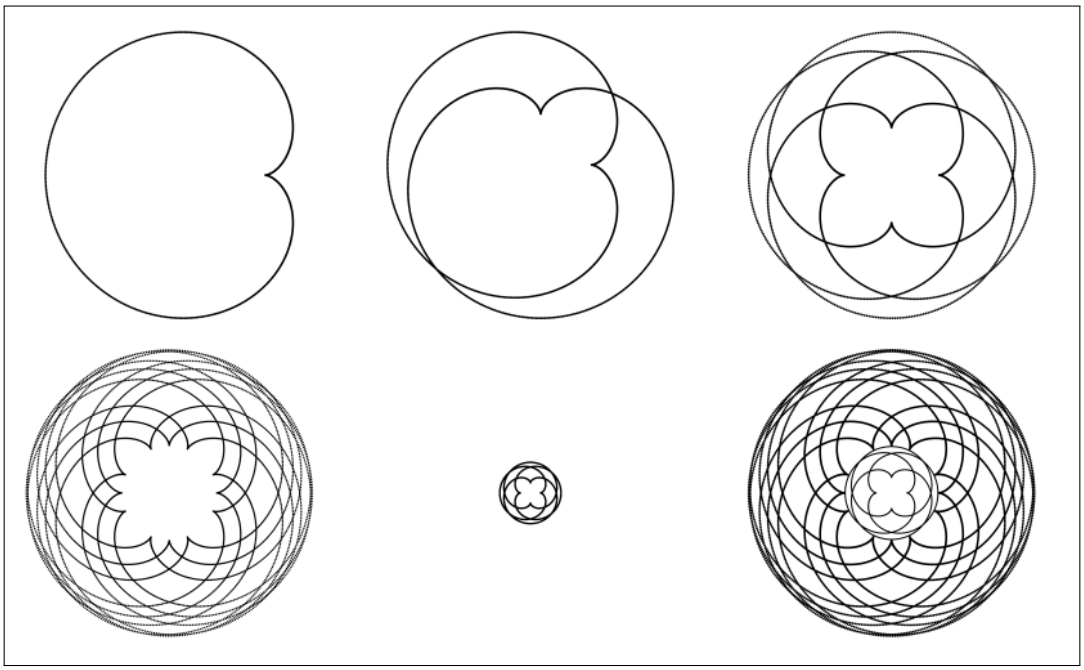


Fonte: Adaptado de (ALCOFORADO et al., 2023).

Agora, a Figura 5.3 ilustra o processo de construção completo, exceto as

colorações.

Figura 5.3: Composição de rotações, e homotetias.



Fonte: Adaptado de (ALCOFORADO et al., 2023).

O item **1.)**, implementado no código anterior, pode ser visualizado na Figura 5.3 (à esquerda e acima). O item **2.)** é completado por meio de uma sequência de operações:

i.) 1 rotação de ângulo $\frac{\pi}{2}$ (ilustrada ao centro da fila 01 de 5.2);

ii.) duas rotações sucessivas de ângulos π e $3\frac{\pi}{2}$ (ilustrada ao fim da fila 01 de 5.3).

Descartando o processo de coloração, a composição de rotações obtida, denominada C_1, é então utilizada diretamente no passo 5.) para obter a construção final por meio de iii.), iv.) e v.) a seguir:

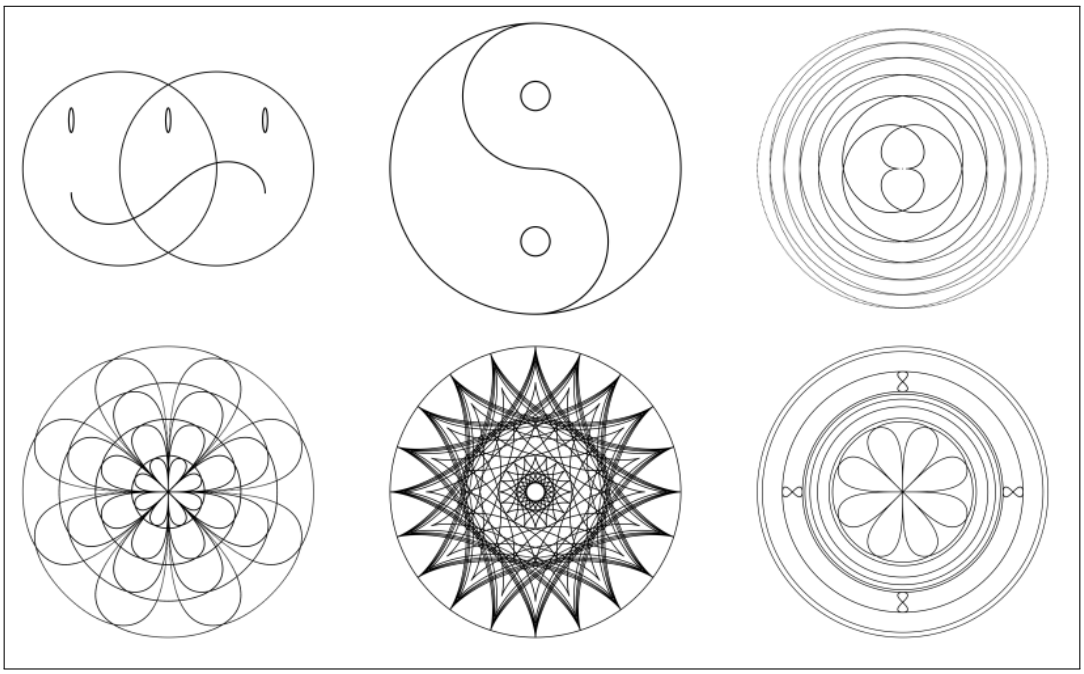
iii.) rotações sucessivas de C_1 por ângulos $\frac{\pi}{8}$ e $-\frac{\pi}{8}$ para obter a composição ilustrada a esquerda e abaixo na Figura 5.3;

iv.) uma homotetia de ii.) para obter a figura ilustrada na segunda fila e ao centro de 5.3;

v.) composição de iii.) e iv.) para obter a composição final, ilustrada a direita e abaixo de 5.3.

Resultados das construções, o passo **v.)**, utilizando uma ou mais curvas planas, são ilustrados na Figura 5.4.

Figura 5.4: Construções com base em curvas planas.



Fonte: Adaptado de (ALCOFORADO et al., 2023).

Os detalhes da figura anterior são descritos a seguir:

- a primeira figura é uma combinação de elipses, circunferências e Lemniscata de Gerono definida para $t \in [0, \pi]$;
- a segunda é uma combinação de circunferências completas e circunferências parciais;
- a terceira figura é uma combinação de espirais;
- a quarta figura é uma combinação de Lemniscatas de Gerono com rotações e homotetias;

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

- a penúltima é uma combinação de deltoides com rotações sucessivas e homotetias;
- por fim, lemniscatas de Gerono com rotações, translações combinadas com rotações e circunferências concêntricas são combinadas em uma única estrutura.

Até o momento as cores não estão em cena pois o destaque é o processo construtivo. Apesar disso, a complexidade emerge das sobreposições dos traçados.

A coloração dos pontos pode ser inserida por meio *colors*(), a qual retorna 657 cores pré-definidas. Cada cor é indexada tanto pelo nome conforme mostrado a seguir, quanto pelo número de identificação na lista ordenada de *colors*().

```
[1] "white"          "aliceblue"      "antiquewhite"
[4] "antiquewhite1" "antiquewhite2" "antiquewhite3"
[7] "antiquewhite4" "aquamarine"     "aquamarine1"
[10] "aquamarine2"   "aquamarine3"    "aquamarine4"
[13] "azure"          "azure1"         "azure2"
[16] "azure3"         "azure4"         "beige"
```

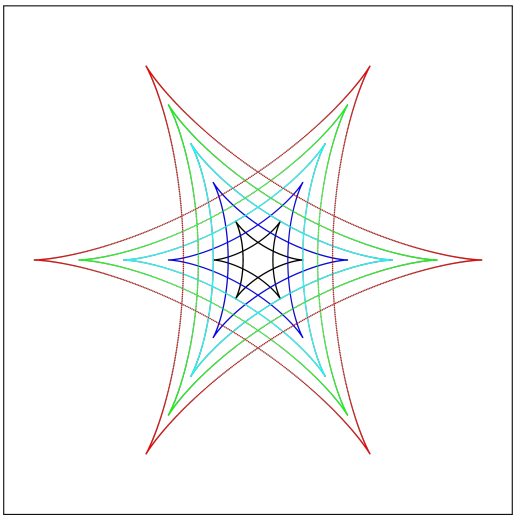
O código a seguir ilustra o método de coloração denominado sequencial. O processo consiste na escolha de um conjunto de cores aplicado de forma sequencial às composições.

```
require(ggplot2)
n=1000; theta=seq(0,2*pi, length.out = n)
x=2*cos(theta)+cos(2*theta);y=2*sin(theta)-sin(2*theta)
z=rep(0,n); dt=tibble::tibble(x,y,z)
step=pi; rotacao=c(seq(0,pi,step)); xt=x; yt=y
p=ggplot()+coord_fixed()+theme_void()
for(i in 1:length(rotacao)){
```

```
    xt=c(xt,x[1:n]*cos(rotacao[i])-y[1:n]*sin(rotacao[i
      ]))
    yt=c(yt,x[1:n]*sin(rotacao[i])+y[1:n]*cos(rotacao[i
      ]))}
red=c(seq(0.2,1,0.2))
cores=c("black","blue","cyan","green","red")
for(i in 1:length(red)){
    xtt=c(xt*red[i]); ytt=c(yt*red[i])
    dt=data.frame(x=c(xt, xtt), y=c(yt, ytt), z="
      astroide")
    p=p+geom_point(data=dt, aes(x=x, y=y),size=0.05,
    color=cores[i])}
p
```

O resultado é a Figura 5.5, um astroide, com identificação das cores empregadas em cada composição e respectivas homotetias. Uma composição, neste caso, consiste do astroide e respectiva rotação.

Figura 5.5: Ilustração do método de coloração denominado sequencial.



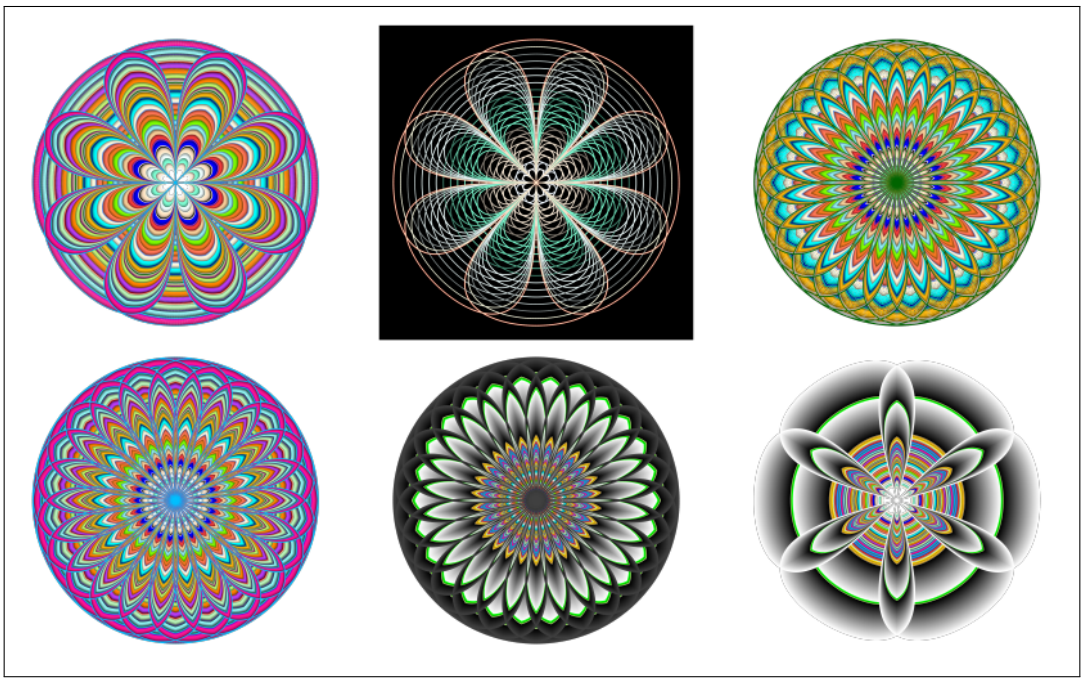
Fonte: O autor.

Note que o resultado final, originado da composição da construção inicial

e respectiva rotação, é colorida com mesma cor. As homotetias da composição são coloridas de cores distintas. Dessa forma, os passos **3.)**, escolha da paleta de cores, e **4.)**, determinação do padrão de coloração, ficam estabelecidos.

A Figura 5.6 mostra algumas possibilidades de construção utilizando lemniscatas com variações do número de rotações, do número de homotetias e o vetor de cores empregado. Detalhes adicionais podem ser consultados em (ALCOFORADO et al., 2023).

Figura 5.6: Construções coloridas utilizando modelo de cores RGB (*Red, Green, Blue*) com o método sequencial.



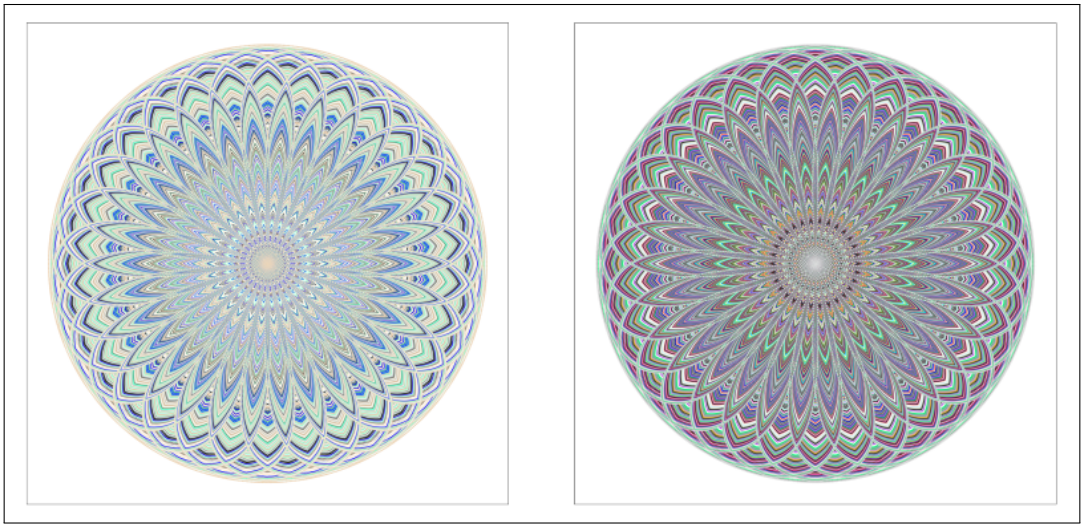
Fonte: Adaptado de (ALCOFORADO et al., 2023).

Em (ALCOFORADO et al., 2023), um outro método de coloração é explorado, o qual envolve realizar amostras aleatórias com reposição do vetor de cores inicialmente escolhido. Neste caso, dado uma lista de cores de interesse, uma amostra com reposição é realizada. O código a seguir ilustra os detalhes de código para a Figura 5.7 (esquerda), enquanto que uma amostra da lista completa de cores é empregada para a Figura 5.7 (direita).

```
step=0.0075
contracao = seq(.0,1.75,by=step);size=0.015
set.seed(10)
cores=sample(colors()[1:27],length(contracao),
  replace=T)
```

O padrão aleatório na escolha da sequência de cores ocasiona resultados com detalhes intrincados. Deve ser notado que as amostras sem reposição também podem ser exploradas, porém aqui não são discutidos ou apresentados.

Figura 5.7: Amostras aleatórias com reposição das cores $colors()[1:27]$ e $colors()$.



Fonte: Adaptado de (ALCOFORADO et al., 2023).

Os resultados apresentados anteriormente não levam em conta cores gradientes ou a coloração para rotações sucessivas em uma composição. Também não levam em conta se a coloração é aplicada nas direções crescentes ou decrescentes das razões das homotetias. Neste contexto, os resultados a seguir contribuem para a expansão das possibilidades de resultados. Esses resultados são distintos daqueles apresentados em (SANTOS; ALCOFORADO, 2023), pois as cores gradientes aqui são aplicadas à composição completa conforme ilustrado na Figura 5.5, enquanto que aqueles aplicam as cores aos pontos da composição, ou seja,

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

cada ponto da composição recebe uma cor, eventualmente, distinta.

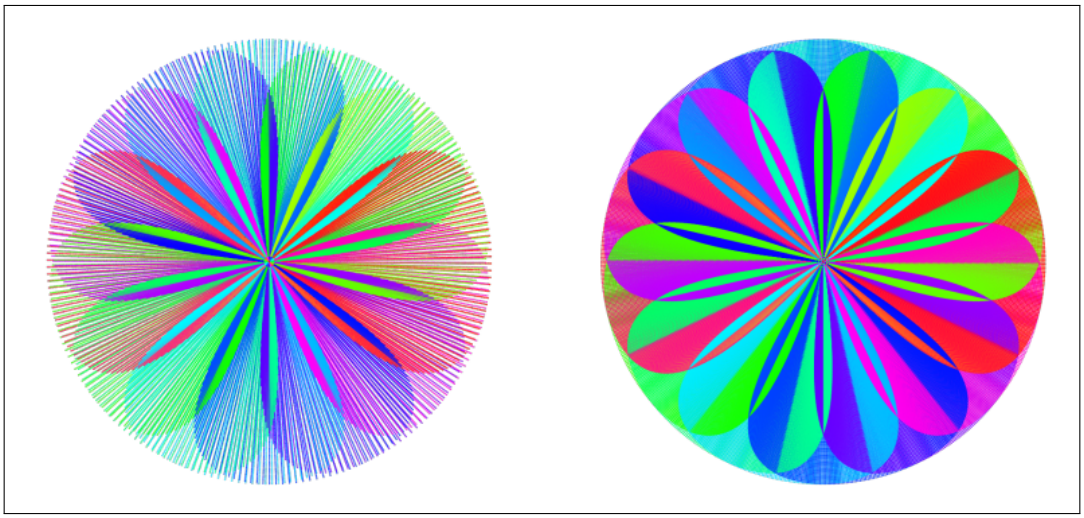
O trecho de código ilustra o mapa gradiente *rainbow*(), cujo resultado é a Figura 5.8 (esquerda).

A Figura 5.8 ilustra o mapa gradiente *rainbow*(), enquanto que a Figura 5.9 ilustra o mapa gradiente *heat.colors*(). A diferença entre as Figuras à esquerda e à direita, para ambas as Figuras 5.8 e 5.9, é apenas a alteração de $n = 200$ para $n = 700$.

O código a seguir ilustra em detalhes a Figura 5.8 (esquerda).

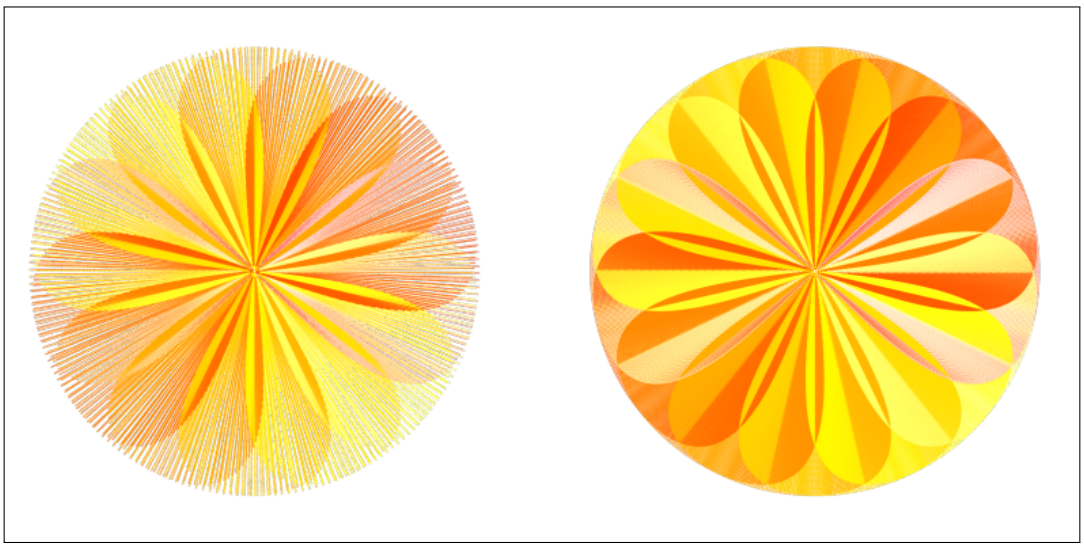
```
nrot=7; n=200
MyColor=rainbow((nrot+2)*length(x), s = 1, v = 1,
   start = 0,
         end = max(1, n - 1)/n, alpha = 1)
rotacao=seq(0.,pi,pi/nrot)
step=0.0075; contracao = seq(.0,1.0,by=step);
```

Figura 5.8: Cores gradientes utilizando o mapa *rainbow*() com distintas quantidades de pontos.



Fonte: O autor.

Figura 5.9: Cores gradientes utilizando o mapa *heat.colors*() com distintas quantidades de pontos.



Fonte: O autor.

A quantidade $(nrot+2)*length(x)$ é o número de pontos em uma das composições, determinado com base na quantidade total de rotações da composição. Esses mapas gradientes, explorados por meio dos mapas *rainbow*() e *heat.colors*(), apenas ilustram as potencialidades. Esse potencial pode ser alavancado levando em conta os que a quantidade de pontos afeta o aspecto visual da figura, bem como a quantidade de rotações. Assim, uma exploração mais detalhada precisa ser conduzida sobre esse método.

Independente do método de coloração que esteja em questão, surge a questão relacionada à direção da razão de homotetia empregada. Nestes casos, direção das razões de homotetias, crescente ou decrescente, referem-se aos valores da sequência de homotetias estarem em ordem crescente ou decrescente, respectivamente.

Considerando uma lista l de cores, é possível ajustar o código de forma que resultados idênticos sejam obtidos tanto para homotetias crescentes ou decrescentes. No entanto, supondo que as cores são aplicadas com mesmo código, conforme aqueles aqui apresentados, então haverá, exceto nos casos em que a lista de cores é aproximadamente simétrica em relação à disposição das cores, mudanças nos

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

resultados obtidos.

A Figura 5.10 explora a dependência da direção das razões das homotetias. A Figura 5.10 (esquerda) ilustra as razões de homotetias $red_c = c(seq(0., 1, 0.0125))$, enquanto que a Figura 5.10 (direita) ilustra as razões de homotetias $red_d = c(seq(1, 0, -0.0125))$. Em ambos os casos, as cores, determinadas por $colors()[i]$, para $i = 1, \cdots, length(red)$, são atribuídas para as homotetias $red_c[i]$ e $red_d[i]$, respectivamente. Em detalhes, o processo de atribuição é sequêncial como segue:

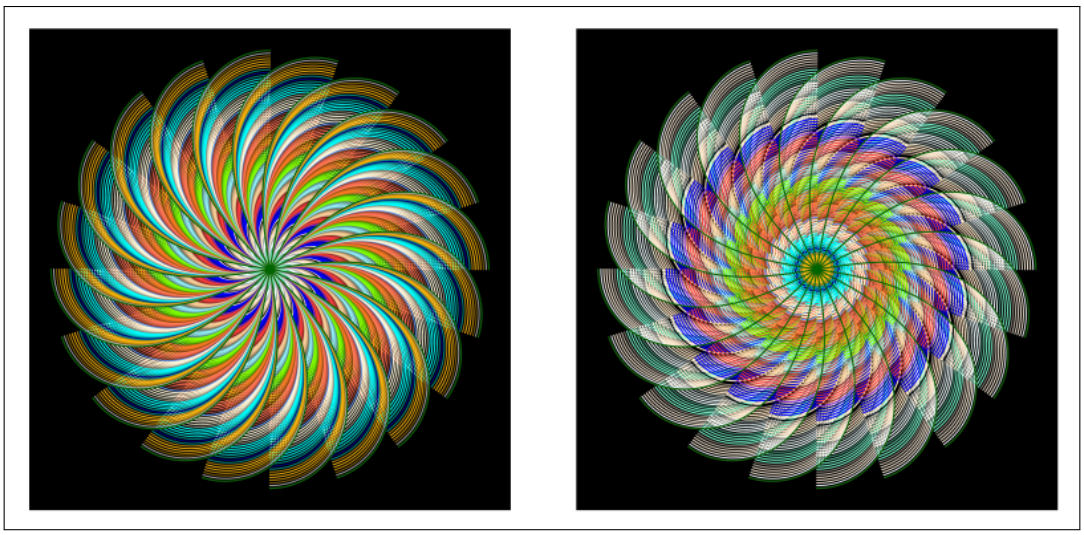
- a primeira cor (*"white"*) é aplicada às homotetias de razões zero ($red_c[1]$) e um ($red_d[1]$), respectivamente;
- a segunda cor (*"aliceblue"*) é aplicada às homotetias de razões ($red_c[2]$) e um ($red_d[2]$), respectivamente;
- analogamente para as demais homotetias.

O código a seguir mostra os detalhes discutidos.

```
require(ggplot2)
n=500; theta=seq(0,pi, length.out = n)
x=cos(theta)/(1+(sin(theta))^2); y=sin(theta)*cos(
  theta)/(1+(sin(theta))^2)
z=rep(0,n); dt=tibble::tibble(x,y,z)
step=pi/10
rotacao=c(seq(0,pi,step));
xt=x; yt=y
p=ggplot()+coord_fixed()+theme_void()
for(i in 1:length(rotacao)){
    xt=c(xt,x[1:n]*cos(rotacao[i])-y[1:n]*sin(
      rotacao[i]))
    yt=c(yt,x[1:n]*sin(rotacao[i])+y[1:n]*cos(
      rotacao[i]))}
```

```
    red=c(seq(0.,1,0.0125))
    for(i in 1:length(red)){
        xtt=c(xt*red[i]);ytt=c(yt*red[i])
        dt=data.frame(x=c(xt, xtt), y=c(yt, ytt), z="
           astroide")
        p=p+geom_point(data=dt, aes(x=x, y=y), color=
           colors()[i],size=0.15)}
    p=p + theme(panel.background = element_rect(fill =
       "black")  )
    p
```

Figura 5.10: Método sequencial com cores aplicadas às homotetias de razão crescente e decrescente, respectivamente.



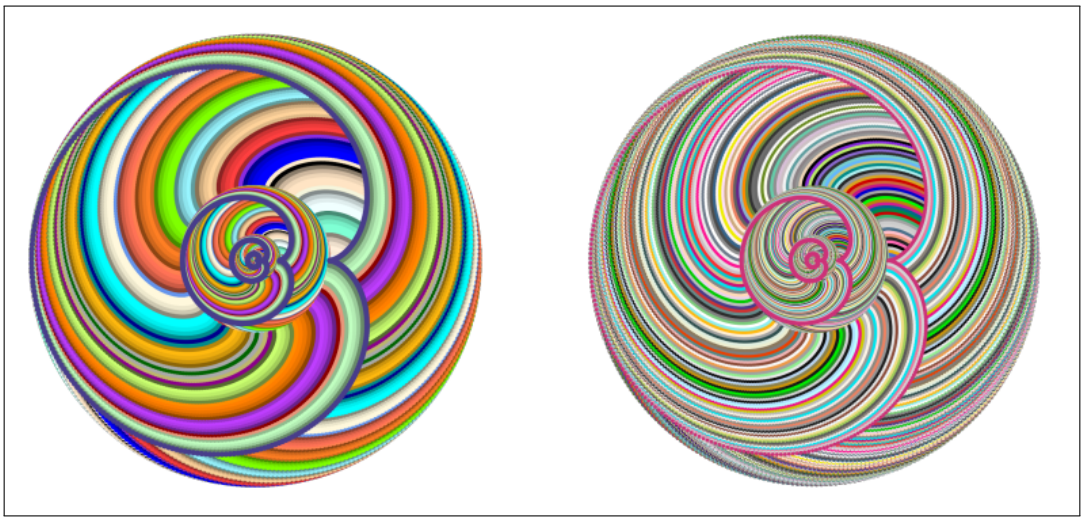
Fonte: O autor.

Um outro aspecto explorado aqui são as colorações para rotações sucessivas. A Figura 5.11 mostra dois exemplos utilizando um cardioide. Para cada rotação do cardioide, duas homotetias com razões determinadas empiricamente, foram adicionadas à composição para obter os resultados. No primeiro caso, rotações sucessivas e homotetias foram utilizadas com as cores determinadas por $colors()[1 : 53]$; no segundo caso, uma amostra aleatória com reposição de 73

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

cores foram obtidas da lista de cores $colors()$. As diferenças são estritamente devido às escolhas das paletas de cores.

Figura 5.11: Rotações sucessivas e respectivas homotetias de um cardioide com cores determinadas em $colors()[1:53]$ e amostra aleatória de 73 elementos com reposição da lista de cores $colors()$.



Fonte: O autor.

Os detalhes são descritos no código a seguir resultam na Figura 5.11 (esquerda); uma alteração do vetor de cores para $cores = sample(colors(), length(rotacao), replace = T)$ resultará na Figura 5.11 (direita).

```
require(ggplot2)
n=400; theta=seq(0,2*pi, length.out = n); raio=1
x=c(2*raio*cos(theta)-raio*cos(2*theta)); y=c(2*
  raio*sin(theta)-raio*sin(2*theta))
dt=tibble::tibble(x,y)
step=pi/53;  rotacao=seq(0,2*pi,step)
xt=x; yt=y
p=ggplot()+coord_fixed()+theme_void()
cores=colors()[1:length(rotacao)]
a=0.325; contracao=c(a,a^2,a^3)
```

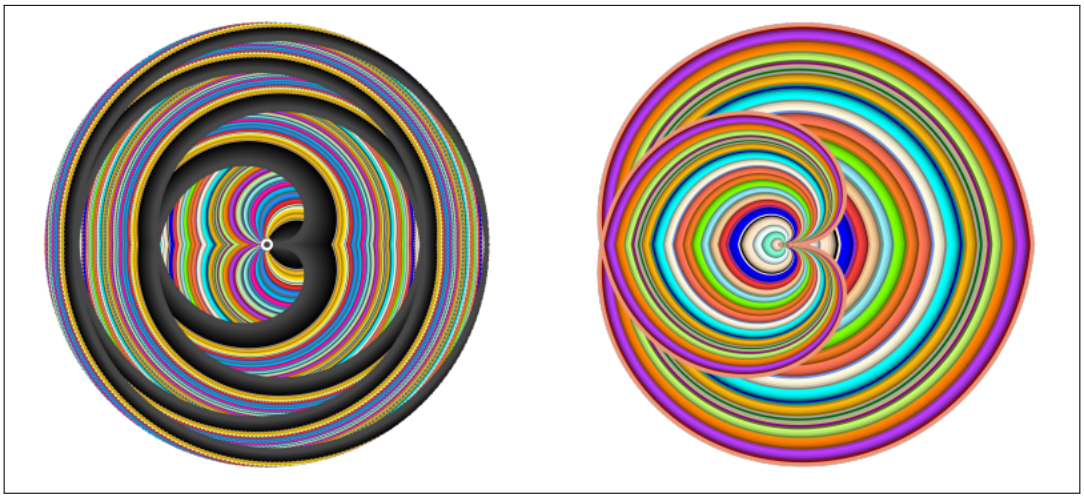
```
    for(i in 1:length(rotacao)){
        xt=c(x[1:n]*cos(rotacao[i])-y[1:n]*sin(rotacao[
          i]))
        yt=c(x[1:n]*sin(rotacao[i])+y[1:n]*cos(rotacao[
          i]))
        dt=tibble::tibble(xt,yt)
      p=p+geom_point(data=dt, aes(x=xt, y=yt), color=
         cores[i],size=.5)
    for(j in 1:length(contracao)){
      xt1=xt*contracao[j]; yt1=yt*contracao[j]
      dt1=tibble::tibble(xt1,yt1)
      p=p+geom_point(data=dt1, aes(x=xt1, y=yt1), color
        =cores[i],size=.5)}   }
      p
```

As Figuras 5.12 e 5.13 mostram resultados para as curvas Espiral de Fermat (apenas parte positiva) e Lemniscata de Bernoulli (Ver (O'CONNOR; ROBERTSON, 2022)). Em 5.12 (esquerda), apenas rotações e reflexões em torno da origem foram utilizadas, enquanto que em 5.12 (direita), homotetias foram incorporadas. As diferenças de construção são devido às homotetias e ao intervalo de varição do parâmetro t da expressão paramétrica, $t \in [0, 4\pi]$ e $t \in [0, 2\pi]$, respectivamente. O código a seguir ilustra os detalhes para a Figura 5.12 (direita).

```
n = 600;  t = seq(0, 2*pi, length.out = n) ; r = 1;
x = r*sqrt(t)*cos(t) ;  y = r*sqrt(t)*sin(t)
dt = data.frame(x,y) ;  steps=90;
rotacao=c(seq(0,2*pi,pi/steps))
xt=x; yt=y
p= ggplot()+coord_fixed()+theme_void()
```

```
  for(i in 1:length(rotacao)){
    xt=c(x[1:n]*cos(rotacao[i])-y[1:n]*sin(rotacao[i
      ]))
    yt=c(x[1:n]*sin(rotacao[i])+y[1:n]*cos(rotacao[i
      ]))
    dt1=tibble::tibble(x=c(xt,xt),y=c(yt,-yt))}
  contracao=c(seq(0,1,.01))
  xt=dt1$x;yt=dt1$y
  for(i in 1:length(rotacao)){
    xt1=xt*contracao[i];yt1=yt*contracao[i]
    dt2=tibble::tibble(xt1,yt1)
    p=p+ geom_point(data=dt2, aes(x=xt1, y=yt1),
      color=colors()[i],size=0.25)
    }
  p
```

Figura 5.12: Rotações sucessivas com cores $colors()[i], i = 1, \cdots, 90$ e $t \in [0, 4\pi]$ (esquerda) e rotações sucessivas com homotetias com $colors()[i], i = 1, \cdots, 90$ e $t \in [0, 2\pi]$ (direita).

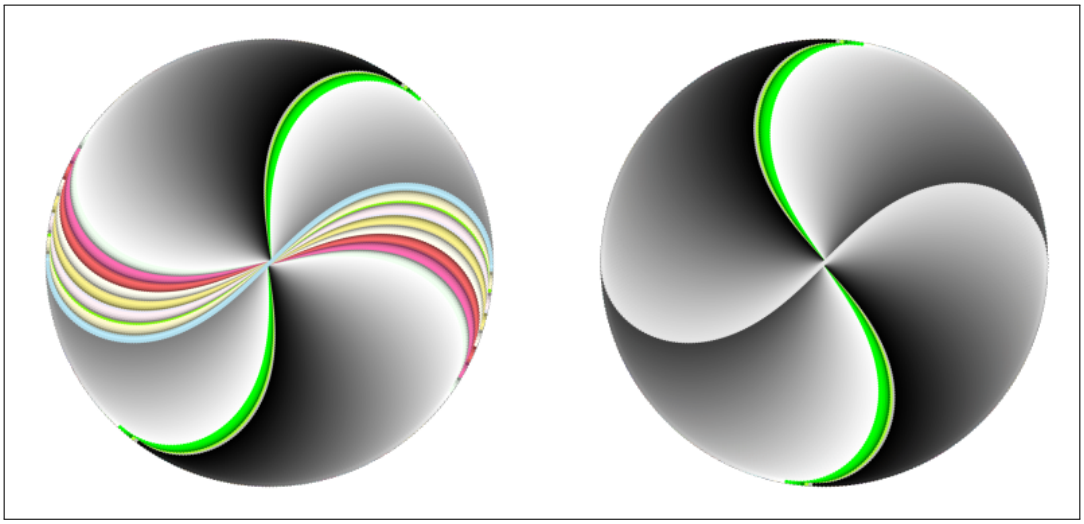


Fonte: O autor.

Resultados adicionais para coloração das rotações sucessivas são mostrados

na Figura 5.13 (esquerda e direita). Ambos os casos consideram a Lemniscata de Bernoulli para $t \in [0, \pi]$.

Figura 5.13: Rotações sucessivas da Lemniscata de Bernoulli para $t \in [0, \pi]$.



Fonte: O autor.

5.4 CONCLUSÕES

A combinação de Matemática, Estatística e Programação permitiu a geração de figuras com aspectos, no mínimo, intrigantes. O entrelaçamento entre as curvas, a sobreposição e combinação das cores ressaltam que a complexidade pode resultar de construções e regras elementares de construção. Essas regras simples, as rotações, as homotetias e as translações expandem a capacidade de utilização do R e, em especial, dos elementos de visualização, propiciados pelo **ggplot2**, e de amostragem aleatória. A complexidade das construções, em princípio, depende da capacidade criativa do indivíduo que, por sua vez, envolve o planejamento e execução da construção inicial, o delineamento das construções sucessivas e o processo de coloração. Este, por sua vez, é sensível as condições e regras, pois pequenas alterações podem gerar efeitos muito distintos. De forma geral, a exploração das construções com simetria radial, levando em conta curvas planas com expressões paramétricas explícitas e as capacidades gráficas do R ilustram um potencial para

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

a combinação de Matemática, Estatística, Programação e Visualização.

5.5 REFERÊNCIAS

ALCOFORADO, Luciane Ferreira et al. **Mandalas, curvas clássicas e visualização com R**. [S.l.]: Universidade de São Paulo. Faculdade de Zootecnia e Engenharia de Alimentos, 2023. Disponível em: https://www.livrosabertos.sibi.usp.br/portaldelivrosUSP/catalog/book/1017. Acesso em 1 agosto 2023. DOI: 10.11606/9786587023335.

MÜLLER, Klaus; WICKHAM, Hadley. **tibble: Simple Data Frames**. [S.l.]: R Project for Statistical Computing, 2022. Disponível em: https://CRAN.R-project.org/package=tibble.

O'CONNOR, J.; ROBERTSON, E. **Famous Curves Index**. [S.l.]: *MacTutor*, 2022. Disponível em: https://mathshistory.st-andrews.ac.uk/Curves/. Accessed 26 October 2022.

SANTOS, João Paulo Martins; ALCOFORADO, Luciane Ferreira. Colorindo mandalas com R: explorando cores e gradientes em curvas planas. In: SILVA, C. e OLIVEIRA, J. X Xornada de Usuarios de R en Galicia. Santiago de Compostela: [s.n.], outubro 2023.

WICKHAM, Hadley. **ggplot2: Elegant Graphics for Data Analysis**. [S.l.]: Springer-Verlag New York, 2016. Disponível em: https://ggplot2.tidyverse.org. ISBN 978-3-319-24277-4.

Capítulo 6

COMO USAR O PACOTE AHPWR PARA A TOMADA DE DECISÃO MULTICRITÉRIO

Autor: Luciane Ferreira Alcoforado [1]
Academia da Força Aérea, Pirassununga, SP
e-mail: lucianea@id.uff.br

Neste capítulo será apresentado uma breve introdução do método de análise hierárquica (AHP) e os principais comandos do pacote AHPWR para realizar o procedimento que vai desde representar a árvore hierárquica do problema, estabelecer os julgamentos necessários para gerar as matrizes paritárias, avaliar a qualidade dos julgamentos e obter os pesos finais das alternativas para auxiliar o tomador de decisão na melhor escolha entre as alternativas.
Palavras-Chave: Tomada de decisão; multicritério; método AHP; pacote AHPWR.

6.1 INTRODUÇÃO

O método AHP (*Analytic Hierarchy Process*) é uma técnica de apoio à decisão multicritério que permite comparar alternativas com base em critérios qualitativos e quantitativos. O método foi proposto por Thomas Saaty na década

[1] Agradecimentos ao Comitê Organizador do VII SER por tornar essa obra possível

de 1970 e consiste em decompor um problema complexo em uma hierarquia de elementos, atribuir pesos aos critérios e às alternativas por meio de comparações pareadas e calcular os escores finais das alternativas. O método AHP também permite verificar a consistência das comparações e realizar análises de sensibilidade, (SAATY, 1980), (SAATY; VARGAS, 2012).

Para facilitar a aplicação do método AHP, foi desenvolvido na linguagem R um pacote denominado AHPWR (*Analytic Hierarchy Process with R*) proposto por (ALCOFORADO; SOUSA; LONGO, 2022), que oferece funções para construir a hierarquia do problema, realizar as comparações pareadas, calcular os pesos e os escores das alternativas, verificar a consistência das comparações e gerar gráficos, auxiliando a produção de relatórios. As funções deste pacote tiveram início durante o projeto de iniciação científica da UFF (Universidade Federal Fluminense) em 2018 e posteriormente foram adequadas e submetidas ao CRAN (*Comprehensive R Archive Network*) em 2022. O pacote AHPWR é uma ferramenta útil para pesquisadores, estudantes e profissionais que desejam utilizar o método AHP em seus projetos de decisão.

6.2 OBJETIVOS

Os objetivos deste capítulo, que versam sobre a apresentação de uma breve introdução ao método AHP e os principais comandos do pacote AHPWR como ferramenta facilitadora da aplicação do método, são:

- Apresentar o conceito e a estrutura do método AHP, bem como suas vantagens e limitações;
- Explicar os passos para a aplicação do método AHP, desde a definição do problema até a obtenção dos resultados;
- Demonstrar o uso do pacote AHPWR para implementar o método AHP na linguagem R, mostrando as funções disponíveis e os exemplos de código;

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

- Ilustrar a aplicação do método AHP e do pacote AHPWR em um caso prático de seleção de alternativas com base em critérios múltiplos.

6.3 APLICAÇÃO

Um problema de decisão multicritério é um problema que envolve a escolha de uma ou mais alternativas com base em vários critérios, que podem ser de natureza qualitativa ou quantitativa. Pode ser formulado da seguinte forma:

- Dado um conjunto de alternativas $A = A_1, A_2, \ldots, A_n$, que representam as possíveis soluções para o problema;
- Dado um conjunto de critérios $C = C_1, C_2, \ldots, C_m$, que representam os atributos ou as dimensões relevantes para avaliar as alternativas;
- Dado um conjunto de pesos $W = W_1, W_2, \ldots, W_m$, que representam a importância relativa dos critérios;
- Dada uma função de avaliação $f : A \times C \rightarrow R$, que atribui um valor numérico para cada alternativa em relação a cada critério;
- Encontrar uma ou mais alternativas que maximizem (ou minimizem) uma função de agregação $g : A \rightarrow R$, que combina os valores das alternativas em relação aos critérios e aos pesos.

Um exemplo de problema de decisão multicritério é a escolha da profissão, considerando os critérios como afinidade, relação candidato-vaga e mercado de trabalho. Cada critério pode ter um peso diferente na preferência do indivíduo e cada profissão pode ter um valor diferente em relação a cada critério. O objetivo é encontrar a profissão que melhor atenda às expectativas do indivíduo.

6.3.1 A estrutura do método AHP

O método AHP é uma das abordagens possíveis para auxiliar o tomador de decisão. Começa por dividir os componentes de um problema em uma estrutura

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

hierárquica. Depois, realiza-se comparações aos pares entre os componentes de um mesmo nível, considerando o critério do nível acima. Essas comparações definem as prioridades e, por meio de uma síntese, as prioridades globais. Avalia-se também a consistência e o tratamento da interdependência. Esses são os passos básicos do método.

A estrutura do método AHP envolve algumas etapas especificadas a seguir:

- Definição do problema e do objetivo (*goal*) da decisão;
- Construção da hierarquia do problema (*hierarchical tree of decision*), identificando os critérios (*criteria*) e as alternativas (*choices*) relevantes;
- Realização das comparações pareadas entre os elementos da hierarquia, usando uma escala numérica ou verbal;
- Cálculo dos pesos e dos escores das alternativas, bem como da consistência das comparações.

As vantagens do método AHP permitem:

- Lidar com problemas complexos e multidimensionais de forma estruturada e sistemática;
- Incorporar tanto dados objetivos quanto julgamentos subjetivos na tomada de decisão;
- Avaliar a consistência das comparações e corrigir as inconsistências;
- Realizar análises de sensibilidade para testar a robustez dos resultados.

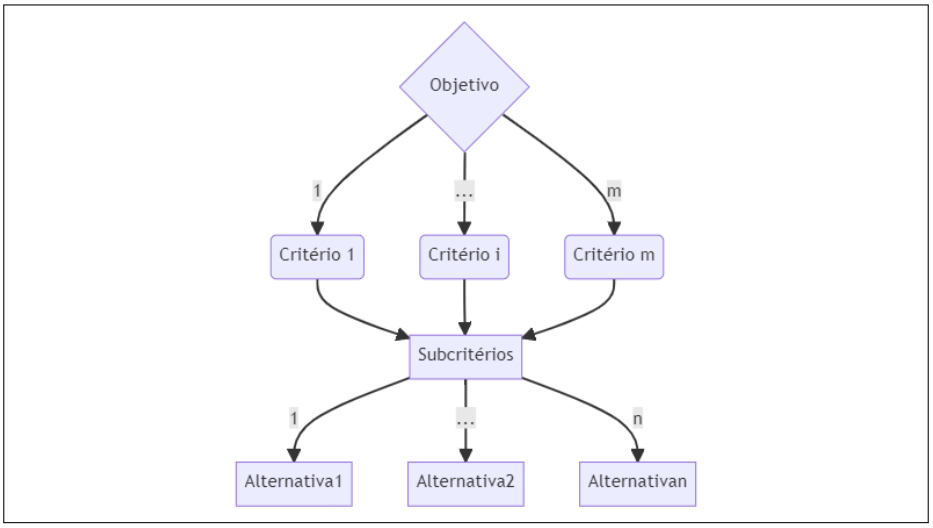
As limitações do método AHP estão ligadas as seguintes requisições:

- Um número elevado de comparações pareadas quando o número de elementos da hierarquia é grande, o que pode gerar fadiga ou erro nos julgamentos;

- Um conhecimento prévio dos critérios e das alternativas, o que pode limitar a criatividade ou a inovação na tomada de decisão;
- Um software específico para realizar os cálculos e as análises, o que pode dificultar o acesso ou a compreensão dos usuários.

Resumidamente, podemos dizer que a estrutura hierárquica do problema é composta de três níveis, conforme Figura 6.1:

Figura 6.1: Estrutura Hierárquica de um problema.



Fonte: A autora.

- Nível 1: Objetivo (*goal*) a ser atingido.
- Nível 2: Critérios (*criteria*) a serem considerados. Cada critério pode ser subdividido em subcritérios, que representam aspectos mais específicos ou detalhados do critério.
- Nível 3: Alternativas (*choices*) a serem ponderadas.

A possibilidade de considerar subcritérios na formulação do problema, permite um maior refinamento e precisão na análise dos critérios, mas também aumenta a complexidade e o número de comparações necessárias para o método

AHP. Os subcritérios podem ter pesos diferentes dentro do critério, e devem ser comparados entre si em relação às alternativas. O pacote AHPWR lida com problemas desse tipo, maiores detalhes devem ser vistos na documentação do pacote, consultando o tutorial elaborado por (ALCOFORADO; LONGO, 2022).

Após a estruturação do problema, o próximo passo é obter os julgamentos que podem ser feitos por um especialista ou um grupo de especialistas.

Por simplificação vamos considerar um único julgador. Caso haja mais de um, cada julgador deverá realizar o mesmo procedimento.

Será necessário obter $m+1$ matrizes paritárias, sendo m o número de critérios do problema.

A primeira matriz será $m \times m$ comparando os m critérios par a par; as demais matrizes serão $n \times n$ comparando as n alternativas à luz de cada critério.

Para realizar as comparações, (T. L. Saaty 1980) propôs uma escala, conhecida como escala fundamental de Saaty, conforme Figura 8.2.

Figura 6.2: Escala Fundamental de Saaty.

Escala numérica	Escala Conceitual	Descrição
1	**Igual**	Os dois elementos comparados contribuem igualmente para o objetivo.
3	**Moderada**	O elemento comparado é ligeiramente importante ao outro.
5	**Forte**	A experiência e o julgamento favorecem fortemente o elemento em relação ao outro.
7	**Muito Forte**	O elemento comparado é muito mais forte em relação ao outro, e tal importância pode ser observada na prática.
9	**Absoluta**	O elemento comparado apresenta o mais alto nível de evidência possível a seu favor.
2,4,6,8	**Valores intermediários entre dois julgamentos, utilizados quando o decisor sentir dificuldade ao escolher entre dois graus de importância adjacentes.**	

Fonte: Adaptado de (SAATY, 1980).

Outra possibilidade para construir os julgamentos é utilizando a proposta de (GODOI, 2014) que consiste no método do julgamento holístico que visa dar auxílio à construção da matriz de julgamentos paritários. O método propõe uma regra de atribuição de pesos aos elementos da hierarquia, usando uma escala de 0 a 10 (ou de 0 a 100), de forma que todos os itens a serem avaliados sejam

colocados lado a lado, cabendo ao julgador atribuir um peso relativo de forma a ser possível ordená-los, do menos importante ao mais importante. Após atribuído os pesos $w_1, ..., w_k$ aos k itens a serem comparados, são aplicados fórmulas para a obtenção da matriz paritária, fazendo:

$$a_{ij} = w_i - w_j + 1$$

, se $w_i \geq w_j$ (isto é, se o item i tem importância maior ou igual ao item j); caso contrário

$$a_{ij} = \frac{1}{(w_j - w_i + 1)}$$

, se $w_i < w_j$.

Todas as matrizes paritárias devem ser verificadas quanto a sua consistência, isto é, deve ser realizado o teste de consistência.

O teste de consistência de Saaty é uma forma de verificar se os julgamentos paritários feitos pelo decisor no método AHP são coerentes e confiáveis. O teste consiste em calcular a razão de consistência (RC) dos julgamentos, que é uma medida que compara o índice de consistência (IC) dos julgamentos com o índice aleatório (IR) esperado para uma matriz de julgamentos aleatórios. A fórmula para calcular a RC é:

$$RC = \frac{IR}{IC}$$

Onde:
IC é o índice de consistência, que é dado por:

$$IC = \frac{\lambda_{max} - n}{n - 1}$$

λ_{max} é o maior autovalor da matriz de julgamentos;
n é a ordem da matriz de julgamentos;
IR é o índice aleatório, que é obtido a partir de uma tabela fornecida por Saaty,

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

consultar (SAATY, 1980).

O teste de consistência de Saaty considera que os julgamentos são consistentes e aceitáveis se a RC for menor ou igual a 0,1. Caso contrário, os julgamentos devem ser revisados pelo decisor para eliminar as inconsistências.

O peso de cada elemento da hierarquia é obtido após o teste de consistência usando o método dos autovetores. Esse método consiste em calcular o autovetor associado ao maior autovalor da matriz de julgamentos. O autovetor representa os pesos relativos dos elementos em relação ao critério do nível superior. Para obter os pesos normalizados, basta dividir cada elemento do autovetor pela soma de todos os elementos. Os pesos normalizados devem somar 1 e refletir as preferências do decisor.

6.3.2 Como utilizar o pacote AHPWR

O pacote auxilia na aplicação do método AHP com funções criadas para facilitar sua implementação.

Vamos considerar dois exemplos básicos. No primeiro o objetivo é a escolha da profissão e no segundo caso o objetivo é a escolha do método construtivo de uma ponte.

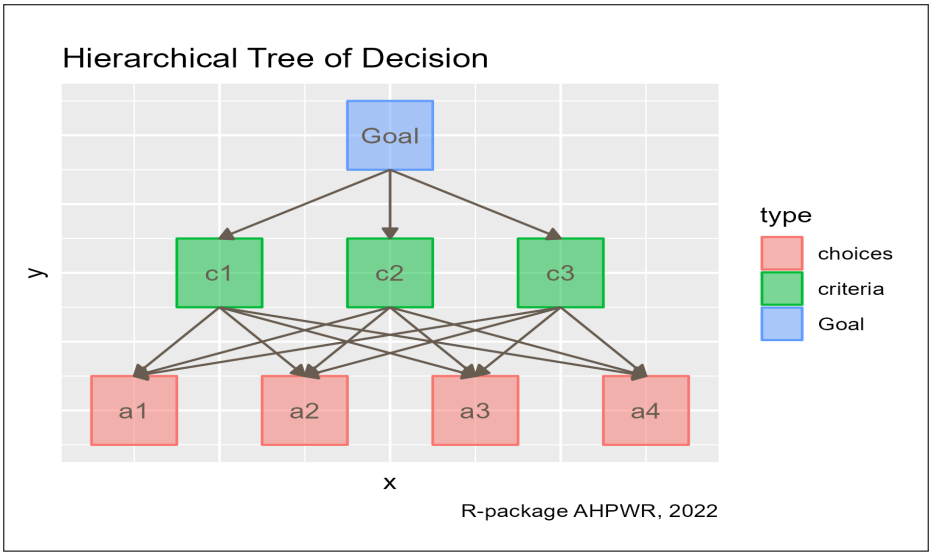
6.3.2.1 Exemplo 1

Objetivo: Escolher a profissão
Critérios para a tomada de decisão: C1 - afinidade; C2 - mercado de trabalho; C3 - Relação Candidato-vaga.
Alternativas: A1 - Estatística; A2 - Engenharia Civil; A3 - Computação; A4 - Administração.

A árvore hierárquica é obtida rapidamente utilizando-se o código a seguir, cujo resultado pode ser visto na Figura 8.1:

```
AHPWR::flow_chart(names=NULL, c=3, a=4)
```

Figura 6.3: Árvore Hierárquica do Exemplo 1.



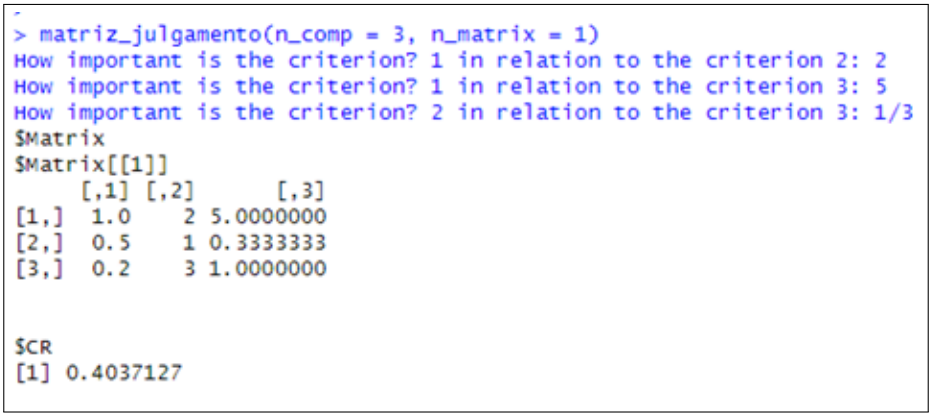
Fonte: A autora.

Como temos 3 critérios e 4 alternativas, devemos obter 4 matrizes paritárias sendo M1 - matriz 3×3 comparando os 3 critérios; M2 - matriz 4×4 comparando as 4 alternativas sob a luz do critério 1; M3 - matriz 4×4 comparando as 4 alternativas sob a luz do critério 2; M4 - matriz 4×4 comparando as 4 alternativas sob a luz do critério 3.

A função **matriz_julgamento()** produz perguntas no console do R, conduzindo o julgador a realizar todas as comparações necessárias no modo tradicional do julgamento proposto por Saaty. Cada pergunta deve ser respondida a nível de console, ao final é gerada a matriz e o resultado do teste de consistência, conforme Figura 8.3.

```
AHPWR::matriz_julgamento(n_comp = 3, n_matrix = 1)
```

Figura 6.4: Console do R para obter matriz de julgamento.



Fonte: A autora.

Vemos que o $CR < 0.1$, indicando uma inconsistência nos julgamentos, ou seja, os julgamentos devem ser refeitos.

Vamos refazer os julgamentos, aplicando agora o método proposto por (GO-DOI, 2014). Agora, o julgador deve atribuir os pesos relativos da importância dos critérios. Supondo que foi atribuído $w_1 = 7, w_2 = 1$ *e* $w_3 = 2$, ou seja, o critério 1 foi julgado o mais importante, depois o critério 3 bem menos importante que o 1 e por fim o critério 2 um pouco menos importante que o 3.

```
x=paste0("C",1:3) #nomes dos critérios C1, C2, C3
y=c(7, 1, 2) #julgamento holístico dos critérios
m1=AHPWR::matrix_ahp(x,y) #matriz paritária de
  julgamentos dos critérios
m1
```

```
          C1 C2  C3
C1 1.0000000  7 6.0
C2 0.1428571  1 0.5
C3 0.1666667  2 1.0
```

```
AHPWR::CR(m1)
```

```
[1] 0.02790226
```

Para cada critério do nível 2 o julgador deve realizar a comparação das alternativas do nível 3.

Considerando o **critério 1**, a afinidade na escolha da profissão, o julgador deve atribuir os pesos para cada curso. supondo que foi atribuído $w_1 = 4, w_2 = 5, w_3 = 3$ *e* $w_4 = 2$, ou seja, a alternativa 2 foi julgado mais importante, depois a alternativa 1 um pouco menos importante seguida da alternativa 3 e a menos importante foi a alternativa 4. Esses julgamentos possibilitam a criação da matriz m_2:

```
x=paste0("A",1:4) #nomes das alternativas A1, A2, A3,
  A4
y=c(4, 5, 3, 2) #julgamento holístico das alternativas
  para o critério 1
m2=AHPWR::matrix_ahp(x,y)
m2
```

```
          A1        A2  A3 A4
A1 1.0000000 0.5000000 2.0  3
A2 2.0000000 1.0000000 3.0  4
A3 0.5000000 0.3333333 1.0  2
A4 0.3333333 0.2500000 0.5  1
```

```
AHPWR::CR(m2)
```

```
[1] 0.01147537
```

Considerando o **critério 2**, o mercado de trabalho na escolha da profissão, o julgador deve atribuir os pesos para cada curso. supondo que foi atribuído

$w_1 = 2, w_2 = 4, w_3 = 3$ *e* $w_4 = 7$, ou seja, a alternativa 4 foi julgada mais importante, depois a alternativa 2 menos importante seguida da alternativa 3 e a menos importante foi a alternativa 1. Esses julgamentos possibilitam a criação da matriz m_3:

```
x=paste0("A",1:4) #nomes das alternativas A1, A2, A3,
  A4
y=c(2, 4, 3, 7) #julgamento holístico das alternativas
  para o critério 2
m3=AHPWR::matrix_ahp(x,y)
m3
```

```
   A1        A2  A3        A4
A1  1 0.3333333 0.5 0.1666667
A2  3 1.0000000 2.0 0.2500000
A3  2 0.5000000 1.0 0.2000000
A4  6 4.0000000 5.0 1.0000000
```

```
AHPWR::CR(m3)
```

```
[1] 0.02436116
```

Considerando o **critério 3**, a relação candidato-vaga na escolha da profissão, o julgador deve atribuir os pesos para cada curso. supondo que foi atribuído $w_1 = 4.9, w_2 = 5, w_3 = 3.3$ *e* $w_4 = 1$, ou seja, a alternativa 2 foi julgada mais importante, depois a alternativa 1 um pouco menos importante seguida da alternativa 3 e a menos importante foi a alternativa 4. Esses julgamentos possibilitam a criação da matriz m_4:

```
x=paste0("A",1:4) #nomes das alternativas A1, A2, A3,
  A4
```

```
y=c(4.9, 5, 3.3, 1) #julgamento holístico das
  alternativas para o critério 3
m4=AHPWR::matrix_ahp(x,y)
m4
```

```
          A1        A2        A3  A4
A1 1.0000000 0.9090909 2.6000000 4.9
A2 1.1000000 1.0000000 2.7000000 5.0
A3 0.3846154 0.3703704 1.0000000 3.3
A4 0.2040816 0.2000000 0.3030303 1.0
```

```
AHPWR::CR(m4)
```

```
[1] 0.01529812
```

Após o estabelecimento dos julgamentos, devemos organizar todas as matrizes numa base de dados.

O pacote AHPWR oferece duas opções, podemos organizar todas as matrizes em uma lista ou podemos organizar todas as matrizes em um arquivo de planilhas de tal forma que cada planilha corresponda a cada matriz.

```
base = list(m1, m2, m3, m3)
base
```

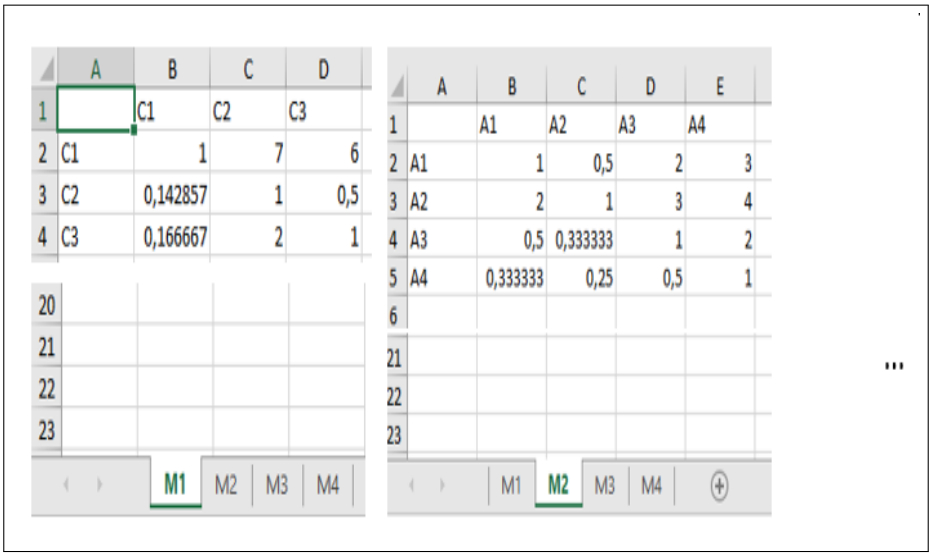
```
[[1]]
          C1 C2  C3
C1 1.0000000  7 6.0
C2 0.1428571  1 0.5
C3 0.1666667  2 1.0
```

```
[[2]]
          A1        A2  A3 A4
A1 1.0000000 0.5000000 2.0  3
A2 2.0000000 1.0000000 3.0  4
A3 0.5000000 0.3333333 1.0  2
A4 0.3333333 0.2500000 0.5  1

[[3]]
   A1        A2  A3        A4
A1  1 0.3333333 0.5 0.1666667
A2  3 1.0000000 2.0 0.2500000
A3  2 0.5000000 1.0 0.2000000
A4  6 4.0000000 5.0 1.0000000

[[4]]
   A1        A2  A3        A4
A1  1 0.3333333 0.5 0.1666667
A2  3 1.0000000 2.0 0.2500000
A3  2 0.5000000 1.0 0.2000000
A4  6 4.0000000 5.0 1.0000000
```

```
#Organizando em arquivo denominado Exemplo_1.xlsx
file1 = AHPWR::xlsx_ahp(m1, file = "Exemplo_1.xlsx",
  sheet = "M1", append = FALSE)
file2 = AHPWR::xlsx_ahp(m2, file = "Exemplo_1.xlsx",
  sheet = "M2", append = TRUE)
file3 = AHPWR::xlsx_ahp(m3, file = "Exemplo_1.xlsx",
  sheet = "M3", append = TRUE)
file4 = AHPWR::xlsx_ahp(m4, file = "Exemplo_1.xlsx",
  sheet = "M4", append = TRUE)
```

Ao optar por armazenar as matrizes no arquivo Exemplo_1.xlsx, o mesmo deve ter sido criado na pasta de trabalho corrente, após rodar os comandos acima, Figura 7.4.

Figura 6.5: Arquivo contendo as matrizes de julgamento em planilhas.



Fonte:A autora.

Para obter o resultado final do método, isto é, os pesos globais das alternativas, utilizamos a função **ahp_geral()**:

```
#Nomes das alternativas no vetor x
x
```

```
[1] "A1" "A2" "A3" "A4"
```

```
#aplicando o método na base de dados
AHPWR::ahp_geral(base, nomes_alternativas = x)
```

```
# A tibble: 4 x 7
  Criteria        Weights      A1      A2      A3      A4      CR
```

```
  <chr>              <dbl>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 ---Alternatives    1       0.228   0.403  0.151  0.218  0.0279
2 --C1               0.758   0.210   0.354  0.121  0.0724 0.0115
3 --C2               0.0905  0.00673 0.0182 0.0109 0.0547 0.0244
4 --C3               0.151   0.0112  0.0304 0.0183 0.0914 0.0244
```

Podemos formatar a tabela de resultados, através do comando $formata\_tabela()$, que possui três versões, sendo a versão padrão exibida na Figura 6.6:

```
tabela = AHPWR::ahp_geral(base, nomes_alternativas = x)
AHPWR::formata_tabela(tabela)
```

Figura 6.6: Tabela Final com os pesos globais, versão padrão.

Criteria	Weights	A1	A2	A3	A4	CR
--- Alternatives	100%	22.81%	**40.29%**	15.06%	21.84%	2.79%
--C1	75.82%	21.02%	35.43%	12.14%	7.24%	1.15%
--C2	9.05%	0.67%	1.82%	1.09%	5.47%	2.44%
--C3	15.12%	1.12%	3.04%	1.83%	9.14%	2.44%

Fonte: A autora.

As versões *GRAY* e *WHITE* podem ser conferidas em 6.7 e 6.8, respectivamente.

```
AHPWR::formata_tabela(tabela, cores = "GRAY")
AHPWR::formata_tabela(tabela, cores = "WHITE")
```

Figura 6.7: Tabela Final com os pesos globais, versão GRAY.

Criteria	Weights	A1	A2	A3	A4	CR
---	100%	22.81%	**40.29%**	15.06%	21.84%	2.79%
Alternatives						
--C1	75.82%	21.02%	35.43%	12.14%	7.24%	1.15%
--C2	9.05%	0.67%	1.82%	1.09%	5.47%	2.44%
--C3	15.12%	1.12%	3.04%	1.83%	9.14%	2.44%

Fonte: A autora.

Figura 6.8: Tabela Final com os pesos globais, versão WHITE.

Criteria	Weights	A1	A2	A3	A4	CR
---	100%	22.81%	**40.29%**	15.06%	21.84%	2.79%
Alternatives						
--C1	75.82%	21.02%	35.43%	12.14%	7.24%	1.15%
--C2	9.05%	0.67%	1.82%	1.09%	5.47%	2.44%
--C3	15.12%	1.12%	3.04%	1.83%	9.14%	2.44%

Fonte: A autora.

6.3.2.2 Exemplo 2

Objetivo: Escolher o método construtivo de uma ponte

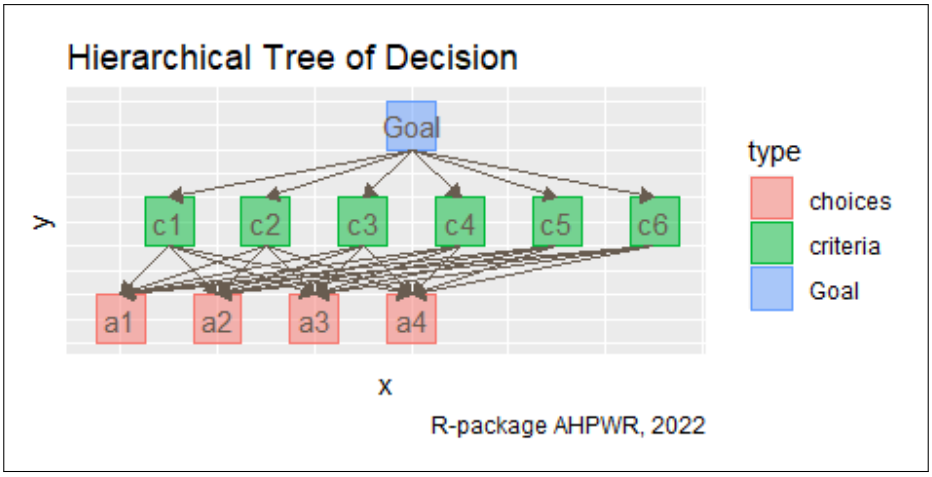
Critérios para a tomada de decisão: C1 - Custo de Construção; C2 - Desenho; C3 - Custo de Manutenção; C4 - Resistência ao Vento; C5 - Resistência ao Tráfego; C6 - Vida Útil.

Alternativas: A1 - Ponte Pênsil; A2 - Ponte Estaiada; A3 - Ponte Treliçada; A4 - Ponte em Arco.

A árvore hierárquica é obtida rapidamente utilizando-se o código a seguir, cujo resultado pode ser visto na Figura 6.9:

```
AHPWR::flow_chart(names=NULL, c=6, a=4)
```

Figura 6.9: Árvore Hierárquica do Exemplo 2.



Fonte: A autora.

Como temos 6 critérios e 4 alternativas, devemos obter 7 matrizes paritárias sendo M1 - matriz 6×6 comparando os 6 critérios; M2 - matriz 4×4 comparando as 4 alternativas sob a luz do critério 1; M3 - matriz 4×4 comparando as 4 alternativas sob a luz do critério 2; M4 - matriz 4×4 comparando as 4 alternativas sob a luz do critério 3; M5 - matriz 4×4 comparando as 4 alternativas sob a luz do critério 4; M6 - matriz 4×4 comparando as 4 alternativas sob a luz do critério 5; M7 - matriz 4×4 comparando as 4 alternativas sob a luz do critério 6.

Vamos aplicar o julgamento holístico proposto por (Godoi 2014). Agora, o julgador deve atribuir os pesos relativos da importância dos critérios. Supondo que foi atribuído $w_1 = 7, w_2 = 9, w_3 = 2.5, w_4 = 10, w_5 = 9.5$ *e* $w_6 = 2$, ou seja, o critério 4 foi julgado o mais importante, depois o critério 5 e assim sucessivamente até o menos importante que foi o critério 6.

```
x=paste0("C",1:6) #nomes dos critérios C1, C2, ..., C6
y=c(7, 9, 2.5, 10, 9.5, 2) #julgamento holístico dos
  critérios
m1=AHPWR::matrix_ahp(x,y) #matriz paritária de
  julgamentos dos critérios
m1
```

```
          C1        C2        C3        C4        C5  C6
C1 1.0000000 0.3333333 5.5000000 0.2500000 0.2857143 6.0
C2 3.0000000 1.0000000 7.5000000 0.5000000 0.6666667 8.0
C3 0.1818182 0.1333333 1.0000000 0.1176471 0.1250000 1.5
C4 4.0000000 2.0000000 8.5000000 1.0000000 1.5000000 9.0
C5 3.5000000 1.5000000 8.0000000 0.6666667 1.0000000 8.5
C6 0.1666667 0.1250000 0.6666667 0.1111111 0.1176471 1.0
```

```
AHPWR::CR(m1)
```

```
[1] 0.03020855
```

Para cada critério do nível 2 o julgador deve realizar a comparação das alternativas do nível 3.

Considerando o **critério 1**, o custo de construção da ponte, o julgador deve atribuir os pesos para cada tipo. Supondo que foi atribuído $w_1 = 4, w_2 = 5, w_3 = 4.5$ *e* $w_4 = 4.9$, ou seja, a alternativa 2 foi julgado mais importante, ou seja, seu custo é o mais importante a ser considerado na tomada de decisão, é o que deverá pesar mais nesse quesito. Esses julgamentos possibilitam a criação da matriz m_2:

```
x=paste0("A",1:4) #nomes das alternativas A1, A2, A3,
  A4
```

```
y=c(4, 5, 4.5, 4.9) #julgamento holístico das
  alternativas para o critério 1
m2=AHPWR::matrix_ahp(x,y)
m2
```

```
    A1        A2        A3        A4
A1 1.0 0.5000000 0.6666667 0.5263158
A2 2.0 1.0000000 1.5000000 1.1000000
A3 1.5 0.6666667 1.0000000 0.7142857
A4 1.9 0.9090909 1.4000000 1.0000000
```

```
AHPWR::CR(m2)
```

```
[1] 0.0006140235
```

Considerando o **critério 2**, o desenho da ponte, o julgador deve atribuir os pesos para cada alternativa. Supondo que foi atribuído $w_1 = 2, w_2 = 4, w_3 = 3$ *e* $w_4 = 5$, ou seja, a alternativa 4 foi julgada mais importante, depois a alternativa 2 menos importante seguida da alternativa 3 e a menos importante foi a alternativa 1. Esses julgamentos possibilitam a criação da matriz m_3:

```
x=paste0("A",1:4) #nomes das alternativas A1, A2, A3,
  A4
y=c(2, 4, 3, 5) #julgamento holístico das alternativas
  para o critério 2
m3=AHPWR::matrix_ahp(x,y)
m3
```

```
   A1        A2  A3        A4
A1  1 0.3333333 0.5 0.2500000
```

```
A2  3 1.0000000 2.0 0.5000000
A3  2 0.5000000 1.0 0.3333333
A4  4 2.0000000 3.0 1.0000000
```

```
AHPWR::CR(m3)
```

```
[1] 0.01147537
```

Considerando o **critério 3**, custo de manutenção, o julgador deve atribuir os pesos para cada alternativa. Supondo que foi atribuído $w_1 = 3, w_2 = 3.1, w_3 = 3.3$ *e* $w_4 = 3.5$, esses julgamentos possibilitam a criação da matriz m_4:

```
x=paste0("A",1:4) #nomes das alternativas A1, A2, A3,
  A4
y=c(3, 3.1, 3.3, 3.4) #julgamento holístico das
  alternativas para o critério 3
m4=AHPWR::matrix_ahp(x,y)
m4
```

```
    A1        A2        A3        A4
A1 1.0 0.9090909 0.7692308 0.7142857
A2 1.1 1.0000000 0.8333333 0.7692308
A3 1.3 1.2000000 1.0000000 0.9090909
A4 1.4 1.3000000 1.1000000 1.0000000
```

```
AHPWR::CR(m4)
```

```
[1] 3.160346e-05
```

Considerando o **critério 4**, resistência ao vento, o julgador deve atribuir os pesos para cada alternativa. Supondo que foi atribuído $w_1 = 2, w_2 = 5, w_3 = 4$ *e* $w_4 = 1$, esses julgamentos possibilitam a criação da matriz m_5:

```
x=paste0("A",1:4) #nomes das alternativas A1, A2, A3,
  A4
y=c(2, 5, 4, 1) #julgamento holístico das alternativas
  para o critério 3
m5=AHPWR::matrix_ahp(x,y)
m5
```

```
    A1   A2        A3 A4
A1 1.0 0.25 0.3333333  2
A2 4.0 1.00 2.0000000  5
A3 3.0 0.50 1.0000000  4
A4 0.5 0.20 0.2500000  1
```

```
AHPWR::CR(m5)
```

```
[1] 0.01791141
```

Considerando o **critério 5**, resistência ao tráfego, o julgador deve atribuir os pesos para cada alternativa. Supondo que foi atribuído $w_1 = 3, w_2 = 3.5, w_3 = 6$ *e* $w_4 = 5.5$, esses julgamentos possibilitam a criação da matriz m_6:

```
x=paste0("A",1:4) #nomes das alternativas A1, A2, A3,
  A4
y=c(3, 3.5, 6, 5.5) #julgamento holístico das
  alternativas para o critério 3
m6=AHPWR::matrix_ahp(x,y)
m6
```

```
    A1        A2        A3        A4
A1 1.0 0.6666667 0.2500000 0.2857143
```

```
A2 1.5 1.0000000 0.2857143 0.3333333
A3 4.0 3.5000000 1.0000000 1.5000000
A4 3.5 3.0000000 0.6666667 1.0000000
```

```
AHPWR::CR(m6)
```

```
[1] 0.006356667
```

Considerando o **critério 6**, a vida útil, o julgador deve atribuir os pesos para cada alternativa. Supondo que foi atribuído $w_1 = 3, w_2 = 4, w_3 = 6$ *e* $w_4 = 8$, esses julgamentos possibilitam a criação da matriz m_7:

```
x=paste0("A",1:4) #nomes das alternativas A1, A2, A3,
  A4
y=c(3, 4, 6, 8) #julgamento holístico das alternativas
  para o critério 3
m7=AHPWR::matrix_ahp(x,y)
m7
AHPWR::CR(m7)
```

```
   A1  A2        A3        A4
A1  1 0.5 0.2500000 0.1666667
A2  2 1.0 0.3333333 0.2000000
A3  4 3.0 1.0000000 0.3333333
A4  6 5.0 3.0000000 1.0000000
```

```
AHPWR::CR(m7)
```

```
[1] 0.02917912
```

Após o estabelecimento dos julgamentos, devemos organizar todas as matrizes numa base de dados.

O pacote AHPWR oferece duas opções, podemos organizar todas as matrizes em uma lista ou podemos organizar todas as matrizes em um arquivo de planilhas de tal forma que cada planilha corresponda a cada matriz.

```
#Organizando em lista
base = list(m1, m2, m3, m4, m5, m6, m7)
base
```

```
[[1]]
          C1        C2        C3        C4        C5  C6
C1 1.0000000 0.3333333 5.5000000 0.2500000 0.2857143 6.0
C2 3.0000000 1.0000000 7.5000000 0.5000000 0.6666667 8.0
C3 0.1818182 0.1333333 1.0000000 0.1176471 0.1250000 1.5
C4 4.0000000 2.0000000 8.5000000 1.0000000 1.5000000 9.0
C5 3.5000000 1.5000000 8.0000000 0.6666667 1.0000000 8.5
C6 0.1666667 0.1250000 0.6666667 0.1111111 0.1176471 1.0

[[2]]
    A1        A2        A3        A4
A1 1.0 0.5000000 0.6666667 0.5263158
A2 2.0 1.0000000 1.5000000 1.1000000
A3 1.5 0.6666667 1.0000000 0.7142857
A4 1.9 0.9090909 1.4000000 1.0000000

[[3]]
   A1        A2  A3        A4
A1  1 0.3333333 0.5 0.2500000
A2  3 1.0000000 2.0 0.5000000
```

```
A3  2 0.5000000 1.0 0.3333333
A4  4 2.0000000 3.0 1.0000000

[[4]]
    A1        A2        A3        A4
A1 1.0 0.9090909 0.7692308 0.7142857
A2 1.1 1.0000000 0.8333333 0.7692308
A3 1.3 1.2000000 1.0000000 0.9090909
A4 1.4 1.3000000 1.1000000 1.0000000

[[5]]
    A1   A2        A3 A4
A1 1.0 0.25 0.3333333  2
A2 4.0 1.00 2.0000000  5
A3 3.0 0.50 1.0000000  4
A4 0.5 0.20 0.2500000  1

[[6]]
    A1        A2        A3        A4
A1 1.0 0.6666667 0.2500000 0.2857143
A2 1.5 1.0000000 0.2857143 0.3333333
A3 4.0 3.5000000 1.0000000 1.5000000
A4 3.5 3.0000000 0.6666667 1.0000000

[[7]]
   A1  A2        A3        A4
A1  1 0.5 0.2500000 0.1666667
A2  2 1.0 0.3333333 0.2000000
A3  4 3.0 1.0000000 0.3333333
```

```
A4  6 5.0 3.0000000 1.0000000
```

```
#Organizando em arquivo denominado Exemplo_2.xlsx
file1 = AHPWR::xlsx_ahp(m1, file = "Exemplo_2.xlsx",
  sheet = "M1", append = FALSE)
file2 = AHPWR::xlsx_ahp(m2, file = "Exemplo_2.xlsx",
  sheet = "M2", append = TRUE)
file3 = AHPWR::xlsx_ahp(m3, file = "Exemplo_2.xlsx",
  sheet = "M3", append = TRUE)
file4 = AHPWR::xlsx_ahp(m4, file = "Exemplo_2.xlsx",
  sheet = "M4", append = TRUE)
file5 = AHPWR::xlsx_ahp(m5, file = "Exemplo_2.xlsx",
  sheet = "M5", append = TRUE)
file6 = AHPWR::xlsx_ahp(m6, file = "Exemplo_2.xlsx",
  sheet = "M6", append = TRUE)
file7 = AHPWR::xlsx_ahp(m7, file = "Exemplo_2.xlsx",
  sheet = "M7", append = TRUE)
```

Ao optar por armazenar as matrizes no arquivo *Exemplo_ 2.xlsx*, o mesmo deve ter sido criado na pasta de trabalho corrente, após rodar os comandos acima.

Para obter o resultado final do método, isto é, os pesos globais das alternativas, utilizamos a função **ahp_geral()**:

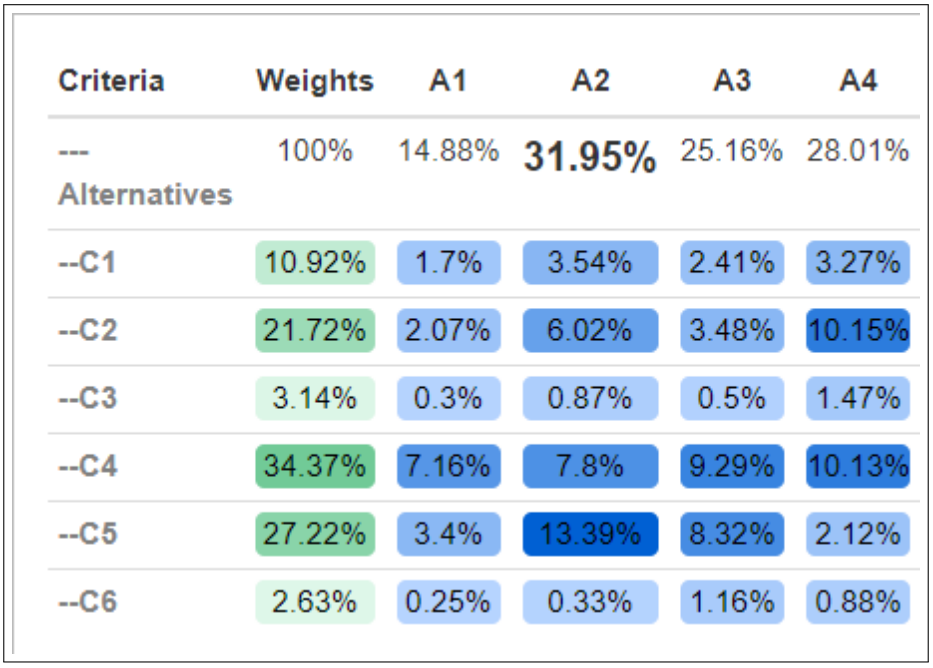
```
#Nomes das alternativas no vetor x
x
#aplicando o método na base de dados
AHPWR::ahp_geral(base, nomes_alternativas = x)
```

```
[1] "A1" "A2" "A3" "A4"

# A tibble: 7 x 7
```

	Criteria	Weights	A1	A2	A3	A4	CR
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	---Alternatives	1	0.115	0.309	0.299	0.276	0.0302
2	--C1	0.109	0.0170	0.0354	0.0241	0.0327	0.000614
3	--C2	0.217	0.0207	0.0602	0.0348	0.102	0.0115
4	--C3	0.0314	0.00653	0.00712	0.00848	0.00924	0.0000316
5	--C4	0.344	0.0429	0.169	0.105	0.0267	0.0179
6	--C5	0.272	0.0262	0.0345	0.120	0.0913	0.00636
7	--C6	0.0263	0.00182	0.00288	0.00678	0.0148	0.0292

A tabela formatada pode ser vista na Figura 6.10

Figura 6.10: Tabela Final com os pesos globais.



Criteria	Weights	A1	A2	A3	A4
--- Alternatives	100%	14.88%	**31.95%**	25.16%	28.01%
--C1	10.92%	1.7%	3.54%	2.41%	3.27%
--C2	21.72%	2.07%	6.02%	3.48%	10.15%
--C3	3.14%	0.3%	0.87%	0.5%	1.47%
--C4	34.37%	7.16%	7.8%	9.29%	10.13%
--C5	27.22%	3.4%	13.39%	8.32%	2.12%
--C6	2.63%	0.25%	0.33%	1.16%	0.88%

Fonte: A autora.

6.4 RESULTADOS E DISCUSSÃO

Os exemplos 1 e 2 descrevem o procedimento para aplicar o método AHP, desde sua estrutura hierárquica conforme Figura 8.1 e Figura 6.9, todo o procedimento de elaboração das matrizes de julgamento com base nos julgamentos de

especialista, até a obtenção do resultado final que é a tabela com os pesos globais de cada alternativa, conforme Figura 6.6 e Figura 6.10.

As tabelas com o peso global facilitam a escolha do tomador de decisão, pois mostram a ordem de preferência das alternativas em relação ao objetivo geral. No exemplo 1, a alternativa A2-Engenharia Civil foi a mais prioritária, com um peso de 40.29%. O critério C1-Afinidade foi o mais relevante para essa decisão, com um peso de 75.82%. As razões de consistência dos julgamentos (CR) foram todas inferiores a 10%, indicando uma boa coerência nas comparações.

No exemplo 2, a alternativa A2-Ponte Estaiada foi a mais prioritária, com um peso de 31.95%. O critério C4-Resistência ao vento foi o mais relevante para essa decisão, com um peso de 34.37%. As razões de consistência dos julgamentos (CR) foram todas inferiores a 10%, indicando uma boa coerência nas comparações.

Note que a soma dos pesos tanto dos critérios como das alternativas resultam em 100%. Os valores intermediários dos critérios versus alternativas mostram a contribuição de cada item na composição final do peso global.

Este método auxilia o tomador de decisão a fazer uma escolha com base no peso global das alternativas ou mesmo obter um ranking de importância dessas alternativas.

6.5 CONSIDERAÇÕES FINAIS

Neste capítulo, apresentamos o método AHP como uma técnica de apoio à decisão que permite comparar e priorizar alternativas com base em critérios múltiplos e subjetivos. O método consiste em construir uma hierarquia de objetivos, critérios e alternativas, e atribuir pesos (ou importâncias) a cada elemento da hierarquia por meio de comparações pareadas. As tabelas com os pesos globais do método AHP são o resultado da agregação dos pesos locais de cada nível da hierarquia, ponderados pelos pesos dos níveis superiores. Essas tabelas permitem visualizar a ordem de preferência das alternativas em relação ao objetivo geral, bem como a consistência das comparações realizadas. Quanto maior o peso global

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

de uma alternativa, maior é a sua prioridade na decisão. Quanto menor o índice de consistência, maior é a coerência das comparações.

O método AHP tem como vantagens a sua simplicidade, flexibilidade e consistência, mas também possui algumas limitações, como a necessidade de um grande número de comparações, a sensibilidade aos pesos dos critérios e a dificuldade de lidar com incertezas. O método AHP pode ser aplicado em diversas áreas do conhecimento e em diferentes tipos de problemas de decisão, como por exemplo: escolha da melhor localização de uma usina hidrelétrica, seleção do melhor fornecedor de um produto ou serviço, avaliação do desempenho de funcionários ou organizações, definição da melhor estratégia de marketing, alocação de recursos militares em um cenário de conflito, intervenção em uma situação de crise, vulnerabilidade de um sistema de defesa, entre outros. O método AHP pode ajudar os líderes a tomar decisões racionais e consistentes em situações complexas e incertas.

Foram explicados os passos para a aplicação do método AHP, desde a definição do problema até a obtenção dos resultados, destacando a importância de uma boa estruturação da hierarquia, da escolha de uma escala de comparação adequada e da verificação da consistência das comparações. Foi demonstrado o uso do pacote AHPWR para implementar o método AHP na linguagem R, mostrando as funções disponíveis e os exemplos de código. O pacote AHPWR facilita a realização das comparações pareadas, o cálculo dos pesos locais e globais, a verificação da consistência e a visualização dos resultados. Foi ilustrada a aplicação do método AHP e do pacote AHPWR em dois casos práticos de seleção de alternativas com base em critérios múltiplos, um envolvendo a escolha da profissão e outro a escolha do melhor método de construção de uma ponte.

Espera-se que este capítulo possa contribuir para a divulgação e o uso do método AHP e do pacote AHPWR na comunidade acadêmica e profissional, bem como estimular novas pesquisas e aplicações na área de apoio à decisão.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

6.6 REFERÊNCIAS

ALCOFORADO, Luciane Ferreira; LONGO, Orlando Celso. **Introduction to AHPWR package**. [S.l.], 2022. R package version 0.1.0. Disponível em: <https://cran.r-project.org/web/packages/AHPWR/vignettes/Intro_to_AHP.html>.

ALCOFORADO, Luciane Ferreira; SOUSA, Lyncoln; LONGO, Orlando Celso Longo. **AHPWR: Compute Analytic Hierarchy Process**. [S.l.], 2022. R package version 0.1.0.

GODOI, Wagner da Costa. Método de construção das matrizes de julgamento paritário no AHP – método de julgamento holístico. **Revista Gestão Industrial**, v. 10, n. 3, p. 474–493, 2014. ISSN 1808-0448. DOI: 10.3895/gi.v10i3.1970.

SAATY, T.L.; VARGAS, L.G. **Models, Methods, Concepts and Applications of the Analytic Hierarchy Process**. New York: Springer, 2012.

SAATY, Thomas L. **The Analytic Hierarchy Process**. New York: McGraw-Hill, 1980.

Capítulo 7

INTELIGÊNCIA ARTIFICIAL APLICADA À ENGENHARIA

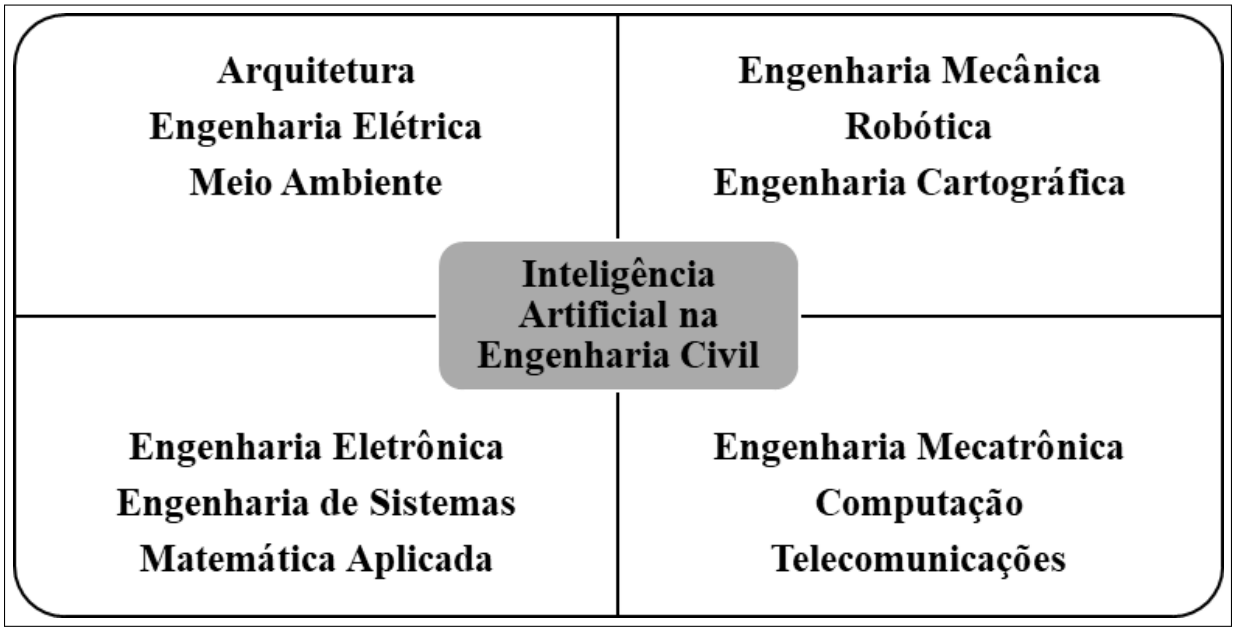
Autor: Marco Aurélio Chaves Ferro
Programa de Pós-Graduação em Engenharia Civil - UFF
e-mail: marcoferro@id.uff.br

O presente trabalho apresenta as principais características e metodologias de Inteligência Artificial adotadas em Engenharia, com destaque para a Engenharia Civil. São descritos os aspectos fundamentais de Aprendizado de Máquinas (*Machine Learning*) e Aprendizado Profundo (*Deep Leaerning*) e suas principais técnicas, com enfoque especial em Redes Neurais Artificiais (*Artificial Neural Network*). Alguns exemplos de utilização da Inteligência Artificial são mostrados, como Gêmeos Digitais, Computação Quântica, Metaverso e ChatGPT. O exemplo de aplicação mostra a predição da resistência do concreto à compressão, usando-se uma planilha disponível na internet (*dataset*), com 1032 experimentos realizados em laboratório. O código R do programa usou a técnica de Redes Neurais Artificiais para obter os resultados que foram plenamente satisfatórios.
Palavras-Chave: Inteligência Artificial; Redes Neurais Artificiais; Linguagem R.

7.1 INTRODUÇÃO

Este trabalho apresenta os principais aspectos da Inteligência Artificial, com suas aplicações à Engenharia de maneira geral, no entanto focado na Engenharia Civil, considerando a sua multidisciplinaridade, quais sejam as disciplinas Arquitetura, Engenharia Mecânica, Engenharia Mecatrônica, Robótica, Computação, Engenharia de Sistemas, Meio Ambiente, dentre outras, conforme ilustrado na Figura 8.1, a seguir.

Figura 7.1: Multidisciplinaridade da IA na Engenharia Civil.



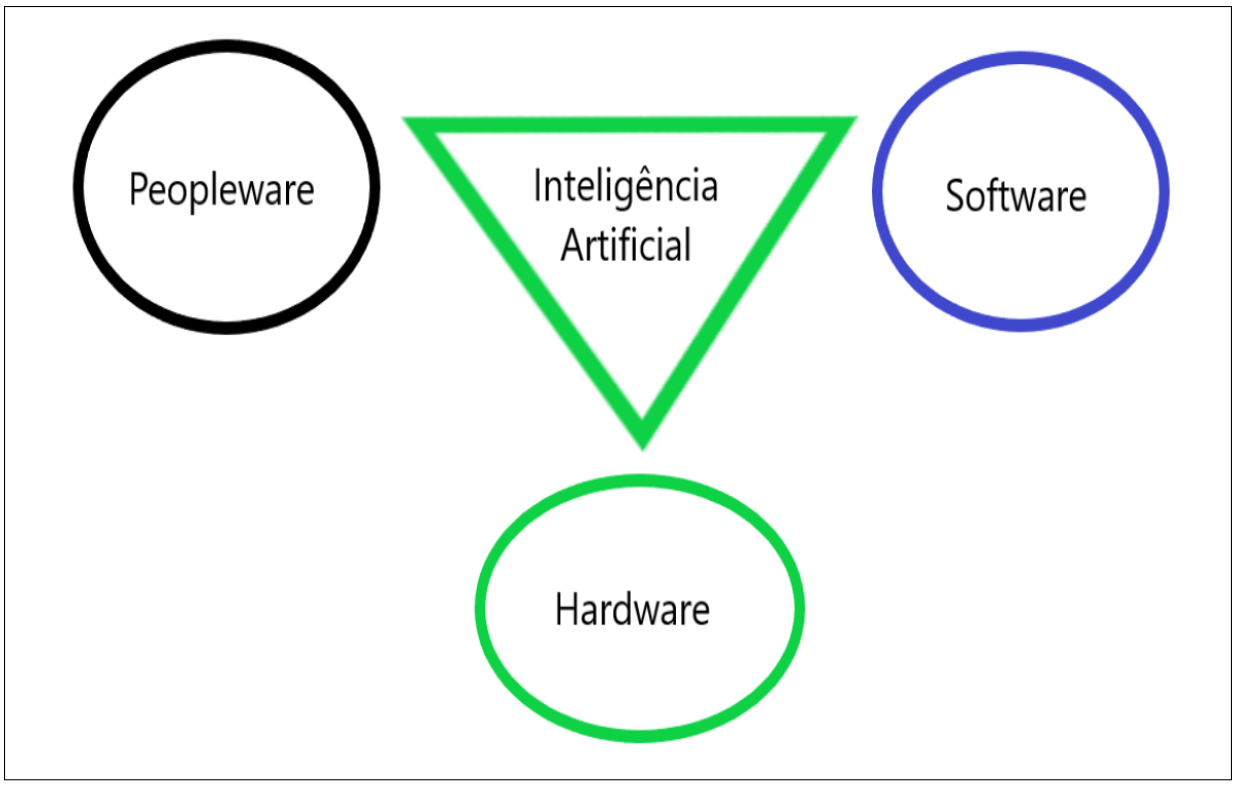
Fonte: O autor.

A simulação numérica é abordada em Engenharia desde o advento do computador e diversos métodos numéricos foram sendo criados por pesquisadores e cientistas ao longo do tempo, destacando-se, dentre outros, Diferenças Finitas, Elementos Finitos, Elementos de Contorno, Volumes Finitos, de acordo com (FERRO, 2020). Há várias definições sobre o significado de Inteligência Artificial na literatura que trata do assunto, destacando-se o descrito no livro de (RUSSELL; NORVIG, 2003).Ela pode ser entendida como a ciência que procura simular o comportamento humano (*peopleware*) com o uso de programas de com-

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

putadores (*softwares*) e máquinas (*hardwares*), conforme ilustrado na Figura 8.2.

Figura 7.2: Tripé da Inteligência Artificial.



Fonte: O autor.

7.2 OBJETIVO

A simulação em Inteligência Artificial é realizada por diversas técnicas, cada uma utilizada sozinha ou em combinação com outras, de acordo com o seu objetivo, destacando-se Redes Neurais Artificiais (RNA)/*Artificial Neural Network* (ANN), Aprendizado de Máquinas (AM)/*Machine Learning* (ML), Aprendizado Profundo (AP) /*Deep Learning* (DL), Algoritmos Genéticos (AG)/*Genetic Algorythms*, dentre outros. Após uma explicação sucinta de algumas aplicações da Inteligência Artificial, destacando-se gêmeos digitais, metaverso, computação quântica e o ChatGPT, será apresentada a técnica de Redes Neurais Artificiais que foi utilizada em um exemplo de aplicação na Engenharia Civil, que é a predição da resistência do concreto, quando são mostradas as considerações finais.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

O Aprendizado de Máquina (*Machine Learning*) é um método de análise de dados que busca a automação, o desenvolvimento e a criação de modelos analíticos. É baseado na hipótese de que sistemas computacionais (*software* e *hardware*) podem aprender com dados genéricos, estruturados ou não, identificar padrões e tomar decisões com o mínimo de intervenção humana (*peopleware*).

Aprendizagem Profunda (*Deep Learning*) é um termo adotado dentro do campo do Aprendizado de Máquinas sendo formado por programas, computadores, sensores e dispositivos inteligentes, conectados entre si e que desempenham suas tarefas sem ou com muito pouca necessidade de ações humanas. É um tipo especializado do Aprendizado de Máquinas, com um linguajar matemático mais especializado. É muito utilizado no reconhecimento de face, de voz e nos processamentos de imagem e de linguagem. Redes Neurais Artificiais (*Artificial Neural Network*) são sistemas de computação com nós interconectados que funcionam como o sistema nervoso do ser humano.

Gêmeos digitais, conforme ilustrado na Figura 8.3, são cópias digitais das estruturas reais. Possuem sensores que permitem a aquisição de grandezas importantes como deslocamentos e temperatura, por exemplo. São capazes de identificar falhas e tomar decisões sobre manutenção preventiva e corretiva.

Os computadores quânticos (ver Figura 7.4) aproveitam o comportamento exclusivo da física quântica, como sobreposição, emaranhamento e interferência quântica, e o aplicam à computação. Isso apresenta novos conceitos para os métodos de programação tradicionais. No entanto, por ser uma tecnologia pouco acessível ao público, o tema ainda causa dúvidas e controvérsias.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 7.3: Gêmeos Digitais: saiba o que é.

Fonte: (PEDERNEIRAS, 2023).

Figura 7.4: Computador Quântico.

Fonte: (MELKO R.; GOLDFARB; BOVA, 2023).

O metaverso é um mundo 3D virtual compartilhado, ou mundos, que são interativos, imersivos e colaborativos (ver Figura 7.5) . Assim como o universo físico é uma coleção de mundos conectados no espaço, o metaverso também pode

ser considerado um conjunto de mundos conectados digitalmente. O metaverso se tornará uma plataforma que não estará vinculada a nenhuma aplicação ou ambiente único, digital ou real.

Figura 7.5: Metaverso.

Fonte: (STOCCO, 2023).

O Chat GPT é um algoritmo baseado em Inteligência Artificial (*machine learning* e redes neurais). Ele foi criado por um laboratório de pesquisas em Inteligência Artificial chamado OpenAI, com sede em San Francisco, EUA. O nome Chat GPT é uma sigla para "*Generative Pre-Trained Transformer*". O sucesso da ferramenta está em oferecer ao usuário uma forma simples de conversar e obter respostas. A gama de assuntos é bem vasta e o Chat GPT é capaz de elaborar textos sobre assuntos variados, dentre inúmeras outras tarefas.

7.3 APLICAÇÕES

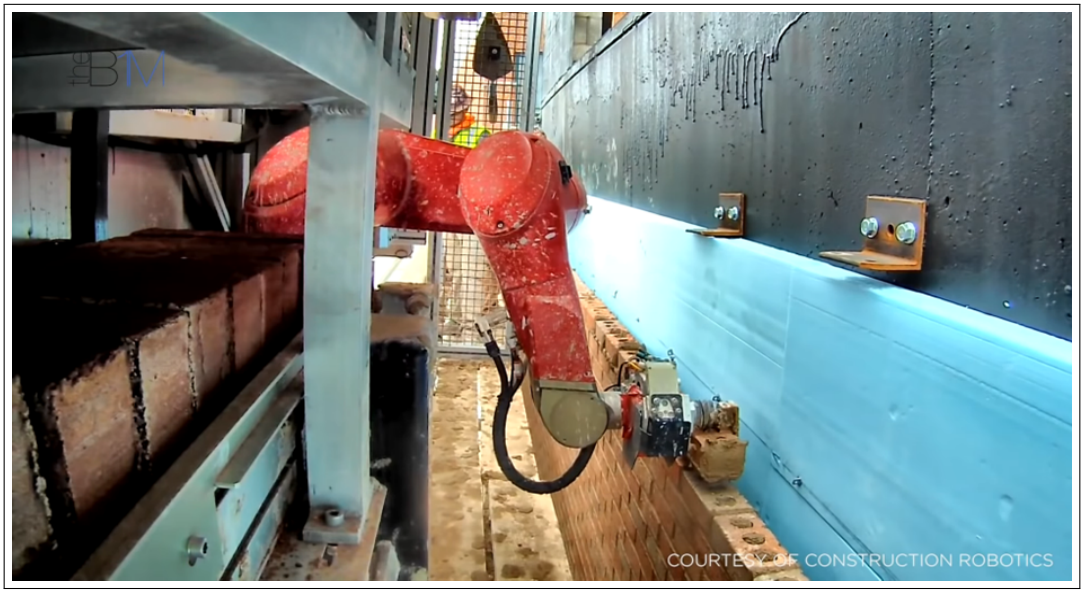
Algumas aplicações da Inteligência Artificial na Engenharia Civil são mostradas nas figuras a seguir. Todas elas demonstram que a Inteligência Artificial traz como resultados extrema precisão, rapidez, segurança, economia de material e de mão de obra, dentre muitas outras vantagens.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

O primeiro exemplo é o uso de robôs para assentamento de alvenaria de tijolos, conforme visto na Figura 7.6.

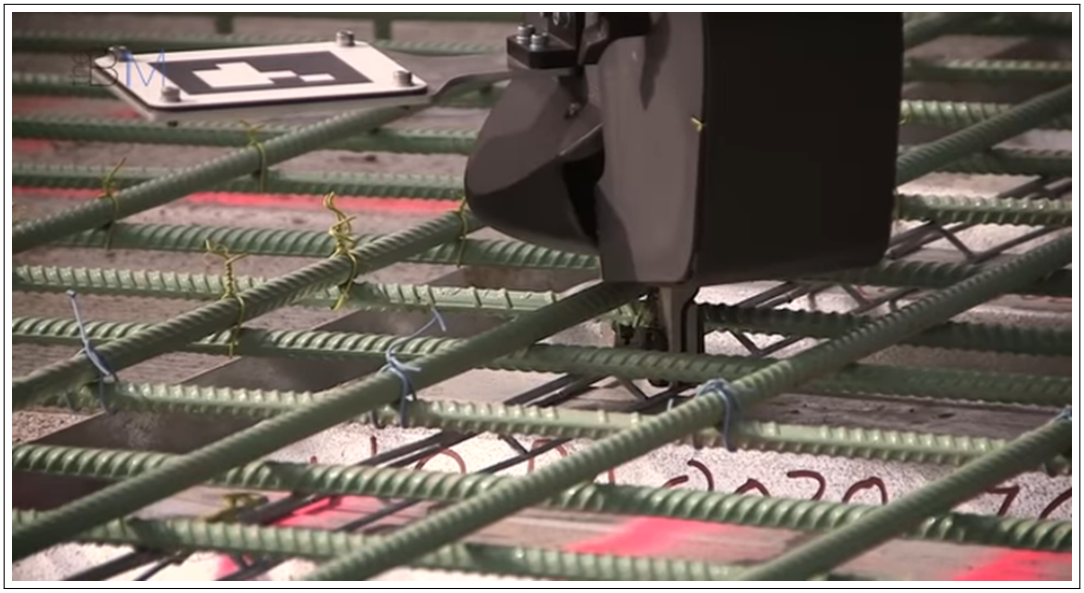
O segundo exemplo é a colocação de armadura de aço e sua amarração com arames, com o uso de robôs, ver Figura 7.7.

Figura 7.6: Assentamento de alvenaria com robô.



Fonte: (THEB1M, 2023).

Figura 7.7: Execução de amarração de armadurta de laje com robô.



Fonte: (THEB1M, 2023).

O terceiro exemplo consiste no transporte de material pesado por um robô, ao invés do uso de carrinho de mão por operários, conforme mostrado na Figura 7.8.

Figura 7.8: Transporte de material pesado em obra por robô.

Fonte: (THEB1M, 2023).

O quarto exemplo mostra o serviço de terraplenagem realizado por máquinas automatizadas, conforme visto na Figura 7.9.

Figura 7.9: Serviço automatizado de terraplenagem.

Fonte: (THEB1M, 2023).

A evolução de fiscalização de obras pode ser evidenciada no quinto exemplo. A grande maioria das medições e verificações realizadas na fiscalização de obras é feita de forma tradicional, utilizando materiais como trenas, esquadros, níveis, calculadoras, pranchetas, dentre outros, como visto na Figura 7.10

Figura 7.10: Fiscalização de obra tradicional.

Fonte: (PORTILHO, 2023).

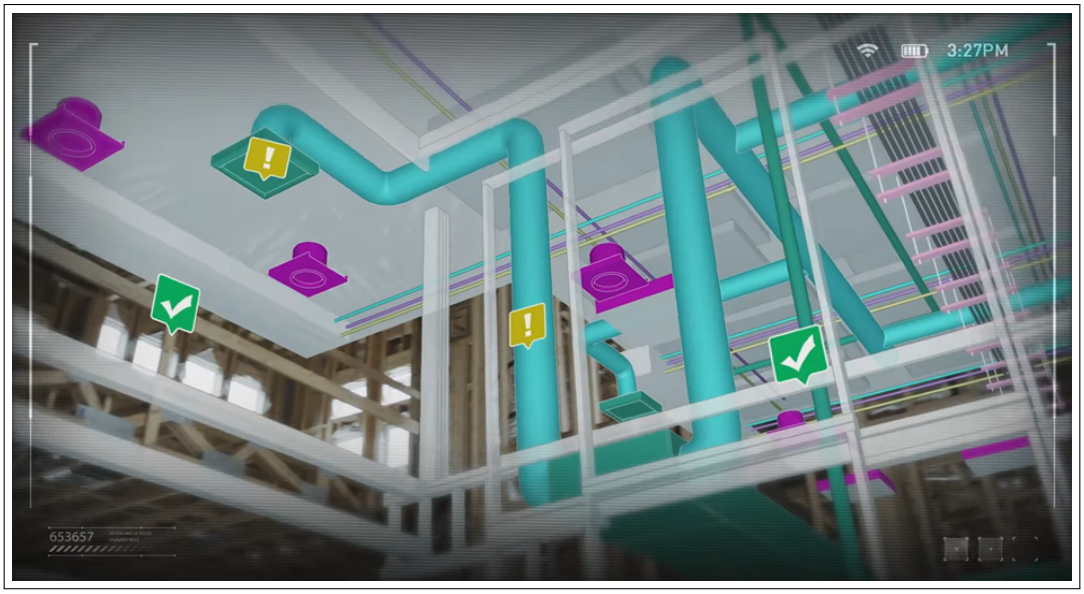
Uma evolução tecnológica da fiscalização de obras é o uso de óculos especiais (ver Figura 7.11) que verificam instalações (ver Figura 7.12), fazem medições, enviam relatórios para o escritório indicando não conformidades, dentre outros.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 7.11: Óculos especial para medição de obras.

Fonte: (SRI, 2023).

Figura 7.12: Verificação de instalações.

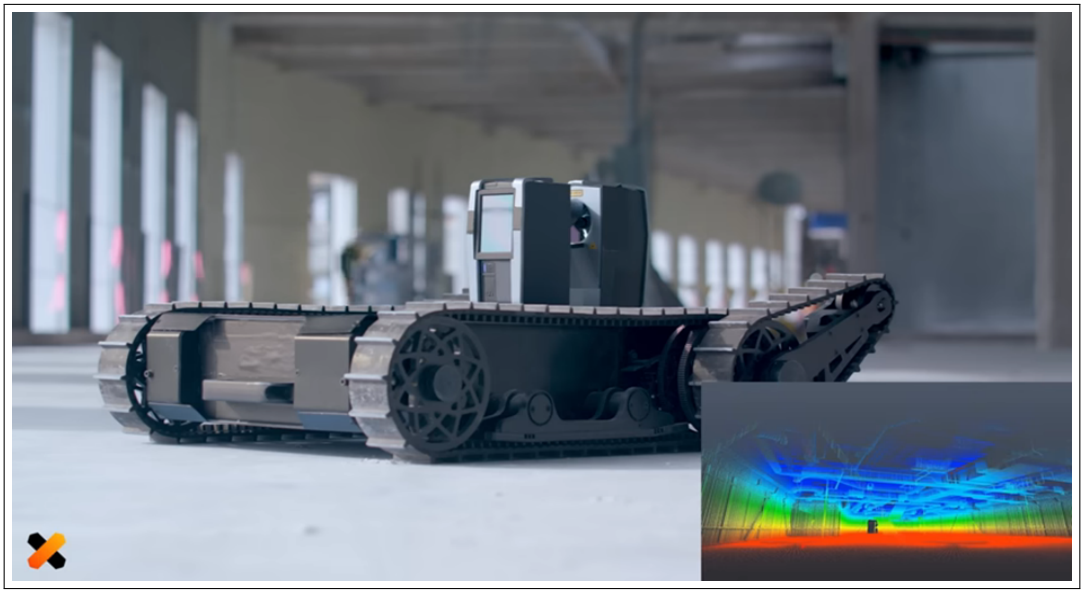


Fonte: (SRI, 2023).

Uma evolução ainda mais avançada é o uso de um robô (ver Figura 7.13) fara realizar os serviços de fiscalização de obras, e na Figura 7.14 podem ser verificadas as medições atualizadas, ou seja, 78,28%, 42,61% e 83,26% do total, executados

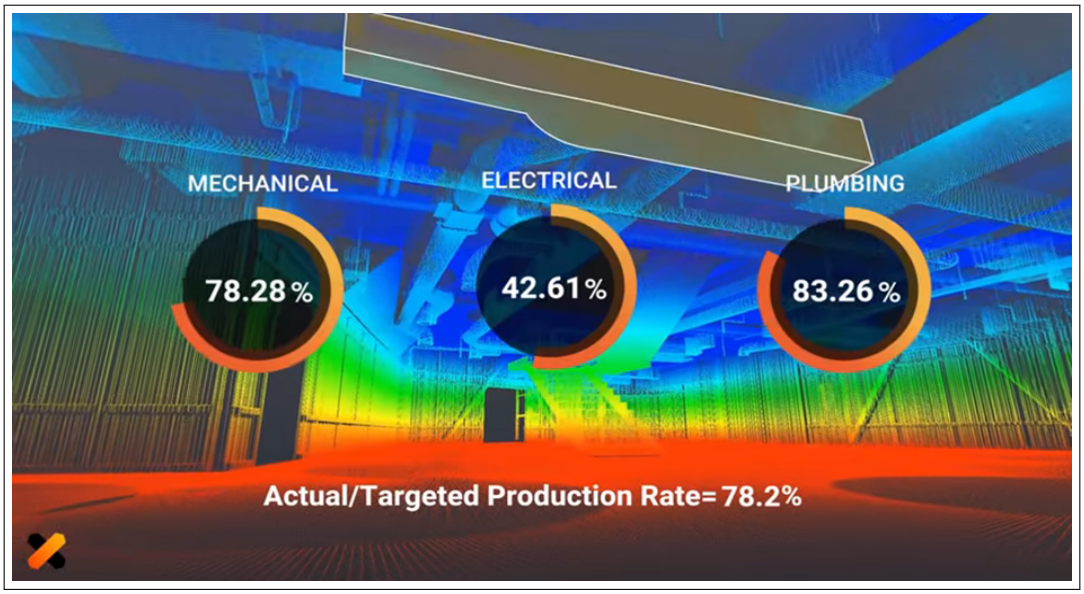
para os serviços mecânicos, elétricos e hidrossanitários, respectivamente. O cronograma está com 78,2% executado, em relação ao previsto, indicando atraso de 21,8%.

Figura 7.13: Robô para fiscalizar obras.



Fonte: (DOXEL, 2023).

Figura 7.14: Fiscalização de obras com robô.



Fonte: (DOXEL, 2023).

O conhecimento da Resistência do Concreto à Compressão é fundamental no Cálculo de Estruturas de Concreto. Essa característica depende de vários fatores, destacando-se a taxa de cimento (kg/m3), taxa de escória (kg/m3), cinzas (kg/m3), água (kg/m3), superplastificante (kg/m3), agregado graúdo ou brita (kg/m3), agregado miúdo ou areia (kg/m3), idade (dias), resistência do concreto (MPa). Um arquivo de dados com 1030 linhas na forma de planilha eletrônica está disponibilizado em (YEH, 2023) e foi criado por (YEH, 1998). O exemplo consta do livro de (LANTZ, 2019), em seu capítulo 7, na página 229. As quatro primeiras linhas do arquivo de dados estão mostradas na Figura 7.15.

Figura 7.15: 4 primeiras linhas do arquivo de dados.

Redes Neurais com R – Predição da Resistência do Concreto – 4 primeiras linhas do arquivo de entrada

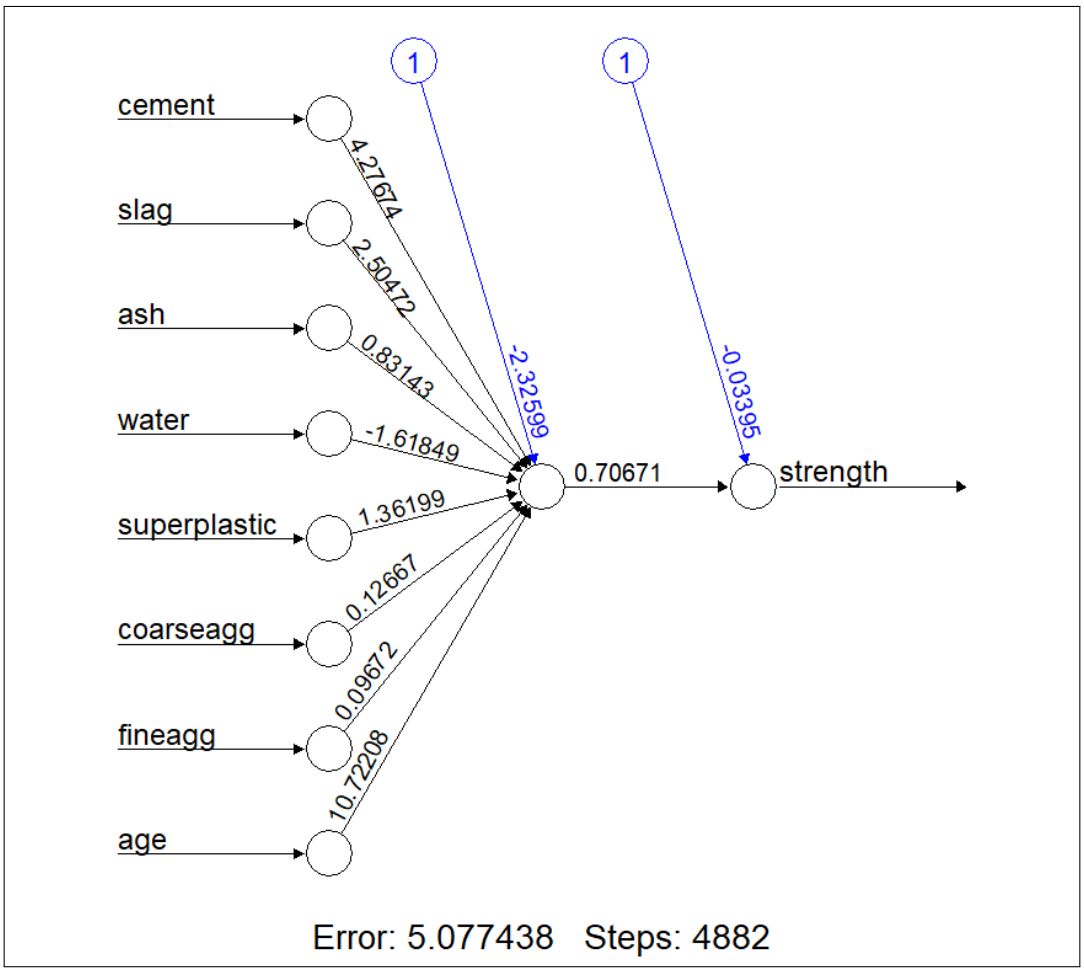
141.3,212,0,203.5,0,971.8,748.5,28,29.89
168.9,42.2,124.3,158.3,10.8,1080.8,796.2,14,23.51
250,0,95.7,187.4,5.5,956.9,861.2,28,29.22
266,114,0,228,0,932,670,28,45.85

Fonte: (LANTZ, 2019).

Uma linha aleatória do conjunto de dados foi separada para verificação da Rede Neural Artificial, cujo resultado final é uma Resistência do Concreto de 30,14 MPa, o que é denominado alvo.

A primeira rede a ser rodada consta de apenas um neurônio (perceptron) e está mostrada na Figura 7.16.

Figura 7.16: Rede Neural com um neurônio.



Fonte: (LANTZ, 2019).

7.4 RESULTADOS E DISCUSSÕES

O resultado indicou um valor predito de 32,59 MPa e coeficiente de correlação (r) igual a 0,806. Em seguida foi rodada uma Rede Neural Artificial com uma camada oculta e cinco neurônios, sendo o valor predito igual a 29,29 MPa e o coeficiente de correlação (r) igual a 0,924. A Rede Neural Artificial pode ser aperfeiçoada ao aumentar-se o número de camadas ocultas e de neurônios por camada, assim como usar outras funções de ativação, de erro e demais hiperparâmetros. Comparado com o alvo, ou seja 30,14 MPa, o resultado de 29,29 MPa dá um erro relativo de -2,82%, o que já pode ser considerado excelente.

7.5 CONCLUSÕES

A Inteligência Artificial está no dia a dia das pessoas. Ela pode ajudá-las a eliminar ou diminuir tarefas repetitivas e mecânicas e tornar os trabalhos mais eficazes, eficientes e efetivos. Várias aplicações da Inteligência Artificial foram apresentadas neste capítulo e um exemplo prático da predição da Resistência do Concreto com Redes Neurais Artificiais foi desenvolvido. Os resultados mostraram-se excelentes e muitas outras aplicações de Inteligência Artificial podem ser elaboradas nas diversas áreas da Engenharia.

7.6 REFERÊNCIAS

DOXEL. **Doxel Uses AI and Robots to Track Construction Projects**. [S.l.: s.n.], set 2023. Disponível em: https://www.youtube.com/watch?v=0369vlp_fjg. Acessado em: 07 set. 2023.

FERRO, Marco Aurélio Chaves. Inteligência Artificial Aplicada à Engenharia Civil: Um Enfoque Multidisciplinar. In: Accessed on 2023-10-20.

LANTZ, BRETT. **Machine Learning with R**. [S.l.]: PACKT Publishing, 2019. ISBN 978-1-78829-586-4.

MELKO R.; GOLDFARB, A.; BOVA, F. **Como a computação quântica desafiará a computação clássica nos negócios**. [S.l.: s.n.], set 2023. Disponível em: https://www.mitsloanreview.com.br/post/como-a-computacao-quantica-desafiara-a-computacao-classica-nos-negocios. Acessado em: 07 set. 2023.

PEDERNEIRAS, GABRIELA. **Gêmeos Digitais**. [S.l.: s.n.], set 2023. Disponível em: https://www.industria40.ind.br/artigo/19798-gemeos-digitais-saiba-o-que-e. Acessado em: 07 set. 2023.

PORTILHO, GABRIELA. **8 coisas que você precisa saber se quer fazer engenharia**. [S.l.: s.n.], set 2023. Disponível em: https://guiadoestudante.abril.com.br/orientacao-profissional/8-coisas-que-voce-precisa-saber-se-quer-fazer-engenharia/. Acessado em: 07 set. 2023.

RUSSELL, Stuart; NORVIG, Peter. **Artificial Intelligence: A Modern Approach**. [S.l.]: Pearson, 2003.

SRI. **Augmented Reality Solutions for Construction Inspection**. [S.l.: s.n.], set 2023. Disponível em: https://www.youtube.com/watch?v=8lY4qaVvR8c. Acessado em: 07 set. 2023.

STOCCO, GUGA. **O que tem de real dentro do buzz do metaverso?** [S.l.: s.n.], set 2023. Disponível em: https://mittechreview.com.br/o-que-tem-de-real-dentro-do-buzz-do-metaverso/. Acessado em: 07 set. 2023.

THEB1M. **The Construction Robots are Coming**. [S.l.: s.n.], set 2023. Disponível em: https://www.youtube.com/watch?v=nKGGHdl3NyQ. Acessado em: 07 set. 2023.

YEH, I. Modeling of Strenght of Hight-Performence Concrete Using Artificial Neural Network. **Cement and Concrete Research**, v. 28, n. 2, p. 1797–1808, 1998.

_____. **UCI Machine Learning Repository**. [S.l.: s.n.], set 2023. Disponível em: http://archive.ics.uci.edu/. Acessado em: 07 set. 2023.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Capítulo 8

O PROBLEMA DO ANIVERSÁRIO, O PACOTE IPSUR E SEU PLUGIN PARA O R COMMANDER: Uma possibilidade para sala de aula

Autor: Felipe Rafael Ribeiro Melo
Universidade Federal do Estado do Rio de Janeiro, UNIRIO-RJ
e-mail: felipe.ribeiro@uniriotec.br

Em uma sala com n pessoas, qual a probabilidade de pelo menos um par destas pessoas fazer aniversário no mesmo dia? Esse é o conhecido Problema do Aniversário. Sua solução engloba conceitos da Probabilidade, como o conceito clássico da Probabilidade e a relação entre as probabilidades de um evento e do seu evento complementar, além de técnicas básicas em Análise Combinatória, como arranjos e combinações. Há pacotes no *software* R com funções que calculam probabilidades associadas ao Problema do Aniversário. Este capítulo aborda inicialmente a solução do Problema do Aniversário, explicitando seus pressupostos, e apresenta os pacotes `IPSUR` e `RcmdrPlugin.IPSUR`, que trazem consigo funções associadas ao Problema do Aniversário. Este último pacote, em particular, é um *plugin* a ser utilizado na interface R Commander, que permite o uso de várias funcionalidades

do R por meio de menus, sem a necessidade de digitação e/ou execução de linhas de comando. Isto pode motivar docentes de Matemática do Ensino Médio e docentes do Ensino Superior em disciplinas de Probabilidade a realizar atividades em sala de aula que abordem o Problema do Aniversário, inclusive associando o valor n ao número de alunos em sala de aula e realizando dinâmicas em sala de aula.
Palavras Chave: Problema do Aniversário; Linguagem R; IPSUR; R Commander; Sala de Aula.

8.1 INTRODUÇÃO

O chamado Problema do Aniversário (*Birthday Problem*) versa sobre a probabilidade de pelo menos um par de pessoas aniversariar em um mesmo dia em uma sala com n pessoas. Em pesquisas sobre este assunto, também é comum se deparar com o termo Paradoxo do Aniversário (*Birthday Paradox*). Esta expressão deriva do fato contraintuitivo de que bastam 23 pessoas para que a probabilidade de ao menos um par de pessoas fazer aniversário no mesmo dia exceda 50%, ou seja, a ocorrência deste evento é mais provável que a sua não ocorrência em um grupo de apenas 23 pessoas - quantitativo pequeno face aos 365 dias possíveis que uma pessoa, aleatoriamente escolhida, possa aniversariar.

O Problema do Aniversário geralmente é atribuído ao matemático inglês Harold Davenport, por volta de 1927. Embora Davenport não o tenha publicado na época, ele não reivindicou ser seu descobridor porque "ele não podia acreditar que não havia sido estabelecido anteriormente" (SINGMASTER, 2004). A primeira publicação de uma versão do Problema do Aniversário se deve a **vonmises**, esperado de pares com o mesmo aniversário é de cerca de 1 quando o grupo tem 29 pessoas, porém não resolve o problema usual (SINGMASTER, 2004). Feller (1950) estabelece a probabilidade de todas as n pessoas aniversariarem em dias diferentes, obtém uma aproximação para esta probabilidade e descobre que, para $n = 23$, tal probabilidade é inferior a 50% e, consequentemente, a probabilidade

de ao menos um par de pessoas aniversariar no mesmo dia num grupo de 23 pessoas excede 50%.

O cálculo da probabilidade de interesse do Problema do Aniversário diretamente pelo evento "pelo menos um par de pessoas na sala aniversaria no mesmo dia" é trabalhoso no sentido de computar a quantidade de casos favoráveis a este evento, sobretudo em salas com muitas pessoas (ou seja, com n grande). Contudo, uma estratégia que facilita substancialmente este cálculo, tornado-o computacionalmente viável e muito mais rápido, é calcular a probabilidade do evento complementar de "pelo menos um par de pessoas na sala aniversaria no mesmo dia", ou seja, calcular a probabilidade do evento "todas as pessoas na sala aniversariam em dias diferentes". Subtraindo esta última probabilidade de 1, chega-se na probabilidade de interesse do Problema do Aniversário.

Uma vez desenvolvida a solução do Problema do Aniversário, o interesse volta-se ao uso de uma ou mais ferramentas computacionais para a realização dos cálculos de interesse. As ferramentas utilizadas neste capítulo são o *software*/linguagem R (TEAM, 2023) e os seus pacotes `IPSUR` (KERNS, 2023a) e `RcmdrPlugin.IPSUR` (KERNS, 2019). Tais pacotes possuem funções específicas voltadas ao Problema do Aniversário. Enquanto o primeiro deles é conveniente para quem tem um domínio mínimo da linguagem de programação R, o segundo é um *plugin* para a interface "point & click" R Commander, contornando a dificuldade de quem tem pouca ou nenhuma experiência com a linguagem R. Em particular, uma aplicação de atividade em sala de aula abordando o Problema do Aniversário com o uso da interface R Commander pode se tornar uma experiência mais agradável tanto para docentes como para discentes em relação a aplicação desta mesma atividade usando apenas linhas de comando.

8.2 OBJETIVO

O Objetivo principal deste capítulo é apresentar o Problema do Aniversário, incluindo sua solução sob certos pressupostos, e ferramentas do *software* R

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

voltadas a este problema. De forma mais específica, deseja-se:

- apresentar, de maneira clara e didática, a solução do Problema do Aniversário, de forma que estudantes de ensino médio e de ensino superior com conhecimentos basilares em Probabilidade e Análise Combinatória possam compreendê-la;
- mostrar como instalar e carregar o pacote `IPSUR` do *software* R, o qual carrega funções voltadas ao Problema do Aniversário.
- explorar as funções do pacote `IPSUR` voltadas ao Problema do Aniversário.
- por meio de representação gráfica, verificar como a probabilidade de interesse do Problema do Aniversário evolui conforme o número de pessoas na sala aumenta;
- apresentar a interface R Commander, proveniente do pacote Rcmdr, e o *plugin* do pacote `IPSUR` nesta interface, denominado `RcmdrPlugin.IPSUR`;
- sugerir dinâmicas em sala de aula envolvendo o Problema do Aniversário, com suporte da interface R Commander e do pacote `RmcdrPlugin.IPSUR` para a realização de cálculos de probabilidades relacionadas a este problema.

8.3 APLICAÇÃO

Esta seção aborda inicialmente a solução do Problema do Aniversário. A solução clássica considera um conjunto de três pressupostos, que seguem abaixo:

1. Desconsiderar 29 de fevereiro como uma data possível de aniversário.
2. Não ter irmãos gêmeos na sala.
3. Assumir a mesma probabilidade de aniversário para cada um dos 365 dias possíveis, ou seja: $P(\text{“01/jan”}) = P(\text{“02/jan”}) = \ \ldots \ = P(\text{“30/dez”}) = P(\text{“31/dez”}) = 1/365$.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Em particular, a última destas suposições permite o uso da interpretação clássica da Probabilidade para a obtenção da probabilidade de interesse. Considerando uma sala com n pessoas, seja o evento

$$B = \{\text{“pelo menos um par de pessoas aniversariar no mesmo dia”}\}.$$

A probabilidade de interesse é $P(B)$, onde B é um evento num espaço amostral finito uniforme Ω composto por 365^n elementos. Cada elemento de Ω é uma n-upla, onde a ordem é relevante. Por exemplo, para $n = 3$, $(18/mai, 16/out, 05/nov)$ e $(18/mai, 05/nov, 16/out)$ são dois elementos diferentes em Ω. O j-ésimo elemento de cada um destes ternos ordenados pode ser pensado como o aniversário da j-ésima pessoa arguida sobre sua data de aniversário.

O fato de Ω ser um espaço amostral finito uniforme faz com que a probabilidade do evento B possa ser calculada por meio da simples razão entre o número de elementos de B e o número de elementos de Ω, ou seja,

$$P(B) = \frac{\#B}{\#\Omega} = \frac{\#B}{365^n}.$$

Contudo, o numerador da fração acima é de difícil obtenção até mesmo para valores de n relativamente pequenos. Voltando ao caso de uma sala com apenas 3 pessoas (ou seja, $n = 3$), o evento B pode ser escrito como a união dos seguintes eventos mutuamente exclusivos:

- $E_1 =$ “aniversário da 1ª e da 2ª pessoa no mesmo dia e da 3ª pessoa em outro dia”,
- $E_2 =$ “aniversário da 1ª e da 3ª pessoa no mesmo dia e da 2ª pessoa em outro dia”,
- $E_3 =$ “aniversário da 2ª e da 3ª pessoa no mesmo dia e da 1ª pessoa em outro dia”,

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

- E_4 = "todas as 3 pessoas fazem aniversário no mesmo dia".

O número de elementos do evento E_1 é o número de arranjos simples de 365 elementos tomados 2 a 2: $A_{365,2} = 365 \times 364 = 132860$. O mesmo vale para os eventos E_2 e E_3, ao passo que o evento E_4 possui 365 elementos. Como $B = E_1 \cup E_2 \cup E_3 \cup E_4$ e os eventos $\{E_j : j = 1, 2, 3, 4\}$ são mutuamente exclusivos,

$$\#B = \#E_1 + \#E_2 + \#E_3 + \#E_4 = 3 \times 132860 + 365 = 398945.$$

Portanto, em uma sala com $n = 3$ pessoas, a probabilidade de ao menos um par de pessoas aniversariarem no mesmo dia é

$$P(B) = \frac{398945}{365^3} = \frac{398945}{48627125} \approx 0,0082.$$

Na tentativa de estender tal raciocínio para valores de n maiores que 3, é possível que se chegue, de maneira equivocada, em

$$\sum_{j=2}^{n} \frac{n!}{j!(n-j)!} \times \frac{365!}{(365-(n-j+1))!} = \sum_{j=2}^{n} C_{n,j} \times A_{365,n-j+1} \tag{8.3.1}$$

como forma geral para $\#B$. Tal equívoco se dá pelo fato da Equação (8.3.1) subestimar $\#B$. Por exemplo, para $n = 4$, a Equação (8.3.1) retorna 289.900.885 elementos, quando a cardinalidade de B para $n = 4$ é, na verdade, 290.299.465. Os 398.580 elementos não computados pela Equação (8.3.1) são os elementos do evento (contido em B) que contempla duas das quatro pessoas aniversariando no mesmo dia e as outras duas pessoas também fazendo aniversário num mesmo dia, porém, numa data diferente (por exemplo, o elemento $(02/ago, 19/nov, 19/nov, 02/ago)$). Portanto, não é difícil concluir que, conforme n aumenta, torna-se cada vez mais difícil o cálculo de $\#B$. Por este motivo, a solução do Problema do Aniversário volta-se ao cálculo do número de elementos do evento complementar

de B, isto é,

$$B^C = \{\text{“todas as } n \text{ pessoas na sala aniversariam em dias diferentes”}\},$$

e a probabilidade do evento B pode ser obtida por:

$$P(B) = 1 - P(B^C) = 1 - \frac{\#B^C}{365^n}.$$

Para qualquer $n \in \{2, 3, \ldots, 365\}$, é direto concluir que:

$$\#B^C = 365 \times \ldots \times (365 - (n-1)) = \frac{365!}{(365-n)!} = A_{365,n}.$$

Para salas com mais de 365 pessoas, o Princípio de Dirichlet (MORGADO et al., 1991) garante que ao menos 2 pessoas nesta sala fazem aniversário no mesmo dia, uma vez que há mais pessoas na sala que datas de aniversário possíveis. Dito isto, $P(B) = 1$ (ou ainda, $P(B^C) = 0$) para $n > 365$. Portanto, em uma sala com n pessoas, a probabilidade de ao menos um par de pessoas aniversariar no mesmo dia é dada por

$$P(B) = \begin{cases} 1 - \dfrac{365!}{365^n \times (365-n)!} & , \text{ se } n \leqslant 365, \\ 1 & , \text{ se } n > 365, \end{cases}$$

ou ainda, utilizando notação de número de arranjos simples,

$$P(B) = \begin{cases} 1 - \dfrac{A_{365,n}}{365^n} & , \text{ se } n \leqslant 365, \\ 1 & , \text{ se } n > 365. \end{cases}$$

Note que a solução acima vale para qualquer n inteiro positivo. Em parti-

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

cular, para $n = 1$, tem-se $P(B) = 0$, o que faz todo o sentido, pois é impossível um par de pessoas aniversariar no mesmo dia se não há sequer um par de pessoas nesta sala.

Uma vez desenvolvida e finalizada a solução do Problema do Aniversário, cabe o uso de alguma ferramenta computacional para calcular a expressão

$$1 - \frac{A_{365,n}}{365^n} \tag{8.3.2}$$

para o(s) valor(es) desejado(s) de n (ou ainda, para uma probabilidade pré-definida, calcular qual o menor n que faz a expressão acima ser maior ou igual a esta probabilidade). A ferramenta utilizada aqui é o *software* R, que consiste em um programa gratuito e de código aberto, comumente usado para tratamento de dados e análises estatísticas. De forma mais genérica, ele também pode ser pensado como uma linguagem de programação, em particular uma linguagem orientada a objetos. Seu *download* pode ser feito em (THE R PROJECT FOR STATISTICAL COMPUTING, 2023).

Uma forte característica do *software*/linguagem R é sua abundante quantidade de pacotes disponíveis. Em particular, dois pacotes que trazem consigo funções associadas ao Problema do Aniversário são objeto de estudo neste capítulo: os pacotes `IPSUR` e `RcmdrPlugin.IPSUR`. Pacotes no *software* R podem ser facilmente instalados por meio do menu *Pacotes* > *Instalar pacote(s)* (ou *Tools* > *Install packages*, para usuários do ambiente de desenvolvimento integrado Rstudio), ou ainda via linha de comando, na forma

```
install.packages("nome_do_pacote")
```

para apenas um pacote ou

```
install.packages(c("nome_do_pacote_1",...,"nome_do_
  pacote_n"))
```

para dois ou mais pacotes, desde que o(s) pacote(s) em questão esteja(m) no repo-

sitório CRAN (*Comprehensive R Archive Network*). Entretanto, nem todo pacote atualmente disponível para o R está listado neste repositório, uma vez que alguns pacotes que outrora pertenceram a ele foram, em algum momento, arquivados. Este é o contexto no qual se encontram os pacotes `IPSUR` e `RcmdrPlugin.IPSUR`. Felizmente, há como instalar pacotes ausentes do repositório CRAN, porém presentes no repositório *Github*, por meio da função `install_github` do pacote `devtools` (WICKHAM et al., 2022), o qual está disponível no repositório CRAN. Ainda em relação ao uso do pacote `devtools`, é necessário que usuários do sistema operacional Windows instalem previamente um programa auxiliar denominado Rtools, conforme explicitado no Passo 2 em (STATISTICAL COMPUTING, 2014), que também esclarece a necessidade da instalação do Xcode para usuários de Mac e de um compilador e várias bibliotecas de desenvolvimento para usuários de distribuições Linux. Os parágrafos seguintes desta seção direcionam-se aos procedimentos para instalação dos pacotes `IPSUR` e `RcmdrPlugin.IPSUR`, considerando que as instalações solicitadas acima tenham sido realizadas conforme o sistema operacional em uso.

O pacote `IPSUR`, cujo nome é uma sigla para *Introduction to Probability and Statistics Using R*, não está mais disponível no repositório CRAN desde 30 de maio de 2019, "pois os problemas de verificação não foram corrigidos a tempo, apesar dos lembretes" (NETWORK, Comprehensive R Archive, 2019). Logo, não é possível instalá-lo pelos métodos convencionais abordados no parágrafo anterior. Uma alternativa para a sua instalação se dá por executar as linhas de comando abaixo, conforme disponibilizado em (KERNS, 2022). Note que o pacote `prob` (KERNS, 2023b), uma das dependências do pacote `IPSUR`, também não está disponível no repositório CRAN, uma vez que é necessário instalá-lo por meio da função `install_github`. Tal pacote foi arquivado pelo CRAN em 29 de abril de 2022, "pois requer o pacote `fAsiaOptions`" (NETWORK, Comprehensive R Archive, 2022a), sendo que este último foi removido, na mesma data, por "deturpação da autoria e propriedade de direitos autorais" (COMPREHENSIVE R

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

ARCHIVE NETWORK, 2022). A dependência do pacote `fAsiaOptions` para o pacote `prob` foi removida pelo autor do último, sem perdas significativas.

```
# INSTALANDO PACOTES DO REPOSITORIO CRAN REFERENCIADOS
  NO IPSUR:
install.packages(c("actuar", "aplpack", "binom", "boot"
  , "coin", "diagram",
                   "distrEx", "e1071", "emdbook", "
                     ggplot2", "HH", "Hmisc",
                   "lmtest", "mvtnorm", "qcc", "reshape
                     ", "RcmdrMisc",
                   "scatterplot3d", "TeachingDemos", "
                     vcd"))

# INSTALAÇÃO DO PACOTE prob, QUE NÃO ESTÁ MAIS NO
  REPOSITÓRIO CRAN:
install.packages("combinat")  # combinat é uma dependê
  ncia do pacote prob
install.packages("devtools")  # caso devtools não
  esteja instalado ainda
devtools::install_github("gjkerns/prob")

# FINALIZANDO INSTALAÇÃO DO PACOTE IPSUR:
devtools::install_github("gjkerns/IPSUR")
```

O pacote `RcmdrPlugin.IPSUR` é um *plugin* do pacote `IPSUR` para a interface R Commander, fornecida pelo pacote Rcmdr (FOX; BOUCHET-VALAT, 2022). Mais detalhes sobre esta interface, incluindo instalação e carregamento do pacote Rcmdr, podem ser consultadas em (MELO, 2019). Em particular, há vários pacotes que são *plugins* para a interface R Commander no repositório CRAN,

todos iniciados em "RcmdrPlugin". Todavia, o pacote `RcmdrPlugin.IPSUR` não é mais um destes pacotes disponibilizados neste repositório desde 10 de agosto de 2022 por conta de "problemas que não foram corrigidos a tempo" (NETWORK, Comprehensive R Archive, 2022b). Assumindo que o pacote `devtools` já está instalado, conforme parte do passo a passo acima, o pacote `RcmdrPlugin.IPSUR` pode ser instalado simplesmente pela execução da linha de comando:

```
devtools::install_github("gjkerns/RcmdrPlugin.IPSUR")
```

Caso algum procedimento de instalação acima falhe por conta de versões antigas de outros pacotes, uma solução é acessar o menu *Pacotes* > *Atualizar pacotes* (ou, equivalentemente, o menu *Tools* > *Check for Packages Updates* do Rstudio) e atualizar todos os pacotes para suas versões mais recentes. Feito isto, refaça o procedimento de instalação de pacote que falhou.

8.4 RESULTADOS E DISCUSSÃO

Uma vez instalado o pacote `IPSUR`, carregue-o por meio dos comandos

```
library(IPSUR)
```

ou

```
require(IPSUR)
```

Para saber quais funções o pacote `IPSUR` fornece, basta executar a linha de comando

```
ls("package:IPSUR")
```

```
[1] "pbirthday.ipsur" "qbirthday.ipsur" "read"
```

Como pode ser visto acima, são apenas três funções. Quanto à função `read`, executar a linha de comando

```
read(IPSUR)
```

abre o arquivo *IPSUR.pdf*, que é a edição mais recente, tomando por base o momento de instalação do pacote, do livro *Introduction to Probability and Statistics Using R* (KERNS, 2021). No sistema operacional Windows, em particular, o arquivo *IPSUR.pdf* encontra-se na pasta de usuário em $AppData\backslash Local\backslash R\backslash win-library\backslash 4.3\backslash IPSUR\backslash doc$ (para usuários de versão 4.3.x do R e que optaram pela instalação padrão deste). Com 15 capítulos e 5 anexos distribuídos em cerca de 350 páginas, a quarta edição do livro *Introduction to Probability and Statistics Using R* traz o Problema do Aniversário em seu Exemplo 4.7, o qual é finalizado com um curto *script* para a geração de um gráfico com o número de pessoas na sala no eixo horizontal e a probabilidade de ao menos um par destas pessoas aniversariar num mesmo dia no eixo vertical. Tal *script* utiliza a função `pbirthday.ipsur`, uma das três funções fornecidas pelo pacote `IPSUR` e detalhada a seguir com seus valores *default*, tal como a função `qbirthday.ipsur`.

- `pbirthday.ipsur(n, classes=365, coincident=2)`: calcula a probabilidade de, numa sala com n pessoas, ao menos 2 fazerem aniversário no mesmo dia, considerando 365 possibilidades de dias.
- `qbirthday.ipsur(prob=0.5, classes=365, coincident=2)`: retorna o menor valor n cuja probabilidade de, numa sala com n pessoas, ao menos 2 fazerem aniversário no mesmo dia, considerando 365 possibilidades de dias, seja maior ou igual ao argumento `prob`.

Os valores *default* dos argumentos `classes` e `coincident` consistem, de fato, no Problema do Aniversário. As linhas de comando abaixo ilustram que a probabilidade de, em uma sala com n pessoas, pelo menos um par delas fazer aniversário no mesmo dia é (aproximadamente): 11,69% para $n = 10$, 50,73% para $n = 23$, 70,63% para $n = 30$, 89,12% para $n = 40$ e 97,04% para $n =$

50. Para $n = 100$, note que é necessário uma precisão de pelo menos 7 casas decimais para que a probabilidade solicitada não retorne 1. Para $n = 120$, note que a saída foi 1, porém isto é um valor aproximado, por conta do padrão de precisão do R ser 7 casas decimais. Alterando a precisão de 7 para 22 casas decimais, a linha de comando que anteriormente retornou 1 passa a retornar 0,9999999997560852227352. De fato, o evento "pelo menos duas de 120 pessoas aniversariarem em um mesmo dia" não é um evento certo para ter probabilidade 1, por mais improvável que seja a sua não ocorrência.

```
pbirthday.ipsur(10)
```

```
[1] 0.1169482
```

```
pbirthday.ipsur(23)
```

```
[1] 0.5072972
```

```
pbirthday.ipsur(30)
```

```
[1] 0.7063162
```

```
pbirthday.ipsur(40)
```

```
[1] 0.8912318
```

```
pbirthday.ipsur(50)
```

```
[1] 0.9703736
```

```
pbirthday.ipsur(100)
```

```
[1] 0.9999997
```

```
pbirthday.ipsur(120)
options(digits=22)
pbirthday.ipsur(120)
```

```
[1] 1
```

```
options(digits=22)
pbirthday.ipsur(120)
```

```
[1] 0.9999999997560852227352
```

Abaixo, seguem resultados obtidos pela função `qbirthday.ipsur`. Note que bastam 15 pessoas na sala para que a probabilidade de ao menos um par destas pessoas aniversariem no mesmo dia seja igual ou ultrapasse 25% (com 14 pessoas, tal probabilidade é menor que 25%). De maneira análoga, a justificativa da expressão "paradoxo do aniversário" se dá ao aplicar esta função em 50%, resultando em apenas 23 pessoas. Aplicando esta função em probabilidades cada vez maiores (75%, 90%, 95% e 99%), claramente os resultados são números cada vez maiores, porém ainda distantes de 365. Basta uma sala ter 57 pessoas, por exemplo, para que a probabilidade de ocorrer ao menos uma coincidência de aniversários atinja ou ultrapasse 99%. Por fim, note que `qbirthday.ipsur(1)` retorna 366, uma vez que este é o menor valor para n que faz com que o evento "Ao menos um par destas n pessoas faz aniversário no mesmo dia" seja um evento certo. Para $n = 365$, este evento não é um evento certo, em que pese ser extremamente improvável não ser satisfeito.

```
qbirthday.ipsur(0.25)
```

```
[1] 15
```

```
qbirthday.ipsur(0.50)
```

```
[1] 23
```

```
qbirthday.ipsur(0.75)
```

```
[1] 32
```

```
qbirthday.ipsur(0.90)
```

```
[1] 41
```

```
qbirthday.ipsur(0.95)
```

```
[1] 47
```

```
qbirthday.ipsur(0.99)
```

```
[1] 57
```

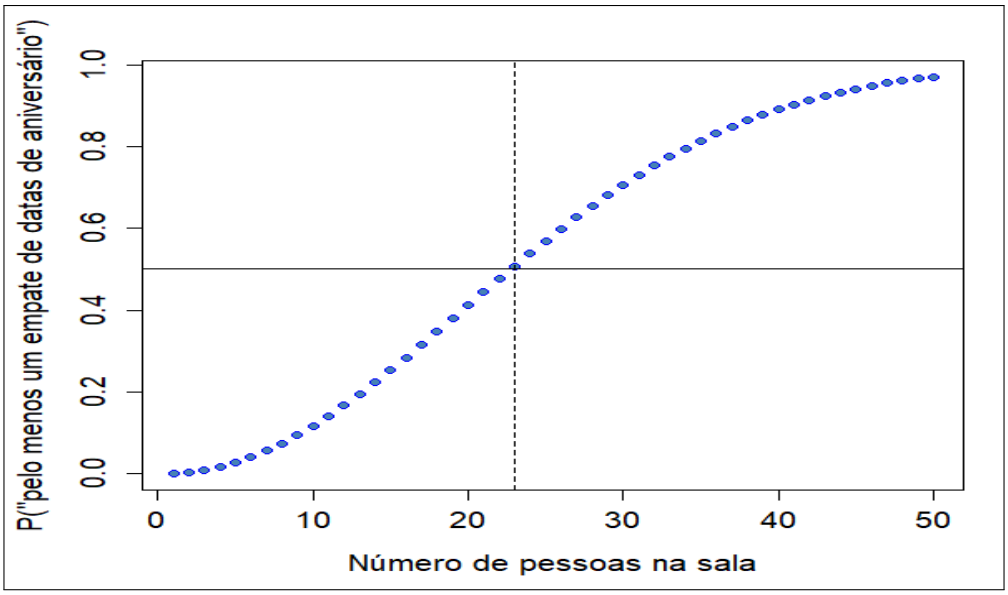
```
qbirthday.ipsur(1)
```

```
[1] 366
```

Voltando ao Exemplo 4.27 em (KERNS, 2021), mais especificamente ao *script* que encerra este exemplo, é possível verificar graficamente a evolução da probabilidade de interesse do Problema do Aniversário conforme n cresce. A Figura 8.1, baseada neste *script*, porém com algumas adaptações, ilustra tal evolução de $n = 1$

até $n = 50$, com destaque para $n = 23$, menor valor de n cuja probabilidade associada ultrapassa 50%.

Figura 8.1: Evolução da probabilidade de interesse do Problema do Aniversário conforme n cresce.



Fonte: O autor.

O *script* que gera a Figura 8.1 segue abaixo:

```
g <- Vectorize(pbirthday.ipsur)
plot(1:50, g(1:50), xlab = 'Número de pessoas na sala',
     ylab = 'P("pelo menos um empate de datas de
       aniversário")',
     pch=21, bg='steelblue', col='blue',cex.lab=1.25,
       cex.axis=1.25)
abline(h = 0.5)
abline(v = 23, lty = 2)
remove(g)
```

Exploradas todas as funções do pacote `IPSUR`, o interesse volta-se ao pacote `RcmdrPlugin.IPSUR`. Conforme já mencionado, este pacote é um *plugin* para a

interface R Commander, proveniente de carregamento do pacote Rcmdr. Como o pacote Rcmdr está disponível no repositório CRAN, sua instalação pode ser feita por um dos métodos convencionais mencionados na Seção 8.3, como por exemplo, executar a linha de comando

```
install.packages("Rcmdr")
```

Feita esta instalação, carregue o pacote Rcmdr por meio dos comandos

```
library(Rcmdr)
```

ou

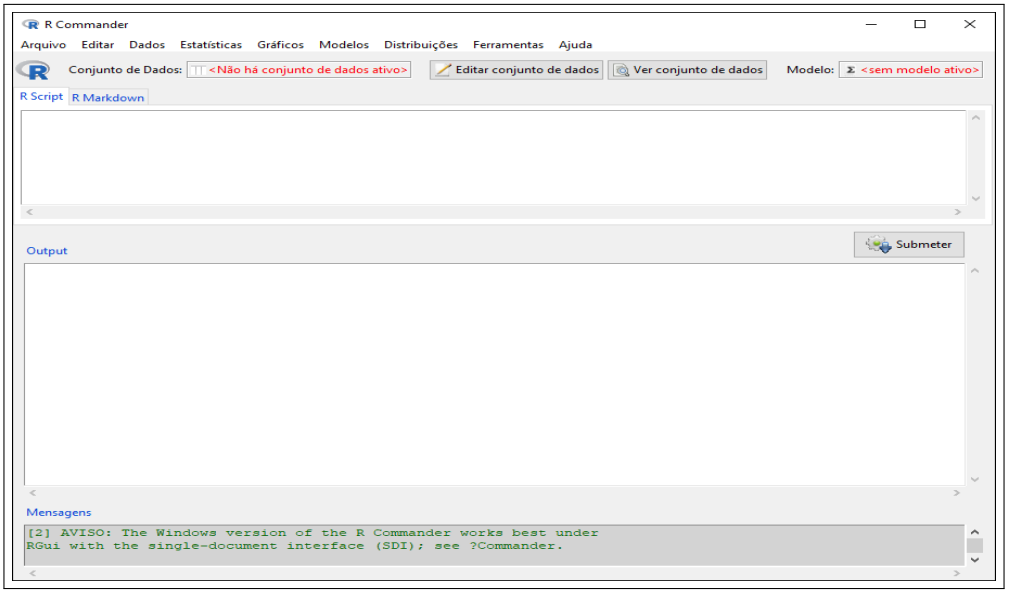
```
require(Rcmdr)
```

para que seja aberta a janela R Commander (Figura 8.2), que fornece ao usuário do R uma interface mais "amigável", com menus que propiciam várias funcionalidades do R sem a necessidade de escrever linhas de comando. Para dinâmicas em sala de aula envolvendo o Problema do Aniversário, seja em turmas de ensino básico ou superior, é recomendável o uso desta interface com menus, de forma a tornar a atividade mais prática, fácil e motivadora, inclusive para os estudantes mais interessados a reproduzirem em outros ambientes fora da sala de aula.

Inicialmente, nenhum dos menus da janela R Commander contém funcionalidades do *plugin* `RcmdrPlugin.IPSUR`, pois este ainda não foi carregado. Uma vez que a janela R Commander esteja aberta, qualquer um de seus *plugins* previamente instalados pode ser carregado por meio do menu (da janela R Commander) *Ferramentas* > *Carregar plug-in(s) do Rcmdr*. Feito este procedimento para carregar o *plugin* `RcmdrPlugin.IPSUR`, a interface R Commander é reiniciada e novas funcionalidades são adicionadas aos menus da janela R Commander, indicadas pelo rótulo "(IPSUR)", como o menu *Distribuições* > *Birthday Problem... (IPSUR)*, que possibilita o uso das funções `pbirthday.ipsur` e `qbirthday.ipsur` sem a necessidade de digitar de linhas de comando (Figura 8.3). Após escolher a opção desejada (Probabilidades ou Quantis) e preencher o campo associado à es-

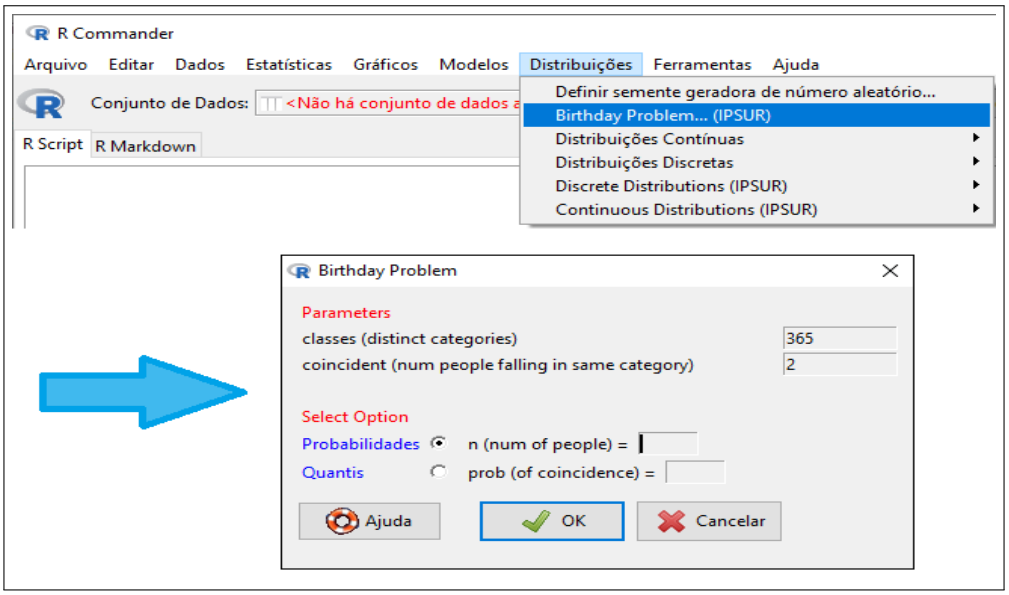
colha marcada, o resultado será mostrado na janela *Output* da janela R Commander. Caso esta janela *Output* não esteja em exibição, o resultado será mostrado na janela onde foi carregado o pacote Rcmdr.

Figura 8.2: Janela R Commander.



Fonte: O autor.

Figura 8.3: Menu do R Commander para o Problema do Aniversário, adicionado pelo *plugin* `RcmdrPlugin.IPSUR`



Fonte: O autor.

O que foi exposto acima considerou o carregamento do *plugin* do pacote `IPSUR` para o R Commander com esta interface aberta, via navegação por menus. Caso a janela R Commander não esteja aberta, carregar diretamente o pacote `RcmdrPlugin.IPSUR` abre a janela R Commander com as funcionalidades deste *plugin* carregado. À título de teste, feche a janela R Commander e a sessão atual do R. Abra uma nova sessão do R e carregue o pacote `RcmdrPlugin.IPSUR` por meio de uma das linhas da comando abaixo:

```
library(RcmdrPlugin.IPSUR)
```

ou

```
require(RcmdrPlugin.IPSUR)
```

Cabe ressaltar um ponto para novos usuários (ou usuários pouco experientes) da interface R Commander: quando a janela R Commander é fechada, não é possível reabri-la na atual sessão do R. Carregar novamente o pacote Rcmdr ou qualquer um de seus *plugins* na atual sessão não surte efeito. Neste cenário, é necessário abrir uma nova sessão do R e carregar novamente o pacote Rcmdr (ou o seu *plugin* de interesse) para acessar novamente a janela R Commander.

Além do menu voltado ao Problema do Aniversário, o pacote `RcmdrPlugin.IPSUR` adiciona ao R Commander outras funcionalidades nos menus *Distribuições*, *Gráficos* e *Estatísticas*, além de três conjuntos de dados que podem ser carregados no menu *Dados* > *Conjunto de dados em pacotes* > *Ler dados de pacote "attachado"*. Mais detalhes sobre as funcionalidades deste pacote seguem em (KERNS, 2014).

8.5 CONCLUSÃO

A solução do Problema do Aniversário pode parecer, a princípio, bastante complicada, sobretudo em salas com um número grande de pessoas, desde que não ultrapasse 365. Contudo, a contagem de n-uplas com ao menos uma coincidência de datas de aniversário é potencialmente facilitada ao se buscar pela diferença

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

entre o total de n-uplas possíveis e o total de n-uplas nas quais todas as datas são diferentes. Ao visualizar a Equação (8.3.2), é natural imaginar valores altos para n (superiores à metade de 365) quando esta equação é igualada a probabilidades maiores que 50%. Porém, conforme verificado na Figura 8.1 e nas saídas ilustradas pela função `qbirthday.ipsur`, tais valores de n estão muito aquém da metade de 365. Por exemplo, a probabilidade de ao menos um par de pessoas fazer aniversário num mesmo dia em uma sala com 41 pessoas ultrapassa 90%, e 41 é inferior a um oitavo de 365.

O fato da solução do Problema do Aniversário ser contraintuitiva torna este problema ainda mais interessante. Em uma sala de aula, o docente pode fazer o questionamento da probabilidade do evento de interesse do Problema do Aniversário considerando os estudantes presentes na sala. Supondo que há 30 alunos presentes, é intuitivo que eles atribuam ("chutem") probabilidades bem baixas, levando em conta que 30 é bem inferior a 365 e também supondo que cada estudante desconheça as datas de aniversário dos colegas. Todavia, para $n = 30$, a solução do Problema do Aniversário estabelece uma probabilidade de 70,63% de que há pelo menos duas pessoas aniversariando num mesmo dia. Ou seja, é bem provável que o professor, ao abordar os alunos, um por um, perguntando cada data de aniversário, receba como resposta, em algum momento, uma data que já foi mencionada, o que pode surpreender os alunos mais incrédulos. Uma dinâmica em sala de aula pode ser iniciada assim para que, na sequência, seja apresentado o Problema do Aniversário, sua solução e aplicações das funções `pbirthday.ipsur` e `qbirthday.ipsur` por meio da interface R Commander, com advento do pacote `RcmdrPlugin.IPSUR`. Ou ainda, para criar um clima de suspense e, desta forma, gerar maior expectativa e interesse da turma, a etapa da arguição das datas de aniversário dos estudantes pode ser deixada para o encerramento da dinâmica.

É importante reforçar, antes e após a solução do Problema do Aniversário, os três pressupostos elencados no começo da Seção 8.3. O primeiro deles é bem razoável, pois ainda que uma pessoa tenha nascido em um dia 29 de fevereiro, seu

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

registro de aniversário consta como 28 de fevereiro ou 1º de março. Sobre não ter gêmeos na sala, isto pode ser adaptado considerando que os irmãos gêmeos constituem um único elemento. Por exemplo, se há 30 alunos presentes na aula e, dentre eles, há um par de irmãos gêmeos, o problema deve ser considerado com $n = 29$ em vez de $n = 30$. O pressuposto restante (de assumir a mesma probabilidade para todas as 365 datas de aniversário possíveis) é essencial para a solução, pois é por conta disto que é possível obter a solução do problema pela interpretação clássica da probabilidade, isto é, para um evento E em um espaço amostral Ω, a probabilidade do evento E é dada por $P(E) = \#E/\#\Omega$. Contudo, dos três pressupostos, este é, na prática, o menos razoável, uma vez que podem haver períodos tipicamente com maior frequência de aniversários quando comparados a outros períodos com a mesma quantidade de dias, dentro de cada ano.

Sobre o desenvolvimento da solução do Problema do Aniversário com a turma, é importante que o docente tenha noção da base matemática da sua turma, seja ela de ensino médio ou ensino superior, sobretudo em conceitos básicos de análise combinatória e interpretação clássica da Probabilidade. Em turmas com maior dificuldade de compreensão destes conceitos, uma estratégia é adotar um conjunto de possibilidades menor. Em vez de 365 dias, adotar os 12 meses de ano, ou os 12 signos do zodíaco, ou ainda os 7 dias da semana. Desta forma, o argumento `classes` das funções `pbirthday.ipsur` e `qbirthday.ipsur` deve ser mudado (de 365 para 12 ou 7), e os pressupostos da solução do Problema do Aniversário devem ser devidamente adaptados, ainda que os meses do ano não tenham todos a mesma quantidade de dias, o mesmo ocorrendo para os períodos de cada signo. Entretanto, a identificação de elementos (n-uplas) de um espaço amostral que satisfazem ou não certo evento pode ser mais plausível quando este espaço amostral possui 7^n ou 12^n elementos em vez de 365^n elementos. Uma vez compreendida a forma geral da solução para ao menos uma coincidência de dias da semana, meses ou signos, a extensão do raciocínio para os 365 dias segue de

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

forma natural.

Por fim, cabe ressaltar que o *software* R possui, de forma nativa (pacote `stats`), as funções `pbirthday` e `qbirthday`, que retornam os mesmos resultados das funções `pbirthday.ipsur` e `qbirthday.ipsur` quando `coincident=2`. Porém, as principais motivações em apresentar e explorar o pacote `IPSUR` neste capítulo se dão pelo livro *Introduction to Probability and Statistics Using R* e como estudo preliminar para apresentar e explorar o pacote `RcmdrPlugin.IPSUR`, de forma a entender quais linhas de comando estão "por trás" do menu dedicado ao Problema do Aniversário propiciado por este pacote à interface R Commander.

8.6 REFERÊNCIAS

COMPREHENSIVE R ARCHIVE NETWORK. **Package 'fAsianOptions' was removed from the CRAN repository**. [S.l.: s.n.], 2022. Disponível em: https://CRAN.R-project.org/package=fAsianOptions. Acesso em: 31 jul. 2023.

FELLER, William. **An Introduction to Probability Theory and Its Applications: Vol 1**. New York, USA: John Wiley e Sons, Inc., 1950. p. 29–30.

FOX, John; BOUCHET-VALAT, Milan. **Rcmdr: R Commander**. [S.l.], 2022.

KERNS, Gary Jay. **Introduction to Probability and Statistics Using R**. [S.l.]: GNU Free Documentation License, 2021.

_____. **IPSUR: Introduction to Probability and Statistics Using R**. [S.l.: s.n.], 2022. Disponível em: https://ipsur.org/. Acesso em: 31 jul. 2023.

_____._____. [S.l.], 2023.

_____. **Package 'RcmdrPlugin.IPSUR'**. [S.l.: s.n.], 2014. Disponível em: http://cran.nexr.com/web/packages/RcmdrPlugin.IPSUR/RcmdrPlugin.IPSUR.pdf. Acesso em: 31 jul. 2023. R package version 0.2-1.

_____. **prob: Elementary Probability on Finite Sample Spaces**. [S.l.], 2023.

_____. **RcmdrPlugin.IPSUR: An IPSUR Plugin for the R Commander**. [S.l.], 2019.

MELO, Felipe Rafael Ribeiro. **Introdução ao R Commander: Notas de Aula**. Rio de Janeiro, Brasil, 2019.

MORGADO, Augusto César de Oliveira et al. **Análise Combinatória e Probabilidade**. Rio de Janeiro: IMPA, 1991. p. 81.

NETWORK, Comprehensive R Archive. **Package 'IPSUR' was removed from the CRAN repository**. [S.l.: s.n.], 2019. Disponível em: https://CRAN.R-project.org/package=IPSUR. Acesso em: 31 jul. 2023.

_____. **Package 'prob' was removed from the CRAN repository**. [S.l.: s.n.], 2022. Disponível em: https://CRAN.R-project.org/package=prob. Acesso em: 31 jul. 2023.

_____. **Package 'RcmdrPlugin.IPSUR' was removed from the CRAN repository**. [S.l.: s.n.], 2022. Disponível em: https://CRAN.R-project.org/package=RcmdrPlugin.IPSUR. Acesso em: 31 jul. 2023.

SINGMASTER, David. **Sources in Recreational Mathematics: An Annotated Bibliography**. [S.l.: s.n.], 2004. Disponível em: https://www.puzzlemuseum.com/singma/singma6/SOURCES/singma-sources-edn8-2004-03-19.htm#_Toc69533810. Acesso em: 31 jul. 2023.

STATISTICAL COMPUTING, The R Project for. **devtools**. [S.l.: s.n.], 2014. Disponível em: https://www.r-project.org/nosvn/pandoc/devtools.html. Acesso em: 31 jul. 2023.

TEAM, R Core. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023.

THE R PROJECT FOR STATISTICAL COMPUTING. **Comprehensive R Archive Network (CRAN)**. [S.l.: s.n.], 2023. Disponível em: https://cloud.r-project.org/. Acesso em: 31 jul. 2023.

WICKHAM, Hadley et al. **devtools: Tools to Make Developing R Packages Easier**. [S.l.], 2022.

Capítulo 9

CLASSIFICATION OF GALICIAN SURNAMES WITH WEB SCRAPING

Autor: Maria José Ginzo Villamayor[1]
Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela.
e-mail: mariajose.ginzo@usc.es

Linguistics considers different classifications of surnames according to their motivation, morphology or semantics. In the case of Galician surnames, Boullón-Agrelo (2008) proposes a classification based on three main groups: appellatives, patronymics and toponymics. In order to classify Galician surnames in these three categories, Web Scraping techniques were used, i.e. a process of extracting content and data from websites, scraping official Galician, Spanish and even Portuguese language dictionaries. These techniques were very useful, especially for appellatives.

Key-words: Surnames; Galicia; Web Scraping; Directional Highest Density Regions; clustering directional.

[1]M.J. Ginzo-Villamayor acknowledges the financial support of Agencia Estatal de Investigación (AEI) del Ministerio de Ciencia e Innovación under grant PID2020-116587GB-I00.

9.1 INTRODUCTION

Linguistics considers different classification of surnames depending on motivation, morphology or semantic. The case more habitual is the semantic, being this last one the most frequent case. (BOULLÓN-AGRELO, 2008) suggests the following classification for Galician surnames: patronymic, toponymic and apelative.

(a) *Patronymic* ending in "–ez", comes from a proper name. For example González means son of Gonzalo. The case of the Portuguese is similar: surnames formed by adding "–es" mean "son of", for example, for Gonzalo the corresponding surname is Gonzales. But not only in the case of languages from Iberian Peninsula, in other languages, patronymic surnames also exist, for example, in Hungarian: the suffix "–i" adjusted to a place- name expresses origin, or to a personal name. The most common method French to form surnames are surnames bases on parent's name, in this case called patronymic and matronymic surnames. The majority of French patronymic and matronymic surnames have no identifying prefix, but in some cases also attach a prefix or suffix that means "son of". In the case of patronymic English surname the suffixes "-son/-s/-kin/-kins/-ken" at the end denote "son o'" or "little". In German surnames the suffix "–sen" means "son of". The Slavic "–ke/–ka" suffix means "son of".

(b) *Toponymic* derives from a place name. There are some toponymics that may be polygenetic (originate in several places) and others can have a unique and local origin (more interesting for this case); among these are found: Cures, Cidrás, Cartelle, Orille, Mourente, Sandiás, Berdiñas, Ageitos. En the case of Galicia, it is very useful to use the Cartography of surnames[2]. In other languages, patronymic surnames also exist, for example, in Hungarian: the suffix "–i" to be found in most of the following examples is regularly attached

[2]Cartography of surnames in Galicia, http://ilg.usc.es/cag/.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

to place names when deriving a surname from it. Its meaning is "to be of, to be from". Finnish surnames which end in "–nen" mean the place where a family lived. Also for Galician and Spanish surnames with particle "De" or "Del" at he beginning "to be of, to be from" for example De Barros, De Villanueva, De León, Del Moral.

(c) Those that have origin in *common names* (professions, characteristics, etc.), physical, nicknames, etc.) for example: Veloso, Blanco, Cordeiro, Negro, Louzao (and Louzán), Conde, Santos. In the previous languages (Portuguese, Hungarian, English, French, ...), there are also this type of surnames.

9.1.1 Surnames, words and language

It should not be forgotten that a surname is still a word, and as such, if it had a meaning, it would appear in a dictionary with its pertinent definition.

It is known that there are a number of surnames that sound phonetically the same but are spelt differently, whose meaning may or may not be the same. Therefore, in a preliminary way, it has been developed a procedure programmed in language *R*, which goes through all the surnames in Galicia and compares them one by one to see if they only differ in one letter, these changes can be changing the letter "b" for the letter "v", as it happens, for example, in the following cases: ALBES vs. ALVES Another example is the change of the letter "c" for the letter "z", when they have a similar pronunciation, as in the case of CELADA vs. ZELADA. Sometimes the change of the letter "c" for the letter "z", although they are not pronounced the same in Spanish or Galician, could be misprints or derivations from Portuguese where the letter "c" could be considered as a "ç". Many words that in Portuguese have the letter "ç" in Spanish or Galician become "z", such as MOUCO vs. MOUZO.

Sometimes the letter "g" and the letter "j" are pronounced similarly, as in the following cases: AGEITOS vs. AJEITOS, AGENJO vs. AJENJO, BORGES

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

vs. BORJES, CEREIGIDO vs. CEREIJIDO, FREIJEDO vs. FREIGEDO, GEREMIAS vs. JEREMIAS, TEIJEIRO vs. TEIGEIRO or the case of VALIJE vs. VALIGE. There are changes of the letter "g" to the letter "j" and they do not correspond to phonetic changes: GAJINO vs. JAJINO.

The letter "y" represents two different phonemes: one equivalent to the letter "i" in surnames like DALI vs. DALY.

It is well known that many words that in Spanish begin with the letter "h" in Galician are written with the letter "f", this is the case of surnames, HIDALGO vs. FIDALGO.

Other curious changes are those of the letter "c" to the letter "q", as reflected in the following surnames: NAVASCUES vs. NAVASQUES.

Usually in Spanish, "m" is always written before the phoneme /p/, as in the case of surnames: PAMPIN, SAMPAYO or SAMPEDRO. Even so, we find the following surnames with the letter "m" before the phoneme /p/: PANPIN, SANPAYO or SANPEDRO. Always write "m" before the phoneme /b/ when it is represented by the letter "b", as in the case of the surnames: ARAMBURO, BRUMBECK, CAMBA, CUMBRADO, MIÑAMBRES, SAMBLAS, TEMBRAS, or WONEMBURGER. Even so, we find the following surnames with the letter "n" before the phoneme /p/: ARANBURO, BRUNBECK, CANBA, CUNBRADO, MIÑANBRES, SANBLAS, TENBRAS or WONENBURGER.

The following groupings of letters are characteristic particles of the Galician language: "-AI-", "-EI-", "-IÑA", "-IÑO" or "-OU-". Thus, their are the following examples of surnames with "-AI-": ABELAIRA, CASAIS or GAITERO. With "-EI-": ACEIRO, BANDEIRA, CABECEIRO, ESPIÑEIRA, GRUEIRO or MACIÑEIRA. With "-IÑA/O", indicating diminutives or affection: AGUSTIÑO, ALBARIÑAS, BESIÑO, CALVIÑO, LAVARIÑAS, LOURIÑO, PATIÑO or TROITIÑA. With "-OU-": BOUZA, COUCE, DOURADA, LOUREIRO, MOURO or VILOUTA.

The Seminario de Onomástica da Real Academia Galega published the book

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

"Os apelidos en Galego" ((RAG, 2016)), a collection of 1500 surnames, representing almost 9% of the population, of Galician tradition, chosen according to their authors by a frequency criterion. In addition, examples of criteria for the standardisation of language-related surnames are presented in (RAG, 2016).

The goal is to characterize surnames according to a certain taxonomy and identify patronymic, apelative, toponymic, as well as finer analysis, foreign, nature-related, ... surnames.

Different methods have been used for this purpose. For the patronymic surnames, we searched for surnames ending in "-ez" and the rest of the endings ("az", "iz" or ''oz", in this case) and thus tried to find out the name from which they come, since, as we have already mentioned, endings of the "-ez" type mean son of. Even so, it cannot be said that all surnames that do not contain one of these endings are not patronymic, as for example the surnames, GARCIA, ALONSO, VICENTE, JORGE, are also patronymic surnames. This type of surname was the first to appear. For apelative surnames, a list of adjectives related to physical or psychological characteristics of persons or family relationships or professions is proposed and those that match are searched for in the set of surnames. This list has been compiled taking into account the criteria for the appearance of this type of surname, since the patronymic surnames were not sufficient. In order to identify toponymic surnames, the surname data set is crossed with the gazetteer data (called "Nomenclátor de Galicia"). Those that coincide in both data sets are studied to see if they are really toponyms. These surnames were incorporated with reference to the origin of the person. For the finer taxonomy, lists have been created with words related to land, vegetation, buildings, animals, etc.

This procedure is just carried out "by hand" and it is proposed to use a more automatic procedure. All of the tools are completely described in (GINZO-VILLAMAYOR, 2022), Section *Weighted Distance*. Mainly, for this part, the R packages ((R CORE TEAM, 2020)) *sqldf* ((GROTHENDIECK, 2017)) and *dplyr* ((WICKHAM et al., 2023)) are used.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Web Scraping

Web Scraping is a technique for converting the data present in unstructured format (HTML tags) on the web to a structured format which can easily be accessed and used. The web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. While Web Scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

Web crawling is a main component of Web Scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet or loaded into a database. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. In particular, Web Scraping tools described in (GINZO-VILLAMAYOR, 2022) have been used to our dataset on the 3-dimensional unit sphere.

Almost all the main programming languages provide ways for performing Web Scraping. In this work, it was used language R for scraping the data for the dictionaries websites. It can be obtained using the *R* ((R CORE TEAM, 2020)) routines `read_html` and `html_nodes` availables in the package *rvest* (see (WICKHAM, 2022)) or *Rcrawler* (see (KHALIL, 2018)).

Once the Web Scraping has been carried out, it is a matter of looking for those surnames that appear in the dictionaries and analysing their definition, to see if they can be assigned to any known taxonomy. Even so, an exact word may not appear and a derivative may appear, and this has also been taken into account when applying this technique.

The Web Scraping technique has been applied to the dictionaries in the Diccionario de la lengua española - Real Academia Española (RAE)

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

https://dle.rae.es/; Diccionario de la lengua galega - Real Academia Galega (RAG) https://academia.gal/dicionario; Dicionário Priberam da Língua Portuguesa (DPLP) https://dicionario.priberam.org/. The Galician dictionary is used because the Galician surnames (The Lei de normalización lingüística (1983) recognises that the only official form of place names is Galician, and one of the first tasks was the creation of a new gazetteer that would include the traditional toponymy and adapt it to the rules of the Galician language, a task that was completed, as far as the major place names (municipalities, parishes, villages and places) are concerned, with the publication of the *Nomenclátor de Galicia* in 2003.) are used, and it could happen that one of them means a word, as well as the use of the Spanish dictionary due to the process of Castilianisation, among others. Due to the influence and proximity of Portugal, it has been considered interesting to use the Portuguese dictionary.

After applying the Web Scraping technique combined with the previously described manual procedure and exchanging conversations with Ana Boullón Agrelo, expert in onomastics at the University of Santiago de Compostela and member of the Instituto da Lingua Galega (ILG), 1711 surnames have been classified into the three large groups, which represent more than 85% of the Galician population.

9.2 OBJECTIVE

The main objective of this work is to classify Galician surnames in the previous three categories using Web Scraping techniques from websites, scraping official Galician, Spanish, and even Portuguese language dictionaries. Once classified, apply directional data techniques to compare the results with others obtained using isonymy measures (possession of the same surname), such as Lasker, Nei, isonymy between ((RODRÍGUEZ-LARRALDE et al., 2003) or (SCAPOLI et al., 2007)).

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

9.3 APPLICATION

The objectives of this Section was applied Directional Highest Density Regions (HDiR) to groups of surnames and represent them on sphere. The second objective was applied spherical clustering. These two techniques will be applied to the set of surnames, once they have been classified, after applying the Web Scraping techniques.

9.3.1 Directional Highest Density Regions

Highest Density Regions estimation in $\mathbb{R}^d$: Given a random sample of points $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ of a random vector X with values in $\mathbb{R}^d$, reconstructing the $t-$ level set

$$G(t) = \left\{x \in \mathbb{R}^d : f(x) \geq t\right\}$$

where f denotes the density function of X and $t > 0$. Or, if the practitioner fixes a value $\tau \in (0, 1)$, estimating the Highest Density Region (HDR) with probability content $1 - \tau$

$$L(\tau) = \left\{x \in \mathbb{R}^d : f(x) \geq f_\tau\right\}$$

where f_τ can be seen as the largest constant such that

$$\mathbb{P}(X \in L(\tau)) \geq 1 - \tau$$

with respect to the distribution induced by f.

There is a method for estimating a HDR, the plug-in methodology. Plug-in methods propose

$$\hat{L}(\tau) = \left\{x \in \mathbb{R}^d : f_n(x) \geq \hat{f}_\tau\right\}$$

as an estimator for $L(\tau)$ where

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_H\left(x - X_i\right)$$

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

where K is a symmetric density function with $K_H(z) = |H|^{-1/2} K\left(H^{1/2} z\right), H$ denotes the bandwidth matrix and $\hat{f}_\tau = f_\tau\left(f_n\right)$ denotes an estimator of the threshold f_τ. (HYNDMAN, 1996) estimated f_τ as the quantile τ of the empirical distribution of $f_n\left(X_1\right), \cdots, f_n\left(X_n\right)$.

HDRs estimation in $\mathbb{S}^{d-1}$: Given a random sample of points $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ of a random vector Y with values in the unit sphere S^{d-1}, reconstructing the $t-$ level set

$$G_g(t) = \left\{y \in S^{d-1} : g(y) \geq t\right\}$$

where g denotes the directional density function of Y and $t > 0$. Or, if the practitioner fixes a value $\tau \in (0, 1)$, estimating the Highest Density Region (HDR) with probability content $1 - \tau$

$$L_g(\tau) = \left\{y \in S^{d-1} : g(y) \geq g_\tau\right\}$$

where g_τ can be seen as the largest constant such that

$$\mathbb{P}\left(Y \in L_g(\tau)\right) \geq 1 - \tau$$

with respect to the distribution induced by g.

There is a method for estimation of directional HDRs. Plug-in methods ((HYNDMAN, 1996; SAAVEDRA-NIEVES, P.; CRUJEIRAS, R. M., 2021)) propose

$$\hat{L}_g(\tau) = \left\{y \in S^{d-1} : g_n(y) \geq \hat{g}_\tau\right\}$$

as an estimator for $L_g(\tau)$ where

$$g_n(y) = \frac{1}{n} \sum_{i=1}^{n} K_{vM}\left(y; Y_i; 1/h^2\right)$$

where $1/h^2 > 0$ is concentration parameter and K_{vM} denotes the von Mises-Fisher kernel density.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

9.3.2 Mixtures of von Mises-Fisher Distributions

This Section presents a data representation in the hypersphere ((MARDIA; JUPP, 2000)) and the application to surname data. Several large-scale data mining applications, such as text categorization and gene expression analysis, deal with high-dimensional data that can be represented on a unit hypersphere ((BANERJEE et al., 2005)).

A hypersphere is a generalisation of a sphere to higher dimensions, denoted by S^n, and can be understood as $S^n = \left\{x \in \mathbb{R}^{n+1}; \|x\| = \alpha\right\}$, for a particular $\alpha \in \mathbb{R}^+$ constant. Independently of α, a normalisation can be made in order to work with a unitary hypersphere, $\alpha = 1$. The case $n = 1$ refers to the circle.

For the case of the sphere, using polar coordinates, it is sufficient to use two different angles in order to cover the set of possible values. One of these will be the variable called longitude ϕ and the other latitude $\frac{\pi}{2} - \theta$, the parameterisation used is as follows:

$$x = (\cos\theta,\, \text{sen}\theta\cos\phi,\, \text{sen}\theta\,\text{sen}\phi).$$

To use this parameterisation, in order to guarantee that two different values of θ and ϕ do not result in the same point (except for 0 and π), restricted to $\phi \in [0, 2\pi]$ and $\theta \in [0, \pi]$. Note that this last interval is not $[0, 2\pi]$ because then we would have the possibility of obtaining a θ' such that $\cos\theta' = \cos\theta$ and $\text{sen}\,\theta' = -\,\text{sen}\,\theta$. Then, there exists ϕ' such that $\sin\phi' = -\,\text{sen}$ and $\cos\phi' = -\cos\phi$ so that they are in the same coordinates. Once we have the polar coordinates, we can consider their projection in the plane.

Modelling data in the sphere: Random variables supported on a sphere can be modelled by different distributions. The most important one is the von Mises-Fisher (vMF) distribution. (BANERJEE et al., 2005) proposes a generative mixture-model approach to clustering directional data based on the vMF distribution, which arises naturally for data distributed on the unit hypersphere.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Let's assume a sample of $x_1, \cdots, x_n$ of data obtained from a sphere. The sample mean in polar version, $\bar{x} = \bar{R}\bar{x}_0$, where $\bar{R}$ is the norm of the mean $\bar{x}$. Thus, when considering the sample mean, if we have two or more distinct observations, we would obtain that $\bar{R} < \alpha$, or in the unitary case $\bar{R} < 1$, or in the unitary case x, we can define its mean length as

$$\rho = \left(\sum_{i=1}^{n} \mathbb{E}\left[x_i\right]^2 \right)^{\frac{1}{2}},$$

and if $\rho > 0$ satisfies, a mean direction can also be defined $\mu = \rho^{-1}\mathbb{E}[x]$, μ corresponds to the normalised mean, i.e. it would indicate the direction and direction of the mean.

9.4 RESULTS AND DISCUSSIONS

9.4.1 Application HDiR to surnames data

A total of 1711 surnames have been classified, representing 8.15% of the total number of surnames and 86.15% of the population, using the Web Scraping technique (Section 9.1. 9.1.1). The procedure carried out is as follows: once the surnames have been classified into 3 groups: apelative, toponymic or patronymic, for each municipality in Galicia the population has been distributed according to these 3 groups, that is to say, to have the population distributed in these 3 groups. Figure 9.1 shows the HDiR from toponymic, patronymic and apelative surnames obtained with `HDiR` package (see (SAAVEDRA-NIEVES, Paula; CRUJEIRAS, Rosa M, 2022)). This package is a R tool for nonparametric plug-in estimation of Highest Density Regions (HDRs) in the directional setting (SAAVEDRA-NIEVES, P.; CRUJEIRAS, R. M., 2021). Table 9.1 shows descriptive statistics for the percentages of the different types of surnames for all councils.

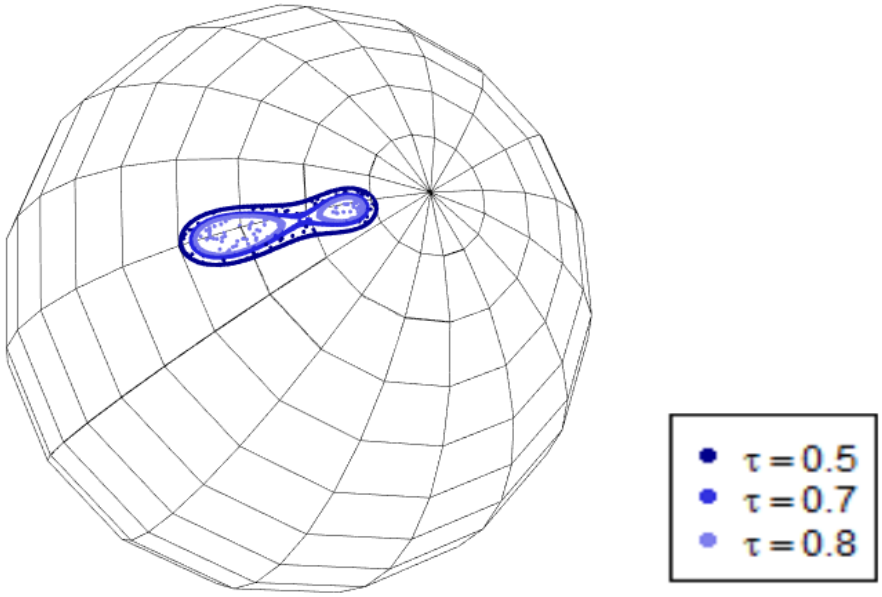
Consider $\tau = 0.8$, on Figure 9.1 and the aim is to find out which municipalities are in each of the two connected components. Figure 9.2 shows the councils

Tabela 9.1: Descriptive statistics for the percentages of the different types of surnames for all councils in Galicia.

	Min.	Median	Mean	Max.
Apelative	1%	13%	13%	35%
Patronymic	31%	58%	55%	92%
Toponymic	3%	28%	30%	56%

Source: (GINZO-VILLAMAYOR, 2022).

Figura 9.1: HDiR from toponymic, patronymic and apelative surnames.



Source: (GINZO-VILLAMAYOR, 2022).

in Galicia,

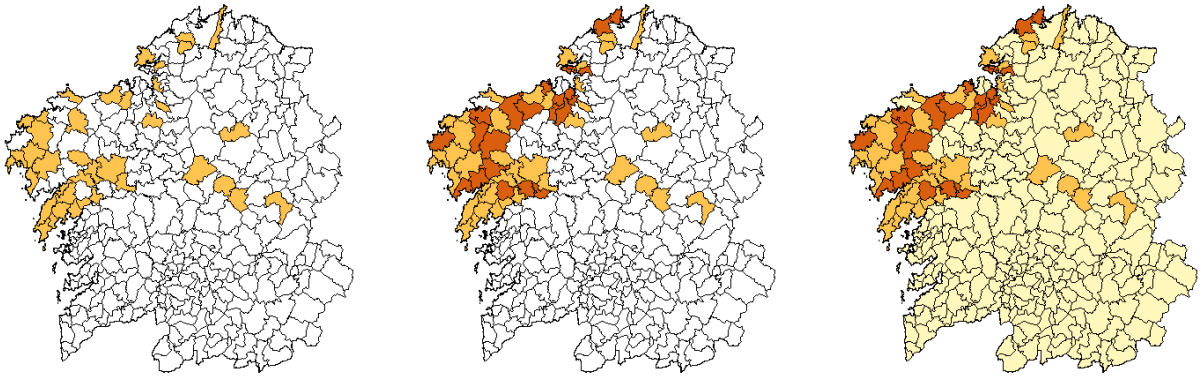
Table 9.2 shows descriptive statistics for the percentages of the different types of surnames, in the related components and for the rest.

Tabela 9.2: Descriptive statistics for the percentages of the different types of surnames, in the related components and for the rest.

	Min.	Median	Mean	Max.	Min.	Median	Mean	Max.	Min.	Median	Mean	Max.
Apelative	12%	17%	17%	28%	1%	17%	17%	29%	3%	12%	13%	35%
Patronymic	36%	46%	46%	61%	35%	46%	50%	91%	31%	59%	60%	92%
Toponymic	21%	36%	36%	47%	8%	33%	33%	44%	3%	27%	27%	56%

Source: (GINZO-VILLAMAYOR, 2022).

Figura 9.2: Left: councils for $\tau = 0.8$ (left connected components - Figure 9.1). Middle: councils for $\tau = 0.8$ (left and right connected components - Figure 9.1). Right: all councils in Galicia.



Source: (GINZO-VILLAMAYOR, 2022).

Mixtures of von Mises-Fisher Distributions

The Mixtures of von Mises-Fisher Distributions fit was obtained with `movMF` package (see (HORNIK; GRÜN, 2022)). This package is a R tool for fit and simulate mixtures of von Mises-Fisher distributions. After this fit, an analysis cluster was carried out.

The clustering obtained from the a-posteriori probabilities is analyzed by comparing the cluster membership with the surnames assigned to a council. Because each surname might have several councils assigned, the surnames and their cluster assignments are suitably repeated.

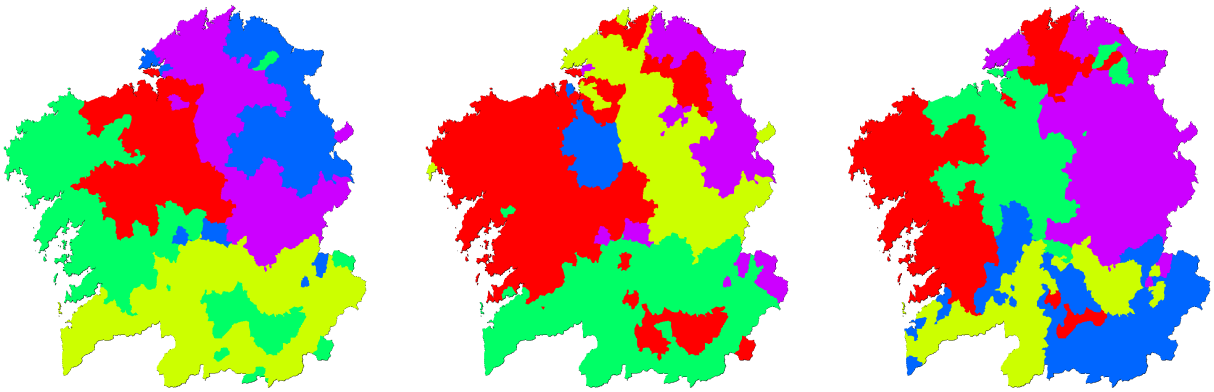
Figures 9.3show the results of the spherical cluster analysis, in 3 different scenarios: with all data (left), only those born in 1965 or earlier (center), and only those born in 1945 or earlier (right).

Figures 9.4 show the results of the spherical cluster analysis, in last 3 different scenarios and removing those ones below and above the 5% and 95% quantiles of the distribution of number of councils. Once these filters are applied, the clusters obtained are more compact, and similar to those obtained in the Lasker distance (LASKER, 1977) cluster.

The results obtained with both techniques are similar to those obtained in

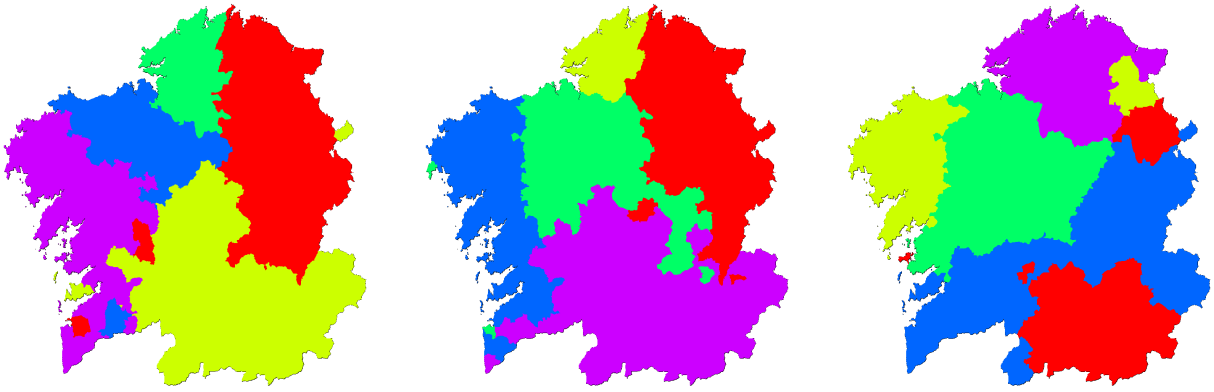
APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Figura 9.3: Spherical cluster results considering all data (left); Spherical cluster results filtering by population born before 1965 (center); Spherical cluster results filtering by population born before 1945 (right).



Source: (GINZO-VILLAMAYOR, 2022).

Figura 9.4: Spherical cluster results considering all data (and removing those ones below and above the 5% and 95% quantiles of the distribution of number of councils) (left); Spherical cluster results filtering by population born before 1965 (and removing those ones below and above the 5% and 95% quantiles of the distribution of number of councils) (center); Spherical cluster results filtering by population born before 1945 (and removing those ones below and above the 5% and 95% quantiles of the distribution of number of councils) (right).



Source: (GINZO-VILLAMAYOR, 2022).

the regionalisation of surnames for Galicia (GINZO-VILLAMAYOR, 2022). In this case, what has been carried out is a cluster analysis once isonymy measures (called also onomastic distances) have been applied to the surname dataset. Some preliminary results reveal a unique and evidence-based regional geography that is

of use in improving our understanding of cultural and social history. The research also contributes a range of methodological insights for future studies concerning spatial clustering of surnames.

9.5 REFERÊNCIAS

BANERJEE, A. et al. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. **Journal of Machine Learning Research**, v. 6, n. 46, p. 1345–1382, 2005.

BOULLÓN-AGRELO, Ana Isabel. I nomi nel tempo e nello spazio - V. Atti del XXII Congresso Internazionale di Scienze Onomastiche Pisa. In: [s.l.]: Edizioni ETS, 2008. v. II The surnames in Galicia today: a characterization and description, p. 299–310. ISBN 9788846729545.

GINZO-VILLAMAYOR, M. J. **Statistical Techniques in Geolinguistics. Onomastic modeling**. 2022. Tese (Doutorado) – Universidade de Santiago de Compostela.

GROTHENDIECK, G. **sqldf: Manipulate R Data Frames Using SQL**. [S.l.: s.n.], 2017. Disponível em: https://CRAN.R-project.org/package=sqldf. R package version 0.4-11.

HORNIK, Kurt; GRÜN, Bettina. **movMF: Mixtures of von Mises-Fisher Distributions**. [S.l.: s.n.], 2022. Disponível em: https://CRAN.R-project.org/package=movMF. R package version 0.2-7.

HYNDMAN, R.J. Computing and graphing highest density regions. **The American Statistician**, v. 50, p. 120–126, 1996.

KHALIL, Salim. **Rcrawler: Web Crawler and Scraper**. [S.l.], 2018. R package version 0.1.9-1.

LASKER, G. W. A coefficient of relationship by isonymy: A method for estimating the genetic relationship between populations. **Human Biology**, v. 49, p. 489–493, 1977.

MARDIA, K. V.; JUPP, P. E. **Directional Statistics**. [S.l.]: John Wiley & Sons, 2000.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020.

RAG. **Os apelidos en galego. Orientacións para a súa normalización**. [S.l.]: Real Academia Galega coa colaboración da Secretaría Xeral de Política Lingüística, 2016.

RODRÍGUEZ-LARRALDE, A. et al. The names of Spain: a study of the isonymy structure of Spain. **American Journal of Physical Anthropology**, Wiley Online Library, v. 121, n. 3, p. 280–292, 2003.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

SAAVEDRA-NIEVES, P.; CRUJEIRAS, R. M. Nonparametric estimation of directional highest density regions. **Advances in Data Analysis and Classification**, 2021.

SAAVEDRA-NIEVES, Paula; CRUJEIRAS, Rosa M. **HDiR: Directional Highest Density Regions**. [S.l.: s.n.], 2022. Disponível em: https://CRAN.R-project.org/package=HDiR. R package version 1.1.3.

SCAPOLI, Chiara et al. Surnames in Western Europe: A comparison of the subcontinental populations through isonymy. **Theoretical Population Biology**, v. 71, p. 37–48, 2007.

WICKHAM, Hadley. **rvest: Easily Harvest (Scrape) Web Pages**. [S.l.: s.n.], 2022. Disponível em: https://CRAN.R-project.org/package=rvest. R package version 1.0.3.

WICKHAM, Hadley et al. **dplyr: A Grammar of Data Manipulation**. [S.l.: s.n.], 2023. Disponível em: https://CRAN.R-project.org/package=dplyr. R package version 1.1.1.

Capítulo 10

USO DO DUCKDB COM R

Autor: Thiago de Oliveira Pires
F&O/CIO/International Business Machines Corporation
e-mail: thop100@hotmail.com

Muitos já tiveram o problema, principalmente relacionado à memória do computador, ao tentar ler e manipular uma base de dados muito grande. Existem várias soluções para lidar com grandes bases de dados, mas uma tem chamado bastante atenção atualmente: `duckdb`. O `duckdb` é um banco de dados simples de operar, por ser um banco de dados embarcado/embutido no estilo SQLite. Contudo, diferentemente do SQLite, o `duckdb` tem suporte para mais de 20 tipos de variáveis. Alguns destes tipos são aninhados, como lista, estrutura e *map*. Pode-se fazer consultas em arquivos `csv` e parquet diretamente. Consultas até em arquivos online salvos em um serviço de armazenamento na nuvem (por exemplo, S3). Tem suporte também para consultas em arquivo `json`. Além de toda a sintaxe padrão do SQL, ele tem algumas funções adicionais e suporta até a sintaxe de *list comprehension* do Python. Este capítulo irá apresentar alguns recursos do `duckdb` e sua interação com a linguagem R. **Palavras-Chave:** duckdb; SQL; R; banco de dados.

10.1 INTRODUÇÃO

O `duckdb` é um sistema de gerenciamento de banco de dados (SGBD) de código aberto (i.e. MIT), escrito em C++ e otimizado para consultas analíticas. Ele foi projetado para fornecer alta velocidade e eficiência em consultas complexas em grandes conjuntos de dados (DUCKDB DEVELOPMENT TEAM, 2023).

Existem duas características principais que distinguem o `duckdb` de outras ferramentas de análise de dados: a arquitetura colunar e o processamento vetorizado. A arquitetura colunar permite armazenar dados em colunas separadas, otimizando o acesso aos dados necessários para consultas específicas. Já o processamento vetorizado realiza operações em lotes, aproveitando as otimizações de hardware e reduzindo a latência de acesso à memória. Com isso, produzem-se consultas mais rápidas e eficientes (RAASVELDT; MÜHLEISEN, 2019).

Outra vantagem do `duckdb` é a sua capacidade de compressão de dados. Ele utiliza algoritmos de compressão especializados, resultando em economia de espaço em disco e também contribuindo com uma maior velocidade na consulta (RAASVELDT, 2022).

Existem várias *API wrappers* em outras linguagens além de C++ que poderão interagir com o `duckdb`. Além disso, é possível rodar o `duckdb` no próprio *browser* utilizando uma versão compilada em Web Assembly (KOHN; MORITZ, 2021).

Todas estas características aliadas à facilidade de instalação tornam o `duckdb` uma ferramenta bastante robusta para a análise de grandes bases de dados.

10.2 OBJETIVO

Este capítulo tem o objetivo de mostrar como podemos usar o `duckdb` com a linguagem `R` (R CORE TEAM, 2023). Discutiremos nas seções seguintes alguns tópicos como:

- Simplicidade

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

- Velocidade
- Recursos
- Casos de uso

10.3 APLICAÇÃO

10.3.1 Simplicidade

Para usar o `duckdb` no `R` é bastante fácil, tendo uma instalação bem simples como vemos a seguir:

```
install.packages("duckdb")
```

Caso necessite de uma versão em desenvolvimento do pacote, a instalação poderá ser feita indicando o repositório do `duckdb`:

```
install.packages('duckdb', repos=c('https://duckdb.r-
  universe.dev',
  'https://cloud.r-project.org'))
```

A versão estável utilizada neste capítulo foi a 0.7.1.1.

10.3.2 Velocidade

Nesta seção, vamos investigar a impressionante velocidade do `duckdb` na análise de dados. Para ilustrar esse desempenho, faremos uso de um conjunto de dados que registra a duração de viagens de táxi na cidade de Nova York. Esses dados foram obtidos a partir da plataforma *Kaggle* e estão disponíveis no formato `csv` através do link https://www.kaggle.com/competitions/nyc-taxi-trip-duration/data. O conjunto de dados compreende cerca de 1.5 milhões de registros de viagens com o arquivo tendo o tamanho de 191MB.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Utilizando apenas o `r-base` e o pacote `dplyr` para criar uma variável com os meses, aplicar um `group_by` e um `summarise` para calcular a média, o tempo foi de aproximadamente 22 segundos.

```
system.time({
  read.csv("../data/nyc-taxi-trip-duration/train.csv")
    |>
    dplyr::mutate(month = lubridate::month(dropoff_
      datetime)) |>
    dplyr::group_by(month) |>
    dplyr::summarise(`Média (s)` = mean(trip_duration,
      na.rm = TRUE))
})
```

O resultado é a média de duração de viagens por mês:

```
usuário   sistema decorrido
 14.323     6.096    21.711
```

Contudo, utilizando o `duckdb`, aplicando a mesma análise, o tempo de execução da leitura de dados e consulta foi um pouco maior do que 2 segundos somente:

```
system.time({
  con <- duckdb::dbConnect(duckdb::duckdb(), "../data/
    nyc-taxi.duckdb")
  duckdb::duckdb_read_csv(con,
    "nyc-taxi", "../data/nyc-taxi-trip-duration/train.
      csv")
  dplyr::tbl(con, "nyc-taxi") |>
  dplyr::mutate(month = dplyr::sql("datepart('month',
    strptime(dropoff_datetime, '%Y-%m-%d %H:%M:%S'))"))
      |>
```

```
  dplyr::group_by(month) |>
  dplyr::summarise(`Média (s)` = mean(trip_duration, na
    .rm = TRUE))
  duckdb::dbDisconnect(con, shutdown = TRUE)
})
```

O resultado é a média de duração de viagens por mês:

```
usuário   sistema decorrido
  2.024     0.145     2.331
```

Os resultados mostram que o `duckdb` é significativamente mais rápido na análise dos dados em comparação com o `R-base` para o conjunto de dados fornecido.

10.4 RECURSOS

10.4.1 Tipos de dados

No `duckdb` há mais de 20 tipos de dados suportados. A lista completa pode ser consultada neste link https://duckdb.org/docs/sql/data_types/overview. A seguir, poderá ser observado alguns dados que foram salvos em alguns tipos no `duckdb` e como os tipos foram preservados ao se fazer a consulta na tabela `dplyr::tbl(con, "examples")`. Uma observação interessante é do tipo `factor` no `R`, e como ele persiste na tabela do `duckdb` como `enum`. Contudo, se esta tabela for lida novamente no `R`, o tipo `factor` permanece.

```
con <- duckdb::dbConnect(duckdb::duckdb(), ":memory:")
dplyr::tibble( boolean = c(TRUE, TRUE, FALSE, TRUE),
  double = c(-1.2, 5.65, 0.91, 100),
  integer = c(3L, 20L, 0L, -2L),
  timestamp = c("2023-04-01 12:13", "2023-05-30 01:45",
```

```
    "2023-06-07 13:01", "2023-09-23 23:02") |>
    lubridate::ymd_hm(),
  varchar = LETTERS[5:8],
  enum = factor(c("Y", "Y", "N", "Y"), levels = c("N",
    "Y"))
) |>
duckdb::dbWriteTable(con, "examples", value = _,
  overwrite = TRUE)
dplyr::tbl(con, "examples")
duckdb::dbDisconnect(con, shutdown = TRUE)
```

10.4.2 Tipos de dados aninhados

Além da extensa diversidade de tipos de dados mencionada anteriormente, o `duckdb` também oferece suporte a tipos de dados aninhados. Esses tipos aninhados viabilizam uma estruturação mais sofisticada dos dados, proporcionando uma estrutura de armazenamento mais complexa.

Os tipos suportados são `list`, `struct` e `map`. É observado que no R existe uma perfeita compatibilidade destes tipos.

Primeiro é criada uma tabela chamada `NEST` com as variáveis:

- `int_list`, com o tipo lista (`[]`) de inteiros (`INT`)
- `varchar_list`, com o tipo lista (`[]`) de caracteres (`VARCHAR`)
- `struct`, do tipo struct (`STRUCT`) com duas variáveis `INT` e `VARCHAR`

Em seguida é feito o `INSERT` de uma observação conforme os tipos que foram definidos na criação da tabela. Foram utilizadas as funções `DBI::dbSendStatement` para pré-definir uma instrução para o banco de dados e `DBI::dbBind` para efetivamente preencher a instrução com os dados que serão inseridos na tabela.

Por último, a consulta mostra como são apresentados estes dados aninhados no R. Após o DBI::dbExecute sempre é mostrado o número de observações afetadas na tabela, como foi feita a criação de uma tabela, nenhuma observação foi alterada.

```
con <- duckdb::dbConnect(duckdb::duckdb(), ":memory:")
DBI::dbExecute(con, "CREATE TABLE NEST (int_list INT[],
    varchar_list VARCHAR[],
  struct STRUCT(i INT, j VARCHAR))"
)
```

```
[1] 0
```

```
stmt <- DBI::dbSendStatement(con, "INSERT INTO NEST
  VALUES (?, ?, ?)")
DBI::dbBind(stmt, list("[1, 2]", "['a', 'b']", "{'i':
  5, 'j': 'c'}"))
dplyr::tbl(con, "nest")
```

```
# Source: table<nest> [1 x 3]
# Database: DuckDB 0.7.1 [root@Darwin 22.6.0:R 4.3.0/:memory:]
int_list varchar_list struct$i $j
<list> <list> <int> <chr>
1 <int [2]> <chr [2]> 5 c
```

```
duckdb::dbDisconnect(con, shutdown = TRUE)
```

10.4.3 Leitura e escrita de arquivos externos

10.4.3.1 csv e parquet

Com o `duckdb` é possível fazer a leitura e escrita de arquivos em estruturas tabulares tradicionais `csv` e `parquet`.

No exemplo abaixo utiliza-se a função `duckdb::duckdb_read_csv` para a leitura de um arquivo `csv` salvo no diretório `../data/nyc-taxi.csv` e criar uma tabela de nome `nyc-taxi` na base de dados.

```
con <- duckdb::dbConnect(duckdb::duckdb(), "../data/nyc
  -taxi.duckdb")
duckdb::duckdb_read_csv(con, "nyc-taxi", "../data/nyc-
  taxi.csv")
```

Em seguida, a partir da tabela `nyc-taxi` na base de dados foi exportado localmente um arquivo no formato `parquet`, utilizando a instrução `COPY`.

```
DBI::dbExecute(con, "COPY 'nyc-taxi' TO '../data/nyc-
  taxi.parquet'
  (FORMAT PARQUET);")
duckdb::dbDisconnect(con, shutdown = TRUE)
```

Por último, temos a função `read_parquet` para leitura de arquivos locais no formato `parquet`.

```
con <- duckdb::dbConnect(duckdb::duckdb(), ":memory:")
DBI::dbGetQuery(con, "SELECT * FROM read_parquet('../
  data/nyc-taxi.parquet')
  LIMIT 2;") |>
  dplyr::as_tibble()
duckdb::dbDisconnect(con, shutdown = TRUE)
```

```
# A tibble: 2 × 11
  id         vendor_id pickup_datetime     dropoff_datetime    passenger_count
  <chr>          <int> <chr>               <chr>                         <int>
1 id2875421          2 2016-03-14 17:24:55 2016-03-14 17:32:30               1
2 id2377394          1 2016-06-12 00:43:35 2016-06-12 00:54:38               1
# i 6 more variables: pickup_longitude <dbl>, pickup_latitude <dbl>,
#   dropoff_longitude <dbl>, dropoff_latitude <dbl>, store_and_fwd_flag <chr>,
#   trip_duration <int>
```

10.4.3.2 json

O `duckdb` é um sistema de gerenciamento de banco de dados relacional que não apenas suporta a leitura de dados tabulares convencionais, mas também é capaz de processar arquivos `json`. Com isso, você pode diretamente lidar com dados não estruturados por meio do `duckdb`.

A seguir um exemplo de arquivo no formato `json`.

```
[
  {"Name" : "Mario", "Age" : 32, "Occupation" : "Plumber"},
  {"Name" : "Peach", "Age" : 21, "Occupation" : "Princess"},
  {},
  {"Name" : "Bowser", "Occupation" : "Koopa"}
]
```

Para leitura do arquivo é criada uma conexão, instalada e carregada a extensão para manipulação dos arquivos `json`.

```
con <- duckdb::dbConnect(duckdb::duckdb(), ":memory:")
DBI::dbExecute(con, "INSTALL json;")
DBI::dbExecute(con, "LOAD json;")
```

Incorporando a função `read_json_auto` na consulta através da função `DBI::dbGetQuery`, o resultado da leitura do arquivo é apresentado como um `data.frame`.

```
DBI::dbGetQuery(con, "SELECT * FROM read_json_auto('../
  data/example.json')")
duckdb::dbDisconnect(con, shutdown = TRUE)
```

É observado no resultado que cada linha do `data.frame` é um documento do arquivo `json` original.

```
    Name Age Occupation
1  Mario  32    Plumber
2  Peach  21   Princess
3   <NA>  NA       <NA>
4 Bowser  NA      Koopa
```

10.4.4 Funções

Existem inúmeras funções à disposição para uso no `duckdb`. Ao trabalhar com dados em um banco deste sistema, utilizar suas funções internas pode oferecer um desempenho muito superior na consulta.

No exemplo a seguir foi criado uma tabela chamada `functions` com duas variáveis `telefone` e `start_date`. Na consulta à esta tabela é aplicado um regex para extrair somente os números de um texto. Observa-se que a função `regexp_extract` do `duckdb` é chamada dentro da função `dplyr::sql` e dentro de um `dplyr::mutate` ou no caso aqui `dplyr::transmute`. No segundo caso temos uma função que irá contabilizar o número de semanas entre duas datas (`datediff`) e no último caso temos uma função que irá trazer o valor de `pi`.

```
df <- data.frame(telefone = c("Meu telefone é:
  21991831234"),
                 start_date = c("1984-10-19") |> as.
                   Date())
con <- duckdb::dbConnect(duckdb::duckdb(), ":memory:")
```

```
# Registra o df como uma tabela virtual (view)
duckdb::duckdb_register(con, "functions", df)

dplyr::tbl(con, "functions") |>
  dplyr::transmute(tel_extract = dplyr::sql("regexp_
    extract(telefone, '[0-9]+')"),
                   weeks = dplyr::sql("datediff('week',
                      start_date, today())"),
                   pi = dplyr::sql("pi()"))
```

```
    # Source: SQL [1 x 3]
# Database: DuckDB 0.7.1 [root@Darwin 22.6.0:R 4.3.0/:memory:]
tel_extract weeks pi
<chr> <dbl> <dbl>
1 21991831234 2030 3.14
```

```
duckdb::dbDisconnect(con, shutdown = TRUE)
```

10.5 RESULTADOS

Nesta seção iremos mostrar alguns exemplos de casos de uso com o `duckdb`.

10.5.1 Mineração de texto

Neste primeiro exemplo será mostrado como pode ser aplicada várias funções do `duckdb` para manipulação de texto.

No primeiro trecho do código é feita a leitura dos textos que irão ser manipulados e salvos no objeto `bible`.

```
bible <- readr::read_lines(
  url("https://www.o-bible.com/download/kjv.txt"),
  skip = 1
  ) |> dplyr::as_tibble()
```

A seguir é criada uma conexão com uma base de dados e uma tabela com o nome `bible`. Esta tabela criada foi uma tabela virtual, onde os dados não são armazenados fisicamente, ou seja, os dados persistem somente enquanto a conexão com o banco permanece ativa. A criação de tabelas virtuais é feita com a função `duckdb::duckdb_register`.

```
con <- duckdb::dbConnect(duckdb::duckdb(), ":memory:")
duckdb::duckdb_register(con, "bible", bible)
```

A tabela denominada `bible` pode ser visualizada abaixo:

```
dplyr::tbl(con, "bible")
```

```
# Source: table<bible> [?? x 1]
# Database: DuckDB 0.7.1 [root@Darwin 22.6.0:R 4.3.0/:memory:]
value
<chr>
1 Ge1:1 In the beginning God created the heaven and the earth.
2 Ge1:2 And the earth was without form, and void; and darkness was upon the fa~
3 Ge1:3 And God said, Let there be light: and there was light.
4 Ge1:4 And God saw the light, that it was good: and God divided the light fro~
5 Ge1:5 And God called the light Day, and the darkness he called Night. And th~
6 Ge1:6 And God said, Let there be a firmament in the midst of the waters, and~
7 Ge1:7 And God made the firmament, and divided the waters which were under th~
8 Ge1:8 And God called the firmament Heaven. And the evening and the morning w~
9 Ge1:9 And God said, Let the waters under the heaven be gathered together unt~
10 Ge1:10 And God called the dry land Earth; and the gathering together of the ~
# i more rows
```

Cada linha da tabela é um versículo da bíblia e em cada início do texto há uma **referência** em que as duas primeiras letras representam o livro, o número antes dos ':' representa o capítulo e os últimos números (após os ':') representam o versículo.

No próximo bloco de códigos são aplicadas um conjunto de funções para a manipulação dos textos:

- `regexp_extract` faz a extração do livro (salvo com o nome `book`), ou seja, somente a parte de texto da referência.
- `regexp_replace` faz uma limpeza e padronização do texto ao substituir a referência do texto (`\\w+\\d+\\:\\d+`, e.g. Ge1:1) ou (`|`) pontuações (`\\;` `\\,` `|\\.` `|\\:)|`) por vazio (''). Foi aplicada a função `trim` para tirar os espaços das extremidades do texto e por último foi aplicada a função `lcase` para por o texto em caixa baixa.
- `regexp_split_to_array` transforma o texto em uma lista de palavras, utilizando o espaço (`\\s`) entre as palavras como o ponto de corte.
- `list_filter` filtra a lista de palavras que não correspondem (`regexp_matches`) com *in*, *the* e *and*.

```
(words <- dplyr::tbl(con, "bible") |>
  dplyr::mutate(book = dplyr::sql("regexp_extract(
    regexp_extract(value,
                   '\\w+\\d+\\:\\d+'), '[A-Za-z]+')"),
                   text = dplyr::sql("lcase(trim(regexp_
                     replace(value,
                   '\\w+\\d+\\:\\d+|\\;|\\,|\\.|\\:', '',
                     'g')))"),
                   word = dplyr::sql("regexp_split_to_
                     array(text, '\\s')"),
```

```
                word_clean = dplyr::sql("list_filter(
                  word, x -> NOT
                regexp_matches(x, 'in|the|and'))")) |>
  dplyr::select(book, text, word, word_clean) |> head
    (1) |> dplyr::as_tibble())
```

```
# A tibble: 1 x 4
book text                                  word word_clean
<chr> <chr>                               <list> <list>
1 Ge in the beginning god created the heaven and the earth <chr> <chr [4]>
```

O resultado da primeira observação pode ser visto com a consulta acima. E abaixo vemos como ficaram estruturadas as variáveis `word` e `word_clean` após a consulta.

```
words$word
```

```
[[1]]
[1] "in" "the" "beginning" "god" "created" "the"
[7] "heaven" "and" "the" "earth"
```

```
words$word_clean
```

```
[[1]]
[1] "god" "created" "heaven" "earth"
```

```
duckdb::dbDisconnect(con, shutdown = TRUE)
```

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

10.5.2 Dados de COVID-19

Neste próximo exemplo será mostrado como trabalhar com dados de COVID-19.

Os dados contidos no link (`url`) abaixo provêm do repositório da John Hopkins University e abrangem informações relacionadas à pandemia de COVID-19. Esta fonte compila e disponibiliza os dados atualizados de todo o mundo.

Abaixo vemos que os dados contém as informações:

- `Province.State`
- `Country.Region`
- `Lat` e `Long`
- Diversas colunas com o padrão mês, dia e ano (`XMM.DD.AA`) tendo o valor acumulado de casos

```
url <- paste0(
  "https://raw.githubusercontent.com/CSSEGISandData/
    COVID-19/master/",
  "csse_covid_19_data/csse_covid_19_time_series/",
  "time_series_covid19_confirmed_global.csv"
  )

read.csv(url, stringsAsFactors = FALSE) |>
  dplyr::as_tibble()
```

```
# A tibble: 289 x 1,147
Province.State Country.Region Lat Long X1.22.20 X1.23.20 X1.24.20
<chr> <chr> <dbl> <dbl> <int> <int> <int>
1 "" Afghanistan 33.9 67.7 0 0 0
2 "" Albania 41.2 20.2 0 0 0
```

```
3 "" Algeria 28.0 1.66 0 0 0
4 "" Andorra 42.5 1.52 0 0 0
5 "" Angola -11.2 17.9 0 0 0
6 "" Antarctica -71.9 23.3 0 0 0
7 "" Antigua and B~ 17.1 -61.8 0 0 0
8 "" Argentina -38.4 -63.6 0 0 0
9 "" Armenia 40.1 45.0 0 0 0
10 "Australian Capital T~ Australia -35.5 149. 0 0 0
# i 279 more rows
# i 1,140 more variables: X1.25.20 <int>, X1.26.20 <int>, X1.27.20 <int>,
# X1.28.20 <int>, X1.29.20 <int>, X1.30.20 <int>, X1.31.20 <int>,
# X2.1.20 <int>, X2.2.20 <int>, X2.3.20 <int>, X2.4.20 <int>, X2.5.20 <int>,
# X2.6.20 <int>, X2.7.20 <int>, X2.8.20 <int>, X2.9.20 <int>, X2.10.20 <int>,
# X2.11.20 <int>, X2.12.20 <int>, X2.13.20 <int>, X2.14.20 <int>,
# X2.15.20 <int>, X2.16.20 <int>, X2.17.20 <int>, X2.18.20 <int>, ...
```

‘ Para a manipulação destes dados criaremos uma conexão e uma tabela virtual com o nome de `"covid19"`.

```
con <- duckdb::dbConnect(duckdb::duckdb(), ":memory:")
duckdb::duckdb_register(con, "covid19", read.csv(url,
   stringsAsFactors = FALSE))
```

Em seguida utilizamos a instrução `PIVOT_LONGER` a fim de mudar a orientação da tabela de *wide* para *long*. Sendo que a principal alteração será agrupar o conteúdo que estão distribuídos em várias colunas separadas para apenas uma coluna com os valores e uma outra coluna com as datas. O argumento `ON` com a função `COLUMNS('X')` declara quais são as colunas que deverão ser transformadas e os argumentos `NAME date` e `VALUE cumulate` declaram a construção das novas colunas que irão receber o novo dado estruturado.

- `replace` irá substituir o valor `'X'` por `''` e a função `strptime` transforma o dado do tipo `VARCHAR` para o tipo `DATE` seguindo o padrão da data `'%m.%d.%y'`.
- É construída uma variável `value` que é o valor acumulado

- o valor do dia anterior (`dplyr::lag(cumulate)`), assim criamos uma variável com o valor exato do dia.
- É aplicado um filtro `date > "2020-02-23"`.
- A função head equivale a declaração LIMIT no SQL.

```
dplyr::tbl(con, dplyr::sql("(PIVOT_LONGER covid19 ON
  COLUMNS('X')
  INTO NAME date VALUE cumulate)")) |>
  dplyr::select(country = Country.Region, date,
    cumulate) |>
  dplyr::mutate(date = dplyr::sql("strptime(replace(
    date, 'X', ''), '%m.%d.%y')"),
                value = cumulate - dplyr::lag(cumulate)
                  ) |>
  dplyr::filter(date > "2020-02-23") |> head(3)
```

```
# Source:   SQL [3 x 4]
# Database: DuckDB 0.7.2-dev2706 [root@Darwin 22.4.0:R 4.2.3/:memory:]
  country     date                cumulate value
  <chr>       <dttm>                 <int> <int>
1 Afghanistan 2020-02-24 00:00:00        5     5
2 Afghanistan 2020-02-25 00:00:00        5     0
3 Afghanistan 2020-02-26 00:00:00        5     0
```

Para utilizar a declaração `PIVOT_LONGER` foi necessário instalar a versão `0.8.0` do pacote em desenvolvimento do `duckdb`.

10.5.3 Lendo dados de um serviço de armazenamento na nuvem

Com o `duckdb` é possível ler arquivos remotos armazenados em um recurso na núvem. O exemplo que utilizaremos aqui é da tabela com as viagens de táxi de

Nova York, em um arquivo de formato `parquet` hospedado em um *Cloud Object Storage (COS)* da IBM.

Primeiro é criada uma conexão em memória e nesta conexão é instalada e carregada a extensão `httpfs`.

```
con <- duckdb::dbConnect(duckdb::duckdb(), ":memory:")
DBI::dbExecute(con, "INSTALL httpfs;")
DBI::dbExecute(con, "LOAD httpfs;")
```

Em seguida são declaradas as informações do serviço onde está hospedado o arquivo no qual se quer consultar.

```
library(glue)
s3_region <- Sys.getenv('S3_REGION')
s3_endpoint <- Sys.getenv('S3_ENDPOINT')
s3_access_key_id <- Sys.getenv('S3_ACCESS_KEY_ID')
s3_secret_access_key <- Sys.getenv('S3_SECRET_ACCESS_
  KEY')
DBI::dbExecute(con, glue("SET s3_region='{s3_region}';"
  ))
DBI::dbExecute(con, glue("SET s3_endpoint='{s3_endpoint
  }';"))
DBI::dbExecute(con, glue("SET s3_access_key_id='{s3_
  access_key_id}';"))
DBI::dbExecute(con, glue("SET s3_secret_access_key='{s3
  _secret_access_key}';"))
```

As variáveis de ambiente são especificadas da seguinte forma:

```
S3_REGION=us-south
S3_ENDPOINT=s3.us-south.cloud-object-storage.appdomain.
  cloud
```

```
S3_ACCESS_KEY_ID=<s3_access_key_id>
S3_SECRET_ACCESS_KEY=<s3_secret_access_key>
```

Elas podem ser salvas em um arquivo `.Renviron` e lidas com a função `readRenviron()`.

Finalmente a tabela poderá ser consultada diretamente de onde ela está armazenada.

```
dplyr::tbl(con, "s3://duckdb-ser/nyc-taxi.parquet")
```

```
# Source:   table<s3://duckdb-ser/nyc-taxi.parquet> [?? x 11]
# Database: DuckDB 0.7.1 [root@Darwin 22.4.0:R 4.2.3/:memory:]
   id         vendor_id pickup_datetime     dropoff_datetime    passenger_count
   <chr>          <int> <chr>               <chr>                         <int>
 1 id2875421          2 2016-03-14 17:24:55 2016-03-14 17:32:30               1
 2 id2377394          1 2016-06-12 00:43:35 2016-06-12 00:54:38               1
 3 id3858529          2 2016-01-19 11:35:24 2016-01-19 12:10:48               1
 4 id3504673          2 2016-04-06 19:32:31 2016-04-06 19:39:40               1
 5 id2181028          2 2016-03-26 13:30:55 2016-03-26 13:38:10               1
 6 id0801584          2 2016-01-30 22:01:40 2016-01-30 22:09:03               6
 7 id1813257          1 2016-06-17 22:34:59 2016-06-17 22:40:40               4
 8 id1324603          2 2016-05-21 07:54:58 2016-05-21 08:20:49               1
 9 id1301050          1 2016-05-27 23:12:23 2016-05-27 23:16:38               1
10 id0012891          2 2016-03-10 21:45:01 2016-03-10 22:05:26               1
# i more rows
# i 6 more variables: pickup_longitude <dbl>, pickup_latitude <dbl>,
#   dropoff_longitude <dbl>, dropoff_latitude <dbl>, store_and_fwd_flag <chr>,
#   trip_duration <int>
```

10.5.4 Análise de Dados Espaciais

Com o `duckdb`, também é possível analisar dados espaciais através da extensão `spatial` (GABRIELSSON, 2023).

Primeiro, é criada uma conexão em memória e nesta conexão é instalada e carregada a extensão `spatial`.

```
con <- duckdb::dbConnect(duckdb::duckdb(), ":memory:")
DBI::dbExecute(con, "INSTALL httpfs;")
DBI::dbExecute(con, "LOAD httpfs;")
```

Os dados são lidos novamente do `COS`, como mostrado na seção anterior, e em seguida, na consulta, são aplicadas um conjunto de funções para análise espacial.

```
nyc_taxi_spatial <- dplyr::tbl(con, "s3://duckdb-ser/
  nyc-taxi.parquet") |>
  dplyr::mutate(
    pickup_point = dplyr::sql("ST_Transform(ST_Point(
      pickup_latitude,
      pickup_longitude), 'EPSG:4326', 'ESRI:102718')"),
    dropoff_point = dplyr::sql("ST_Transform(ST_Point(
      dropoff_latitude,
      dropoff_longitude), 'EPSG:4326', 'ESRI:102718')")
        ,
    aerial_distance = dplyr::sql("ST_Distance(pickup_
      point, dropoff_point)/3280.84")
  ) |> dplyr::as_tibble()
```

- `ST_Point` recebe como input as geolocalizações
- `ST_Transform` faz a conversão das coordenadas geográficas de latitude e longitude, que utilizam o sistema "WGS84"(`EPSG:4326`), em coordenadas de projeção específicas para a região de Nova York, denominadas "NAD83 / New York Long Island ftUS"(`ESRI:102718`). Essa transformação é útil ao lidar com dados geoespaciais relacionados à área de Nova York e é necessária para assegurar uma representação precisa das coordenadas nessa região, minimizando a distorção.

- `ST_Distance` calcula a distância em linha reta entre o ponto de embarque e o ponto de desembarque. A razão entre a distância e o valor 3280.84 é um ajuste para apresentar o valor em Km.

Abaixo vemos apenas um ajuste na apresentação dos resultados e a tabela com as informações extraídas e calculada.

```
nyc_taxi_spatial |>
  dplyr::select(pickup_longitude, pickup_latitude,
                dropoff_longitude, dropoff_latitude,
                aerial_distance) |>
  dplyr::slice(1) |>
  tidyr::pivot_longer(tidyr::everything()) |>
  dplyr::mutate(value = tibble::num(value, digits = 5))
```

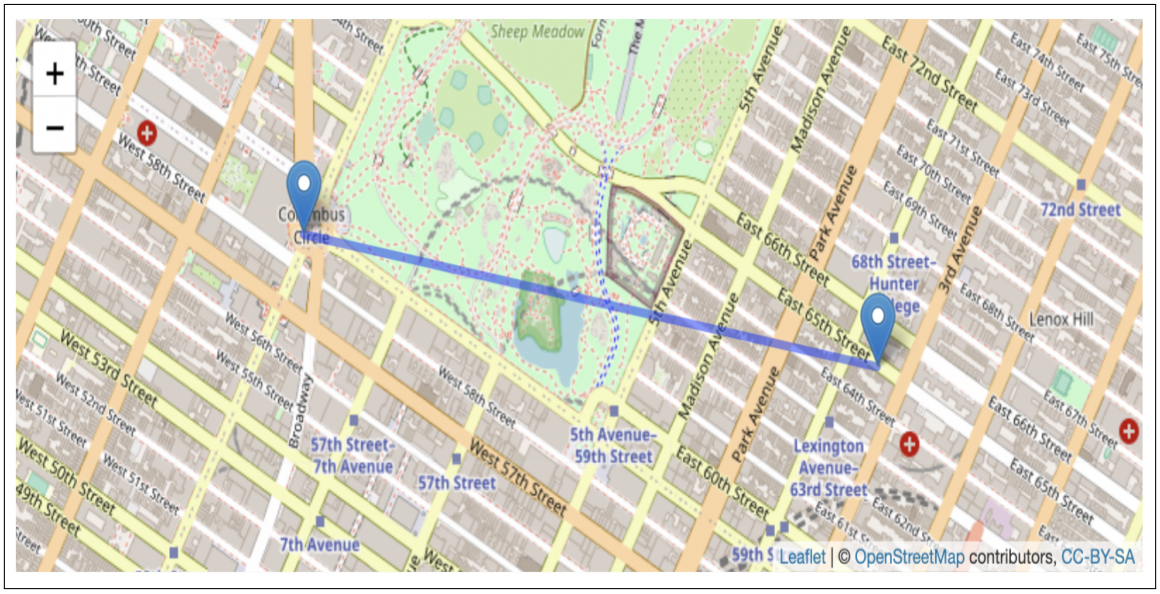
```
# A tibble: 5 × 2
  name                 value
  <chr>            <num:.5!>
1 pickup_longitude  -73.98215
2 pickup_latitude    40.76794
3 dropoff_longitude -73.96463
4 dropoff_latitude   40.76560
5 aerial_distance     1.50216
```

Na 10.1 vemos o mapa com os pontos de embarque e desembarque, ligados por uma linha reta representando a distância que foi calculada.

10.6 UM BANCO EMBARCADO EM UMA API

O `duckdb` é um banco de dados embarcado, o que significa que ele é projetado para ser incorporado diretamente em aplicativos ou sistemas, em vez de ser executado como um serviço separado em um servidor, por exemplo.

Figura 10.1: Pontos de embarque e desembarque.



Fonte: O autor.

Nesse exemplo será mostrado como incorporar o `duckdb` em uma API construída com `plumber` (SCHLOERKE; ALLEN, 2022).

A API é estruturada em uma função no arquivo `api.R`. Esta função irá consultar uma tabela no formato `parquet` que está hospedada no COS. A função tem um argumento `input` e receberá um `id` para filtrar as informações de uma viagem de táxi.

```
#* @apiTitle Mostrar informações segundo ID
#* @param input
#* @get /info
function(input) {
  # Ler variáveis de ambiente
  readRenviron(".Renviron")
  s3_region <- Sys.getenv("S3_REGION")
  s3_endpoint <- Sys.getenv("S3_ENDPOINT")
  s3_access_key_id <- Sys.getenv("S3_ACCESS_KEY_ID")
  s3_secret_access_key <- Sys.getenv("S3_SECRET_ACCESS_
```

```
    KEY")

# Criar conexão com o banco
con <- duckdb::dbConnect(duckdb::duckdb(), ":memory:"
    )
invisible(DBI::dbExecute(con, "INSTALL httpfs;"))
invisible(DBI::dbExecute(con, "LOAD httpfs;"))
invisible(DBI::dbExecute(
  con,
  glue::glue("SET s3_region='{s3_region}';")
))
invisible(DBI::dbExecute(
  con,
  glue::glue("SET s3_endpoint='{s3_endpoint}';")
))
invisible(DBI::dbExecute(
  con,
  glue::glue("SET s3_access_key_id='{s3_access_key_id
    }';")
))
invisible(DBI::dbExecute(
  con,
  glue::glue("SET s3_secret_access_key='{s3_secret_
    access_key}';")
))
# Consulta
resposta <- dplyr::tbl(con, "s3://duckdb-ser/nyc-taxi
    .parquet") |>
  dplyr::filter(id == input) |>
```

```
    dplyr::as_tibble() |>
    as.data.frame()
  duckdb::dbDisconnect(con, shutdown = TRUE)
  # Resultado
  return(resposta)
}
```

Para iniciar a API deve ser executado o comando abaixo, onde é informado o nome do script e a porta em que a API estará disponível:

```
plumber::plumber("api.R") |>
  plumber::pr_run(port=8010)
```

Para fazer a requisição na API podemos usar o pacote `httr`. No exemplo a API roda localmente `http://127.0.0.1` na porta `8010` e foi requisitada as informações da viagem com o `id` de `id2875421`.

```
httr::GET("http://127.0.0.1:8010/info?input=id2875421")
    |>
  httr::content() |>
  jsonlite::toJSON(auto_unbox = TRUE, pretty = TRUE)
```

No resultado observa-se as informações da viagem segundo o `id` informado:

```
[
  {
    "id": "id2875421",
    "vendor_id": 2,
    "pickup_datetime": "2016-03-14 17:24:55",
    "dropoff_datetime": "2016-03-14 17:32:30",
    "passenger_count": 1,
    "pickup_longitude": -73.9822,
```

```
    "pickup_latitude": 40.7679,
    "dropoff_longitude": -73.9646,
    "dropoff_latitude": 40.7656,
    "store_and_fwd_flag": "N",
    "trip_duration": 455
  }
]
```


10.7 CONCLUSÃO

O `duckdb` certamente se destaca como uma ferramenta robusta e versátil para análise de grandes conjuntos de dados. Suas características únicas, facilidade de instalação e extensibilidade através de *plugins* o tornam uma escolha atraente para muitos cenários de análise de dados. Além disso, seu suporte a vários tipos de dados e estruturas complexas aninhadas amplia ainda mais suas capacidades.

A capacidade de processamento em grande escala do `duckdb` o torna adequado para lidar com volumes substanciais de dados, o que é fundamental em muitos projetos de análise de dados. A capacidade de consultar dados usando SQL facilita a integração com outras ferramentas e sistemas que utilizam essa linguagem.

No geral, o `duckdb` é uma solução eficaz e versátil para análise de dados, e sua combinação de desempenho excepcional e ampla gama de recursos faz com que seja uma escolha valiosa para profissionais que trabalham com grandes conjuntos de dados e análise de dados complexa.

10.8 REFERÊNCIAS

DUCKDB DEVELOPMENT TEAM. **DuckDB**. [S.l.: s.n.], 2023. Avilable at: https://duckdb.org/. Accessed on September 17, 2023.

GABRIELSSON, Max. **PostGEESE? Introducing The DuckDB Spatial Extension**. [S.l.: s.n.], abr. 2023. Avilable at: https://duckdb.org/2023/04/28/spatial.html. Accessed on September 17, 2023.

KOHN, André; MORITZ, Dominik. **DuckDB-Wasm: Efficient Analytical SQL in the Browser**. [S.l.: s.n.], out. 2021. Avilable at: https://duckdb.org/2021/10/29/duckdb-wasm.html. Accessed on September 17, 2023.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: [s.n.], 2023. Avilable at: https://www.R-project.org/.

RAASVELDT, Mark. **Lightweight Data Compression in DuckDB**. [S.l.: s.n.], out. 2022. Avilable at: https://duckdb.org/2022/10/28/lightweight-compression.html. Accessed on September 17, 2023.

RAASVELDT, Mark; MÜHLEISEN, Hannes. Duckdb: an embeddable analytical database. In: PROCEEDINGS of the 2019 International Conference on Management of Data. [S.l.: s.n.], 2019. p. 1981–1984.

SCHLOERKE, Barret; ALLEN, Jeff. **plumber: An API Generator for R**. [S.l.: s.n.], 2022. Avilable at: https://CRAN.R-project.org/package=plumber. R package version 1.2.1.

Capítulo 11

EDITANDO OS GRÁFICOS DO PACOTE LIKERT

Autor: Ariel Levy, Marcus Antonio Cardoso Ramalho
Universidade Federal Fluminense, UFF-RJ
e-mail: alevy@id.uff.br e marcusantonio@id.uff.br

A maioria dos pesquisadores ou analistas se deparam com dados em idiomas diferentes daqueles em que produzem seus relatórios. A situação não é diferente quando os dados se apresentam no formato de uma escala Likert. Embora existam muitos instrumentos que permitam a análise deste tipo de pesquisa, é no R, através do pacote *likert*, que são facilmente elaborados e incorporados a um relatório. O objetivo é guiar a estruturação de uma pesquisa com escala Likert, desde a formulação do questionário até a análise com o pacote *likert* no R. Utilizanos o *Quarto* (ALLAIRE et al., 2022), na construção do documento e os pacotes *likert* (BRYER; SPEERSCHNEIDER, 2016), *gt* (IANNONE et al., 2023), *tidyverse* (WICKHAM; AVERICK et al., 2019) e *scales* (WICKHAM; SEIDEL, 2022) na manipulação dos dados e elaboração dos gráficos e tabelas. A base de dados utilizada foi a do PISA 2009 (*Programa internacional de avaliação de estudantes (PISA) : resultados nacionais - PISA 2009* 2010), disponível no pacote *likert*. O pacote *likert* data de 2016 no CRAN e apesar de suas atualizações alguns recursos não estão em pleno funcionamento. Detalhamos as etapas e construção de uma

pesquisa Likert, incluindo um roteiro que facilite a análise dos resultados confrontando com a teoria. Na parte dos códigos procuramos empregar os conceitos atualizados utilizando o tidyverse e sempre que possível nos remetendo à base do pacote *likert*. Para as próximas etapas procuraremos completar a análise focando em pacotes que permitam o tratamento das variáveis categóricas.

Palavras-Chave: Likert; Variáveis categóricas; *ggplot2*; *dplyr*; Funções.

11.1 INTRODUÇÃO

Quem respondeu ou elaborou uma pesquisa provavelmente já se deparou com uma escala Likert. Embora, a utilização de um questionário no formato Likert encontra-se amplamente explorado na literatura, nosso objetivo será elucidar algumas questões e tratamentos desde a formulação do questionário até a análise descritiva dos dados com a utilização do pacote *likert* disponível para linguagem R (R CORE TEAM, 2023).

Motivadas por curiosidade ou para tomar decisões, as pessoas com frequência desejam ou precisam conhecer o que os outros pensam e como se sentem em relação aos mais variados temas, produtos e serviços. Então, utilizam-se de experimentos para coletar essas informações. Um dos formatos mais usuais para obtenção das respostas é uma escala Likert que apresenta normalmente cinco pontos, podendo variar de 3 a 9. Por exemplo, variando de "discordo totalmente" a "concordo totalmente". Outras variações destes descritores textuais são possíveis, conforme o que se deseja medir, como a frequência, a importância ou probabilidade da ocorrência de um evento.

Esta escala tem utilização nas ciências sociais onde é predominante. A escala psicométrica foi criada por Remus Likert em 1932, com a finalidade de melhor representar as opiniões quando comparada aos itens dicotômicos, sim e não, utilizados até então. Assim, através das respostas os prospectos podem expressar a discordância ou concordância, ou mesmo sua neutralidade, ou não aplicabilidade do item em questão.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

Com isso explicado, é importante lembrar que os dados do tipo Likert são categóricos ordinais. Ainda que, possam ser representados por números, só se pode afirmar que uma pontuação é maior do que outra, e nada sobre a distância entre os pontos. Embora, haja aqueles que defendem a conversão dos dados ordinais em dados contínuos, o que não é recomendado, pois essa conversão poderá acarretar conclusões errôneas. Por exemplo, se uma pessoa responde 5 em uma escala de 1 a 7, não significa que ela concorda mais do que alguém que respondeu 4. Assim, a média não será uma medida de posição adequada do grupo de respondentes em relação a esta afirmativa ou questão. Utilizar a mediana ou a moda será mais recomendado (SULLIVAN; ARTINO, 2013).

11.1.1 Objetivo

Apesar de populares e amplamente utilizadas, as escalas Likert são frequentemente mal interpretadas. Neste capítulo o objetivo é estruturar uma pesquisa que utilize a escala Likert desde sua formulação até a análise descritiva com o pacote Likert da linguagem R. Para isso, será apresentado o processo de elaboração de um questionário, a aplicação e a análise descritiva dos dados com o pacote *likert*.

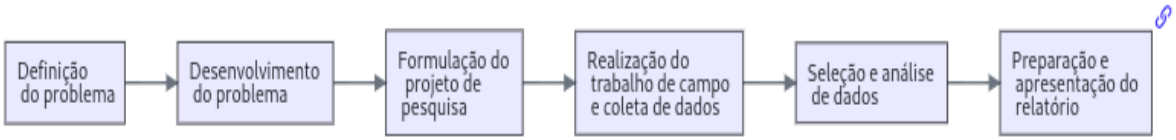
11.1.2 Aplicação

Ao iniciar sua pesquisa, busque a literatura disponível sobre o tema que você pretende pesquisar, procure identificar se há revisões sistemáticas, quais os pesquisadores mais citados, principais revistas científicas, etc. Desse modo, abreviará o esforço de produzir seu referencial teórico. Igualmente importante será verificar a existência de dados secundários, ou seja, se já existe alguma pesquisa que corrobora com dados, ou questionários validados no seu tema. Quais as lacunas ainda precisam ser preenchidas e quais são as contribuições que você pretende com sua pesquisa, como sua pesquisa irá se diferenciar das demais e como as complementará ou aprofundará.

Normalmente, o processo de uma pesquisa envolve as seguintes etapas: de-

finição do problema, o desenvolvimento do problema, a formulação do objeto de pesquisa, a realização do trabalho de campo e coleta dos dados, a seleção e análise dos dados e por fim a preparação do relatório.

Figura 11.1: Fluxo do trabalho.



Fonte: Adaptado de (MALHOTRA, 2006).

Na sequência da revisão da literatura o problema estará mais definido, tanto quanto às questões a serem respondidas, bem como, quanto às limitações às quais o pesquisador estará sujeito.

Antes de partir para a elaboração do questionário é necessário ter conhecimento do contexto e das experiências similares relatadas na literatura. Procurando estabelecer vínculos ao desenvolver suas hipóteses.

A elaboração do questionário é uma etapa importante, pois é a partir dele que você irá coletar os dados para sua pesquisa. Assim, é fundamental que você tenha clareza sobre o que pretende pesquisar e quais as questões que você pretende responder com sua pesquisa.

Ao formular uma pergunta para seu questionário você deverá alinhar: o item, afirmação que se está solicitando ao prospecto resposta; a hipótese que se estará testando e sua justificativa, conforme a literatura estudada. Respeite este formato para cada item do seu questionário. Isto facilitará a análise dos dados e a interpretação dos resultados e evitará um número excessivo de itens.

Estruture os itens de forma organizada e coerente, de modo que o questionário tenha uma sequência lógica e que seja de fácil compreensão para o respondente. Evite perguntas que possam ser interpretadas de forma ambígua. Evite também perguntas que possam ser interpretadas erroneamente por diferentes grupos ou

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

origens. A clareza e organização irão agregar valor a sua pesquisa.

O pacote likert (BRYER; SPEERSCHNEIDER, 2016) da linguagem R (R CORE TEAM, 2023) apresenta um conjunto de funções capaz de facilitar uma análise a partir de visualizações das respostas dos itens de forma isolada ou em conjunto. Este pacote traz um banco de dados exemplo de respostas da pesquisa de 2009 do PISA, Programa Internacional de Avaliação de Estudantes, que é um estudo internacional comparativo coordenado pela OCDE, Organização para Cooperação e Desenvolvimento Econômico, realizado a cada três anos, que tem como objetivo avaliar o desempenho de estudantes de 15 anos em leitura, matemática e ciências. O PISA 2009 (*Programa internacional de avaliação de estudantes (PISA) : resultados nacionais - PISA 2009* 2010) foi aplicado em 74 países e economias, incluindo o Brasil, onde a coordenação é do INEP, Instituto Nacional de Estudos Anísio Teixeira.

Então, ao utilizar este estudo apresentado como exemplo de dados secundários enfrentaremos alguns desafios para adaptar o questionário a nossa pesquisa e posterior publicação no nosso idioma. Este é um pequeno obstáculo que você poderá enfrentar ao utilizar dados secundários em uma língua estrangeira ou mesmo ao buscar publicar sua pesquisa com dados em português em um periódico internacional.

Um editor de periódico certamente irá solicitar que o questionário, seus gráficos e figuras sejam apresentados no idioma do mesmo, o que não é uma tarefa simples, pois, o questionário original foi elaborado em outro idioma e para que seja traduzido será necessário um bom conhecimento do idioma de destino com o uso de expressões e termos adequados. Além disso, é importante que o questionário traduzido seja validado, o que não será abordado neste capítulo.

A seguir, adequaremos a partir do questionário original e das soluções obtidas através do pacote *likert* para o português.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

11.2 RESULTADOS E DISCUSSÃO

Apresentaremos a seguir os resultados e discussão a partir do questionário original, do PISA 2009 e das soluções obtidas através pacote *likert* para o português.

Antes de adotar qualquer solução é importante que você tenha os pacotes instalados e carregados no R. Para isso, utilize o comando *install.packages("likert")* e *library(likert)*, respectivamente, ou utilize as instruções mostradas no código abaixo.

```
if (!require(pacman)){install.packages("pacman")}
pacman::p_load(likert, gt, tidyverse, scales)
```

A primeira instrução verifica se o pacote *pacman* encontra-se instalado, caso não esteja, o pacote será instalado e carregado. A segunda instrução fará o mesmo com os pacotes *likert* (BRYER; SPEERSCHNEIDER, 2016), (**gt**) (IANNONE et al., 2023), (**tidyverse**) (WICKHAM; AVERICK et al., 2019) e (**scales**) (WICKHAM; SEIDEL, 2022), que utilizaremos oportunamente.

Como o banco de dados do PISA que se encontra disponível no pacote (**likert**) é extenso, com 81 colunas que representam as afirmações codificadas, itens, e 66690 respostas dos estudantes, linhas. Seus metadados estão disponíveis no pacote *likert* e podem ser acessados através do comando: *?pisaitems*.

Analisaremos um dos itens para verificar sua distribuição e as opções da variável conforme mostrado.

```
data(pisaitems)
group_by(pisaitems, ST24Q01) %>% summarise(n = n())
```

```
# A tibble: 5 x 2
  ST24Q01                n
  <fct>              <int>
```

```
1 Strongly disagree 14947
2 Disagree          23515
3 Agree             20000
4 Strongly agree     7029
5 <NA>               1199
```

Podemos observar que a variável *ST24Q01*, correspondente a um dos itens, apresenta 4 opções de respostas, que são: "Discordo totalmente"(14947), "Discordo"(23515), "Concordo"(20000), "Concordo totalmente"(7029), na forma original, e o R adicionou o "NA"(1199). Estes resultados representam a contagem de respostas para cada uma das opções. E a resposta "NA"corresponde aos dados faltantes, que refere-se aos indivíduos que deixaram de responder a questão.

Escolhemos apenas algumas variáveis para exemplificar a utilização do pacote seguindo o mesmo procedimento adotado por (KOMPERDA, 2017) de quem nos diferenciaremos ao utilizarmos *tidyverse* (WICKHAM; AVERICK et al., 2019). Com a utilização da função *select()* do pacote *dplyr*, iremos selecionar as variáveis que serão utilizadas em nosso exemplo. Com estas variáveis selecionadas, criaremos um banco de dados que denominamos mini_pisa.

```
mini_pisa <- select(pisaitems, c(CNT, ST24Q01:ST24Q06))
glimpse(mini_pisa)
```

```
Rows: 66,690
Columns: 7
$ CNT     <fct> Canada, Canada, Canada, Canada, Canada,
   Canada, Canada, Canada~
$ ST24Q01 <fct> Disagree, Agree, Strongly agree,
  Disagree, Strongly disagree, ~
$ ST24Q02 <fct> Strongly agree, Strongly disagree,
  Strongly disagree, Disagree~
```

```
$ ST24Q03 <fct> Strongly agree, Strongly disagree,
  Strongly disagree, Agree, S~
$ ST24Q04 <fct> Strongly disagree, Strongly agree,
  Agree, Strongly disagree, D~
$ ST24Q05 <fct> Strongly agree, Strongly disagree,
  Strongly disagree, Disagree~
$ ST24Q06 <fct> Strongly disagree, Agree, Strongly
  agree, Disagree, Disagree, ~
```

A função *dplyr::glimpse()*, notação que informa que a função glimpse pertence ao pacote *dplyr*, apresenta um resumo do banco de dados, com o nome das variáveis, o tipo de variável e as primeiras observações. A partir deste resumo podemos observar que as variáveis são categóricas, que no R são representadas pelo tipo factor, com quatro níveis de respostas, todas com características iguais às examinadas anteriormente.

Ficaria muito difícil analisar as respostas com os itens representados pelas variáveis codificadas. Procederemos a troca dos rótulos dos itens que serão utilizados por padrão nos gráficos, com a função *dplyr::rename()*. E neste caso já poderíamos traduzi-los para o português.

- *ST24Q01: I read only if I have to.* / Eu leio apenas se eu tiver que.
- *ST24Q02: Reading is one of my favorite hobbies.* / Ler é um dos meus hobbies favoritos.
- *ST24Q03: I like talking about books with other people.* / Eu gosto de conversar sobre livros com outras pessoas.
- *ST24Q04: I find it hard to finish books.* / Eu acho difícil terminar livros.
- *ST24Q05: I feel happy if I receive a book as a present.* / Eu fico feliz se eu receber um livro de presente.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

- *ST24Q06: For me, reading is a waste of time.* / Para mim, ler é uma perda de tempo.

```
mini_pisa_t1 <- rename(mini_pisa, "País" = CNT,
    "Eu leio apenas se eu tiver que" = ST24Q01,
    "Ler é um dos meus hobbies favoritos." = ST24Q02,
    " Eu gosto de conversar sobre livros com outras
      pessoas." = ST24Q03,
    "Eu acho difícil terminar livros." = ST24Q04,
    "Eu fico feliz se eu receber um livro de presente."
       = ST24Q05,
    "Para mim, ler é uma perda de tempo." = ST24Q06)
glimpse(mini_pisa_t1)
```

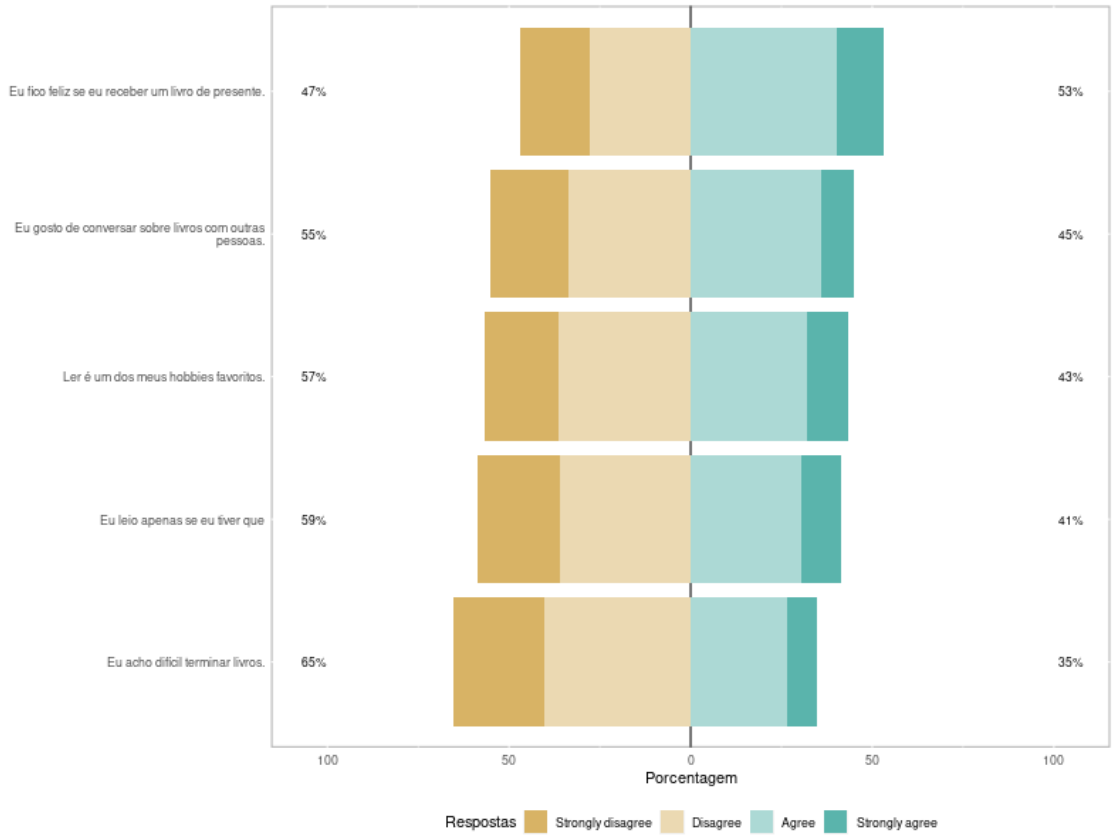
```
Rows: 66,690
Columns: 7
$ País

  <fct> Canada, Cana~
$ `Eu leio apenas se eu tiver que`
                            <fct> Disagree, Ag~
$ `Ler é um dos meus hobbies favoritos.`
                      <fct> Strongly agr~
$ ` Eu gosto de conversar sobre livros com outras
  pessoas.` <fct> Strongly agr~
$ `Eu acho difícil terminar livros.`
                           <fct> Strongly dis~
$ `Eu fico feliz se eu receber um livro de presente.`
        <fct> Strongly agr~
```

```
$ `Para mim, ler é uma perda de tempo.`
                        <fct> Strongly dis~
```

Iniciaremos selecionando as variáveis da base cujos rótulos estão traduzidos e atribuímos ao objeto criado *likert_out* utlizando a função *dplyr::select()*. A seguir, iremos utilizar a função *likert::likert()*, cujo primeiro argumento é o banco de dados que será utilizado esta deverá ser um data frame, a função não aceita um *tibble*. O objeto de saída do pacote likert será atribuído ao objeto *likert_out*.

```
#| results: hide
likert_out <- select(mini_pisa_t1, 2:6)
likert_out <- likert(as.data.frame(likert_out))
plot(likert_out) + # os eixos são trocados no likert_
  out
    labs(y = "Porcentagem") +
    guides(fill = guide_legend("Respostas"))
```

Figura 11.2: Gráfico após a tradução.



Fonte: Os autores.

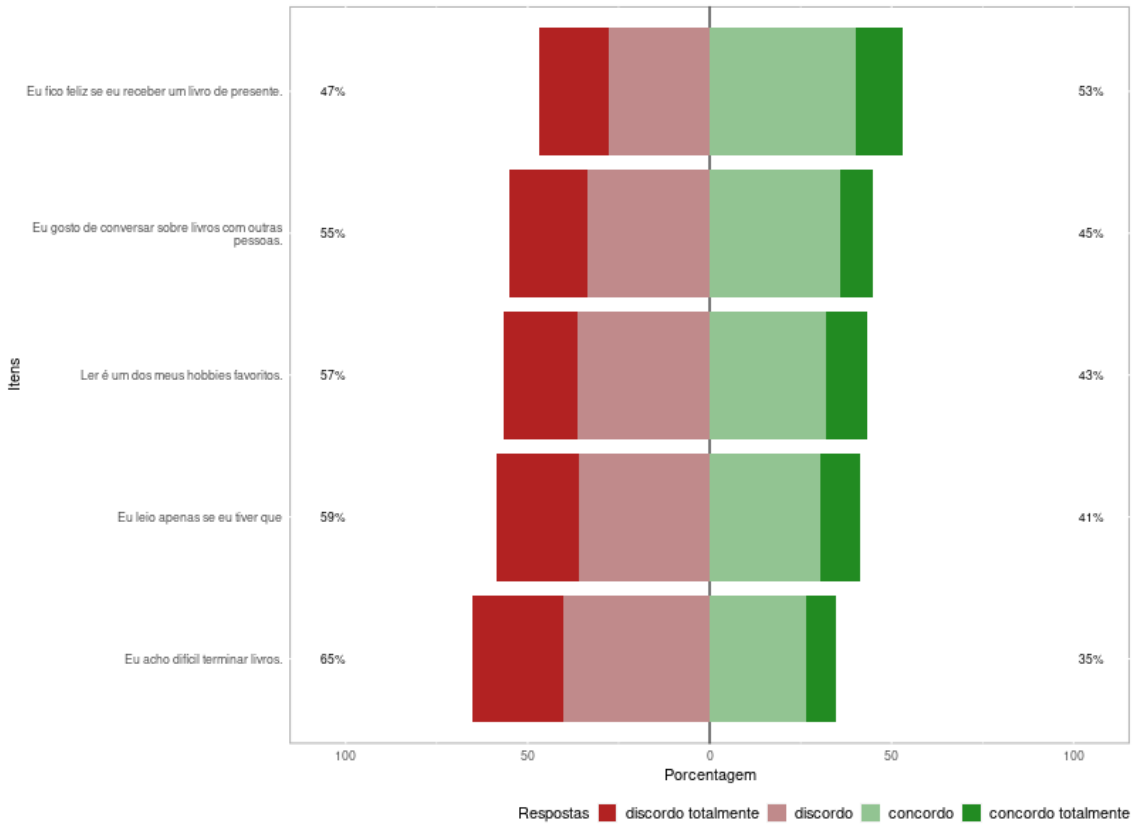
O pacote apresenta um gráfico de desenho moderno e de fácil interpretação onde as perguntas partem da neutralidade, centro das quatro opções embora não ofertada ao respondente, e mostram suas posições. Veja que ao final da instrução plot utilizamos o *"+"* característico da adição de *layers* com o pacote *ggplot2* o que nos remete que a base dos gráficos no pacote likert utiliza ele como base. Entretanto, um olhar mais cuidadoso observará que tanto o rótulo dos eixo x como a legenda e suas opções de respostas aparecem em inglês.

Além dessa correção iremos alterar as cores do gráfico, objetivo que pode ser realizado de diversas formas, através do uso de paletas, pelo nome das cores ou pelos códigos. Também reforçamos que a legenda apareça numa ordem desejada. Este conjunto de alterações será alcançado a função *ggplot2::scale_fill_manual()* cuja sintaxe será simplificada pela utilização de dois objetos criados anteriormente

color e *legend_ order*. Os eixos terão seus rótulos corrigidos com a utilização do comando labs. Exploraremos na sequência um conjunto de opções presentes na função *theme()*, onde poderemos alterar: o tamanho da fonte , a posição da legenda, margens entre muitos outros parâmetros.

```
color <- c("firebrick", "#C08A8B", "#92C492", "
  forestgreen")
legend_order <- c("Strongly disagree",
                  "Disagree",
                  "Agree",
                  "Strongly agree")
plot(likert_out, centered = TRUE,) +
    scale_fill_manual(name = "Respostas",
        values = color, breaks = legend_order,
        labels = c("discordo totalmente",
                   "discordo",
                   "concordo",
                   "concordo totalmente")) +
    labs(y = "Porcentagem", x = "Itens") +
    theme(text = element_text(size = 11),
        legend.position = "bottom",
        legend.justification = "right",
        legend.direction = "horizontal",
        legend.margin = margin(t = 3),
        axis.text.y = element_text(angle = 0),
        legend.key.size = unit(0.5, "cm"),
        legend.text = element_text(size = 11),
        plot.margin = margin(t = 15, r = 5, b = 10, l =
            5) )
```

Figura 11.3: Alteração de cores e legendas.



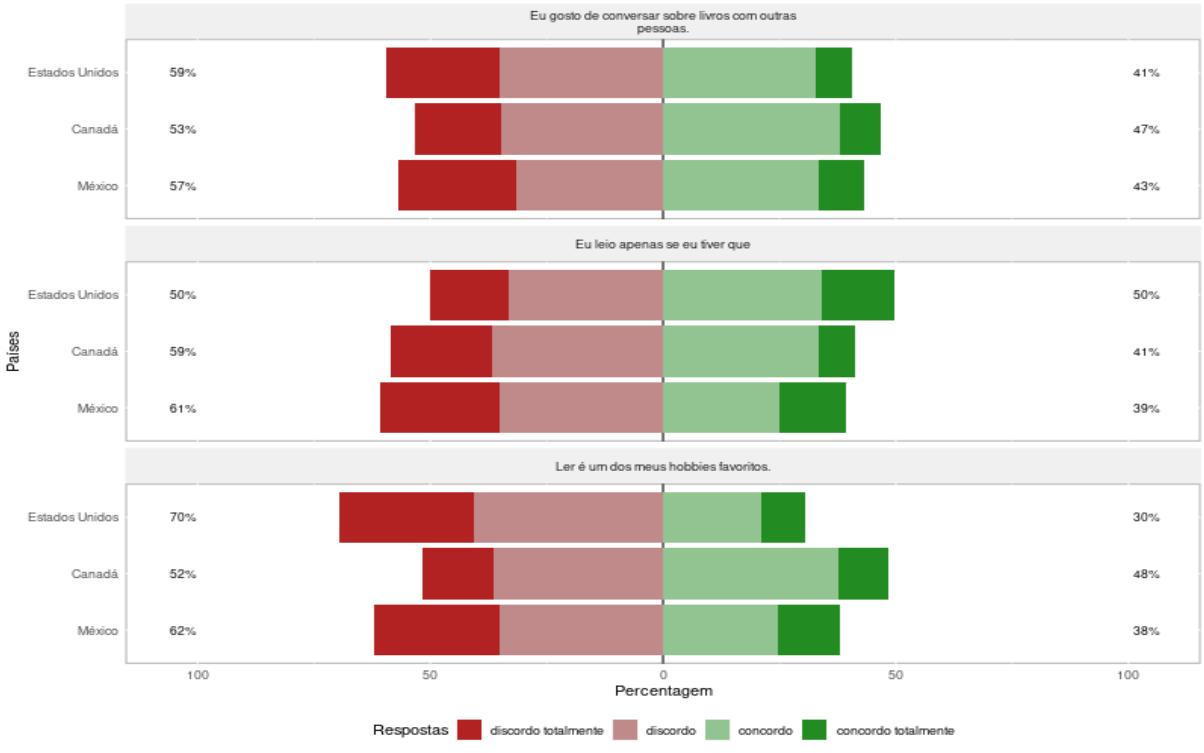
Fonte: Os autores.

O pacote *likert* apresenta recursos para segmentação da análise, bastando atribuir a variável categórica na função *likert::likert()* ao parâmetro *grouping*. Na função *likert::plot()* com o parâmetro *group.order* podemos ordenar a apresentação no gráfico. E desta feita optamos pela função *ggplot2::scale_x_discrete()* com a qual trocaremos os rótulos dos países traduzindo-os no exemplo para o português.

```
likert_out_group <- likert(as.data.frame(mini_pisa_t1[,
    2:4]),
     grouping = mini_pisa_t1$País)
plot(likert_out_group,
     group.order =c("Mexico",
```

```
                    "Canada",
                    "United States")) +
  scale_x_discrete(labels = c("México", "Canadá",
  "Estados Unidos")) +
  scale_fill_manual( name = "Respostas",
      values = color, breaks = legend_order,
      labels = c("discordo totalmente",
      "discordo",
      "concordo",
      "concordo totalmente" )) +
  labs(y = "Percentagem", x = "Países")
```

Figura 11.4: Tradução do nome dos países.



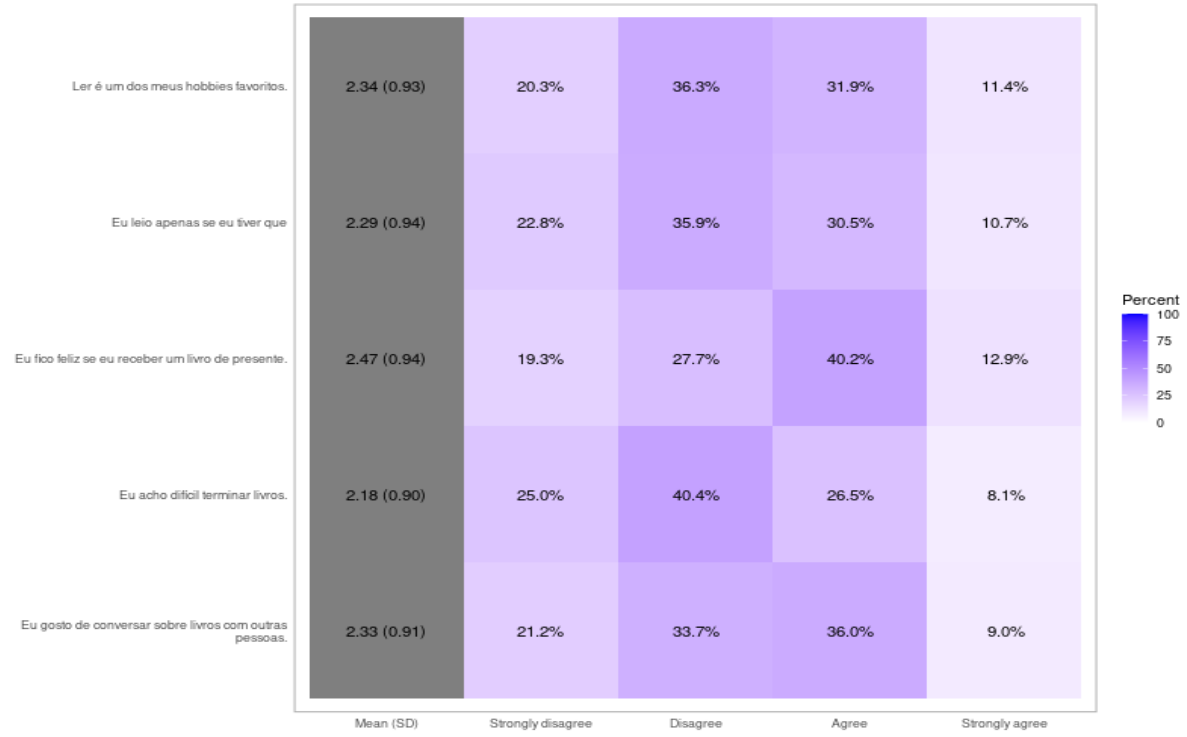
Fonte: Os autores.

Na sequência exploraremos outro recurso do pacote *likert*: um gráfico de

calor.

```
likert_out <- select(mini_pisa_t1, 2:6)
likert_out <- likert(as.data.frame(likert_out))
(heat <- plot(likert_out, type = "heat"))
```

Figura 11.5: Gráfico de calor com média.



Fonte: Os autores.

Aqui nos deparamos com uma coluna representando a média, contudo, conforme havíamos pontuado no início do capítulo a escala Likert se utiliza de variáveis categóricas e acreditamos que a média não seja uma medida de posição adequada. A retirada da coluna mean do gráfico no pacote *Likert* não é algo trivial e preferimos refazer o gráfico de calor utilizando o pacote *ggplot2*.

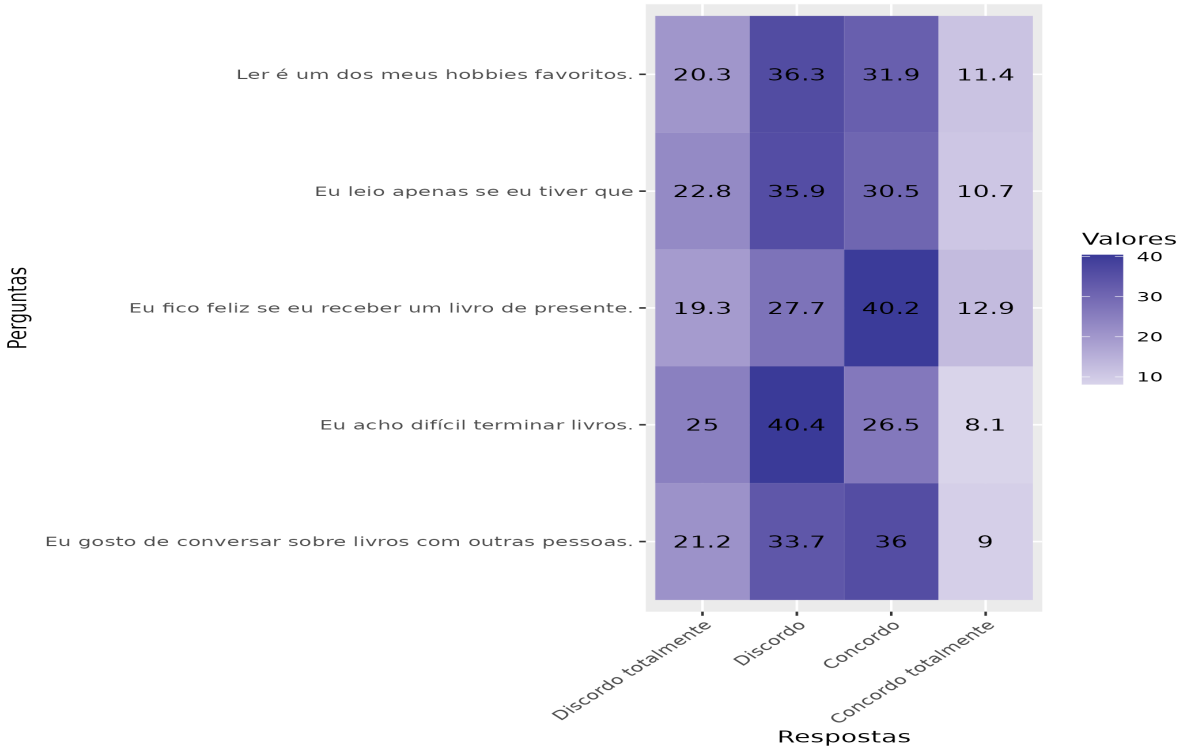
```
dados_calor <- likert_out$results %>%
    pivot_longer(cols = -Item) %>%
```

```
    filter(name != "Mean") %>%
    mutate(name = factor(name, levels = rev(unique(name
      )))) %>%
    mutate(name = factor(name,
    levels = c("Strongly disagree","Disagree","Agree",
      "Strongly agree" )))
```

O objetivo do código apresentado é a retirada da coluna *mean* dos dados e ajustar a ordem das respostas. Assim, na primeira instrução criamos um objeto que denominamos *dados_calor*. Na sequência transformamos *likert_out$results* de um dataframe largo para longo utilizando a função *pivot_longer()*. A manipulação que se segue filtra linhas indesejadas cujos o valor da coluna *name* são iguais a *Mean*. Finalizamos ao reordenar os níveis da variável categórica *name* presentes em *results* na ordem desejada.

```
dados_calor %>% ggplot(aes(x = name, y = Item, fill =
  value)) + geom_tile() +
    scale_fill_gradient2(low = muted("red"), mid = "
      white",
        high = muted("blue"), midpoint = 0) +
    geom_text(aes(label = round(value, digits = 1))) +
    theme(axis.text.x = element_text(angle = 45, hjust
      = 1)) +
    scale_x_discrete(labels = c( "Discordo totalmente",
       "Discordo",
        "Concordo", "Concordo totalmente")) +
    xlab("Respostas") + ylab("Perguntas") +
    guides(fill = guide_colorbar(title = "Valores"))
```

Figura 11.6: Gráfico de calor traduzido e sem média.



Fonte: Os autores.

11.2.1 Outra solução

Outra solução para resolver o problema de publicar no idioma desejado pode ser alcançada se modificarmos os fatores. Então, criamos a função *translate_data()* que utiliza a função *dplyr::recode_factor()*(Wickham 2023). A função *recode_factor()* permite recodificar os fatores, ou seja modifica-lo ou reorganizá-lo. A função *translate_data()* retorna o banco de dados com a variável recodificada e foi utilizada para facilitar a leitura do código.

```
# Define a função para traduzir os dados
translate_data <- function(x){ dplyr::recode_factor(x,
    `Strongly disagree` = "Discordo totalmente",
    `Disagree` = "Discordo",
    `Neither agree nor disagree` = "Nem concordo nem
```

```
      discordo",
    `Agree` = "Concordo",`Strongly agree` = "Concordo
      totalmente")}
```

A seguir, iremos utilizar a função *dplyr::mutate()* para criar uma nova variável no banco de dados *mini_pisa_t1* traduzindo a base de dados para o português, e atribuímos o resultado a *minipisa_t2*.

```
# Translate the data using mutate and the translate_
  data function
mini_pisa_t2 <- mini_pisa_t1 %>% mutate(
        across(
            `Eu leio apenas se eu tiver que`:`Para mim,
               ler é uma perda de tempo.`,
            ~ translate_data(.x)  )  )
glimpse(mini_pisa_t2)
```

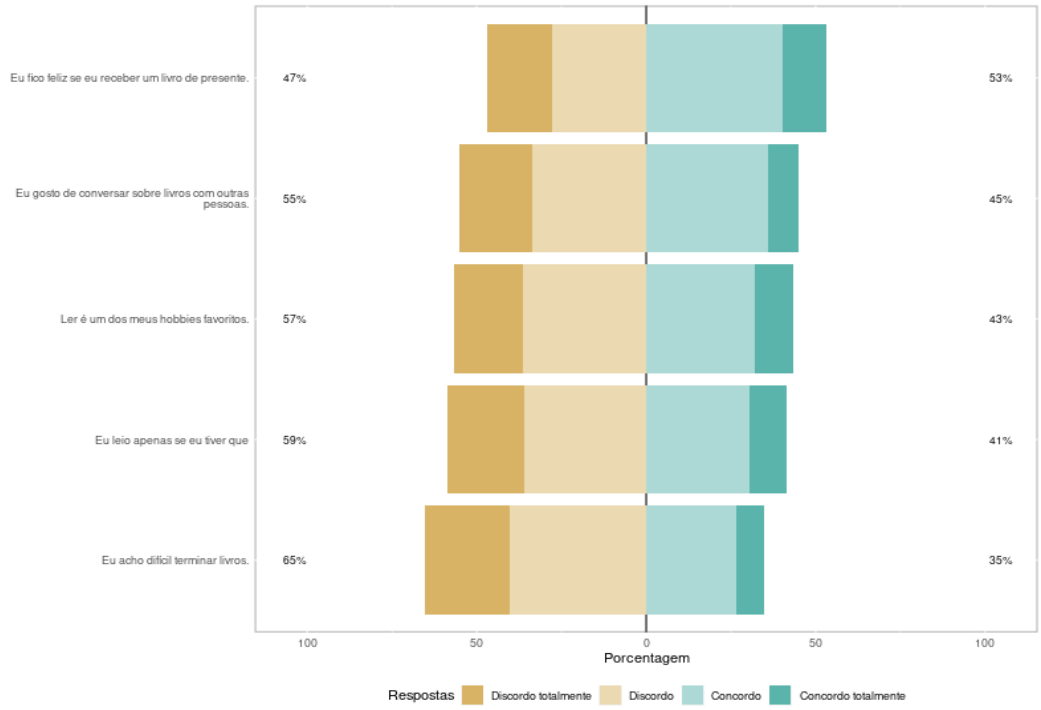
```
Rows: 66,690
Columns: 7
$ País

  <fct> Canada, Cana~
$ `Eu leio apenas se eu tiver que`
                            <fct> Discordo, Co~
$ `Ler é um dos meus hobbies favoritos.`
                      <fct> Concordo tot~
$ ` Eu gosto de conversar sobre livros com outras
  pessoas.` <fct> Concordo tot~
$ `Eu acho difícil terminar livros.`
```

```
                              <fct> Discordo tot~
$ `Eu fico feliz se eu receber um livro de presente.`
        <fct> Concordo tot~
$ `Para mim, ler é uma perda de tempo.`
                           <fct> Discordo tot~
```

Visualizando o gráfico:

```
likert_out <- select(mini_pisa_t2, 2:6)
likert_out <- likert(as.data.frame(likert_out))
plot(likert_out) +
    # os eixos são trocados no likert_out
    labs(y = "Porcentagem") +
    guides(fill = guide_legend("Respostas"))
```

Figura 11.7: Outra opção para tradução.



Fonte: Os autores.

11.2.2 Tabelas

Oportunamente ao apresentarmos os resultados de uma pesquisa necessitamos lançar mão de tabelas e o pacote *gt* nos possibilita este tratamento para os dados com extrema simplicidade.

```
likert_out$results[, 2:5] <- round(likert_out$results[,
    2:5], digits = 2)
gt(likert_out$results) %>%
    cols_label(`Discordo totalmente` = md("Discordo \n
      totalmente")) %>%
    as_latex()
```

Item	Discordo totalmente	Discordo	Concordo	Concordo totalmente
Eu leio apenas se eu tiver que	22.82	35.91	30.54	10.73
Ler é um dos meus hobbies favoritos.	20.32	36.32	31.93	11.42
Eu gosto de conversar sobre livros com outras pessoas.	21.25	33.74	35.96	9.05
Eu acho difícil terminar livros.	24.96	40.39	26.51	8.14
Eu fico feliz se eu receber um livro de presente.	19.28	27.65	40.17	12.89

Finalizaremos este capítulo, na seção seguinte mas certamente haverão muitos outros recursos a explorar na análise de suas pesquisas com o R.

11.3 CONCLUSÃO:

A proposta deste capítulo foi apresentar como contexto a pesquisa com a escala *Likert* quando os dados não estão no idioma desejado, e como utilizar minimamente o pacote *likert* para sua análise no ambiente *tidyverse* utilizando a linguagem R.

Sugerimos um modo operacional que facilita a formulação do questionário e análise de suas hipóteses. Esse procedimento também será útil ao comparar seus resultados com outros trabalhos pertencentes ao seu referencial teórico.

Um dos pacotes mais centrais nos estudos de pesquisas *Likert* utilizando a linguagem R tem sido o pacote de mesmo nome (ROPENSCI, 2023). Porém,

sua publicação no **CRAN** ocorreu em 2016. Assim, é possível que nem todos os recursos mostrados na documentação possam ser usados dependendo da versões do **ggplot2** e de outras dependências instaladas.

Contudo, representa uma simplicidade de utilização que o mantém como uma das principais escolhas.

Exploramos algumas de suas principais características sem esgotar todas as possibilidades e aproveitamos para brevemente introduzir algumas notações do R, e funções do pacote **tidyverse** (WICKHAM; AVERICK et al., 2019), bem como do pacote **gt** (IANNONE et al., 2023).

Para um próximo estudo sugerimos a continuidade da pesquisa para além de uma análise descritiva com a utilização de métodos mais adequados a análise de variáveis categóricas.

11.4 REFERÊNCIAS

ALLAIRE, J. J. et al. **Quarto**. [S.l.: s.n.], 2022. Disponível em: https://doi.org/10.5281/zenodo.5960048.

BRYER, Jason; SPEERSCHNEIDER, Kimberly. **likert: Analysis and Visualization Likert Items**. [S.l.: s.n.], 31 dez. 2016. Disponível em: https://cran.r-project.org/web/packages/likert/index.html.

IANNONE, Richard et al. **gt: Easily Create Presentation-Ready Display Tables**. [S.l.: s.n.], 31 mar. 2023. Disponível em: https://cran.r-project.org/web/packages/gt/index.html.

KOMPERDA, Regis. Likert-Type Survey Data Analysis with R and RStudio. In: [s.l.]: American Chemical Society, 1 jan. 2017. v. 1260, p. 91–116. (ACS Symposium Series, 1260). Section: 7 DOI: 10.1021/bk-2017-1260.ch007.

MALHOTRA, Naresh K. **Pesquisa de marketing: uma orientação aplicada**. [S.l.]: Bookman, 2006.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: [s.n.], 2023. Avilable at: https://www.R-project.org/.

ROPENSCI. **R-Universe: Personal Package Repositories for R!** [S.l.: s.n.], 2023. Disponível em: https://r-universe.dev/search/. Acesso em: 6 out. 2023.

APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

SULLIVAN, Gail M.; ARTINO, Anthony R. Analyzing and Interpreting Data From Likert-Type Scales. **Journal of Graduate Medical Education**, v. 5, n. 4, p. 541–542, dez. 2013. ISSN 1949-8357, 1949-8349. DOI: 10.4300/JGME-5-4-18. Acesso em: 6 out. 2023.

WICKHAM, Hadley; AVERICK, Mara et al. Welcome to the Tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 21 nov. 2019. DOI: 10.21105/joss.01686.

WICKHAM, Hadley; SEIDEL, Dana. **scales: Scale Functions for Visualization**. [S.l.: s.n.], 20 ago. 2022. Disponível em: https://cran.r-project.org/web/packages/scales/index.html.

ISBN 978-65-87023-39-7 (e-book)