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."

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.

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.

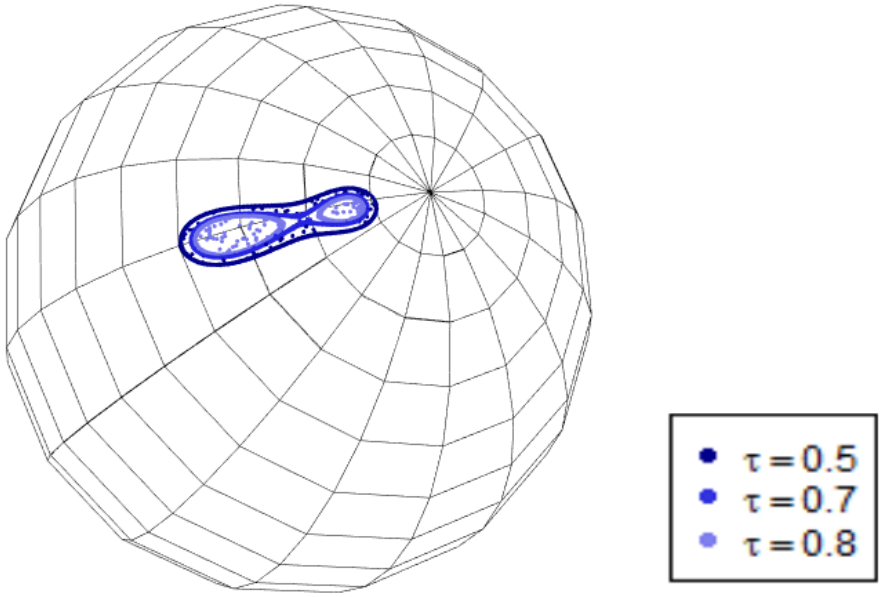
Consider $\tau = 0.8$, on Figure 9.1 and the aim is to find out which municipalities are in each of the two connected components. Figure 9.2 shows the councils

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,

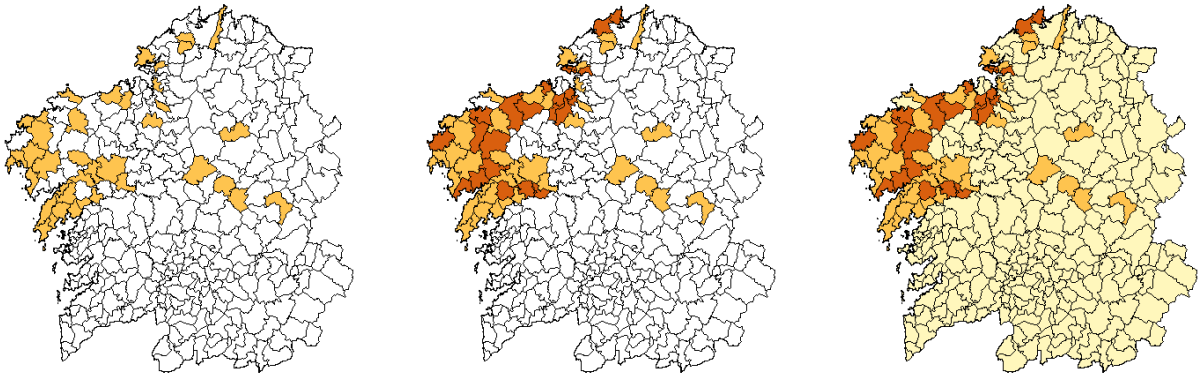
Table 9.2 shows descriptive statistics for the percentages of the different types of surnames, in the related components and for the rest.

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.

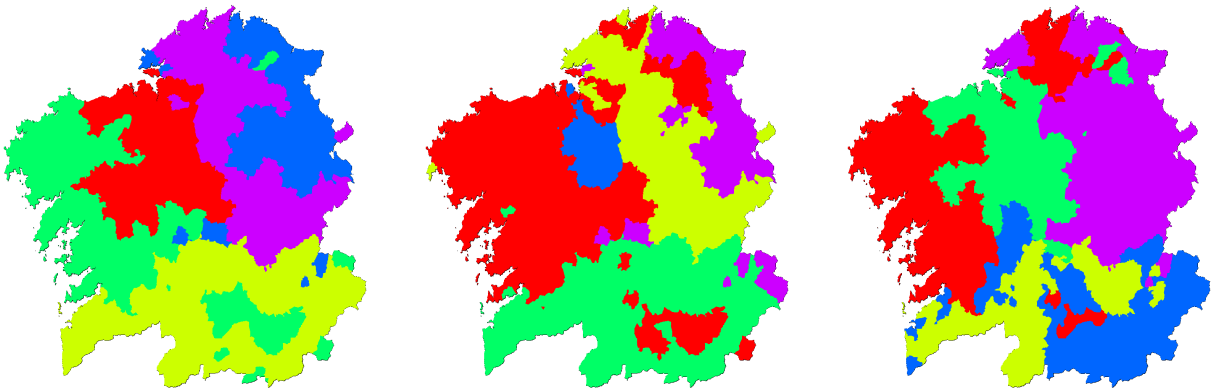
Figures 9.3show the results of the spherical cluster analysis, in 3 different scenarios: with all data (left), only those born in 1965 or earlier (center), and only those born in 1945 or earlier (right).

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

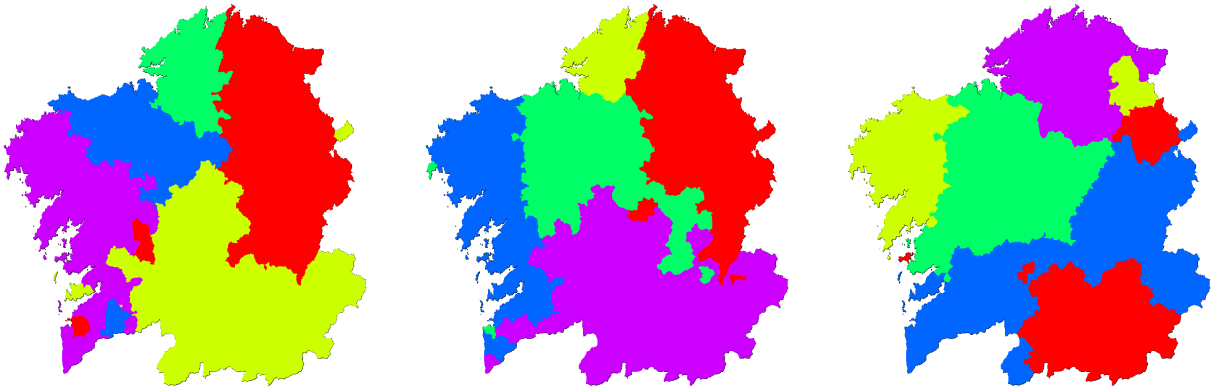
APLICAÇÕES EM R: ENCURTANDO DISTÂNCIAS NAS CIÊNCIAS Portal de Livros Abertos da USP, Pirassununga: Faculdade de Zootecnia e Engenharia de Alimentos, 2024. 286 p. ISBN 978-65-87023-39-7(e-book). Disponível em: https://doi.org/10.11606/9786587023397.

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.

ISBN 978-65-87023-39-7 (e-book)