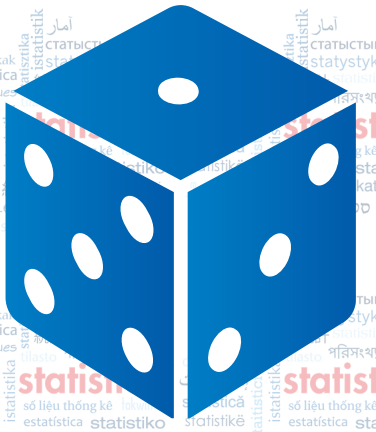
Boletim

SPE
Sociedade Portuguesa de Estatística

Publicação semestral — outono de 2024

Replicabilidade e Controlo de Confidencialidade

Informação Editorial
Endereço: Sociedade Portuguesa de Estatística.
Campo Grande. Bloco C6. Piso 4.
1749-016 Lisboa. Portugal.
Telefone: +351.217500120
e-mail: *spe*@spestatistica.pt
URL: https://www.spestatistica.pt
ISSN: 1646-5903
Depósito Legal: 249102/06
Tiragem: Edição digital
Execução Gráfica e Impressão: Gráfica Sobreirense
Editor: Fernando Rosado, *fernando.rosado@fc.ul.pt*



SPE

Sociedade Portuguesa de Estatística

https://www.spestatistica.pt

Sociedade Portuguesa de Estatística desde 1980

Junta-te à

SPE
Sociedade Portuguesa de Estatística

"Se já és sócio da SPE, incentiva os teus colegas, colaboradores e alunos a juntarem-se à SPE"

A SPE

- Oferece descontos em congressos e outros eventos organizados pela SPE.
- Oferece distinções e reconhecimento através dos seus prémios.
- Oferece oportunidades para ampliares a tua rede de contactos através da comunidade SPE.
- Oferece aos sócios acesso a um sistema de acreditação internacional.
- Valoriza sócios, comunidade e profissão apostando na educação, formação e inovação.
- Defende a profissão e molda o seu futuro.

Junta-te à SPE:

https://www.spestatistica.pt/socios/admissao-formulario

Quota anual

- Regular: 30€
- Estudante: 15€
- Promotor*: 0€
- Recém-licenciado†: grátis!

† Até um ano após terminar a licenciatura.
* Um "sócio promotor" angaria 2 novos sócios por ano.

Editorial

… "dias mundiais, internacionais, nacionais"…

1. Um determinado tema central no Boletim SPE para além da divulgação e atualização científica de um assunto específico – um objetivo estatutário da SPE (Cf. artigo 1º, números 2 e 3) – é também a concretização de direitos dos sócios. No *Boletim SPE outono 24* unem-se, ainda mais, aqueles desideratos de comunicação e direitos dos sócios; pois, ele é fruto de generosa colaboração de colegas, na sua maior parte, colaboradores do Banco de Portugal. Ao incluir um *Tema Central* de um dos seus primeiros sócios institucionais, esta edição releva a (eventual) falha, que pode ressaltar ao não dar o devido relevo a todos aqueles sócios que "contribuem para o progresso e prestígio da SPE" (Cf. artigos 6 e 7 dos Estatutos da SPE).
Embora alguns dos seus membros já tenham colaborado em diversas edições, esta é a primeira vez que o Banco de Portugal, como instituição, é incluído na listagem dos autores e dos Temas Centrais do Boletim SPE, com o título, bem apelativo: *Replicabilidade e Controlo de Confidencialidade*. Esta edição do Boletim SPE teve a preciosa coedição da colega Rita Sousa a quem é devido um agradecimento especial pela generosa colaboração.

2. O **Dia das Nações Unidas**, ou **Dia da ONU**, é comemorado em 24 de outubro, desde 1948, como aniversário da entrada em vigor, em 1945, da Carta das Nações Unidas. A celebração foi determinada em 1947 pela Assembleia Geral de forma que o dia fosse "dedicado a fazer serem conhecidos pelos povos do mundo os objetivos e conquistas da ONU e a obter apoio para" a sua obra. A data compõe a Semana da ONU, que vai de 20 a 26 de outubro. O que esperar desta realização?
As Nações Unidas elegem também dias específicos para marcar acontecimentos ou assuntos relevantes com o objetivo de promover, através da consciencialização e da ação, os objetivos da Organização. Geralmente, são um ou mais Estados-membros que propõem estas comemorações e a Assembleia Geral estabelece-as através de uma resolução.
Ocasionalmente, estas datas são declaradas pelas agências especializadas das Nações Unidas, quando se referem a questões que se enquadram no âmbito das suas competências. Algumas datas podem ser posteriormente adotadas pela Assembleia Geral.
Sobre dias internacionais temos, uma listagem, em https://unric.org/pt/dias-internacionais/.
Em cada mês de outubro temos o **Dia Mundial dos Professores**: "World Teachers Day is held annually on 5 October to celebrate all teachers around the globe. It commemorates the anniversary of the adoption of the 1966 ILO/UNESCO Recommendation concerning the Status of Teachers, which sets benchmarks regarding the rights and responsibilities of teachers, and standards for their initial preparation and further education, recruitment, employment, and teaching and learning conditions. The Recommendation concerning the Status of Higher-Education Teaching Personnel was adopted in 1997 to complement the 1966 Recommendation by covering teaching personnel in higher education. World Teachers' Day has been celebrated since 1994".
Em 20 de Outubro e sobre o lema "Serviço – Profissionalismo – Integridade" "surgiu" o **Dia Mundial da Estatística**, criado em 20.10.2010 e a ser celebrado de cinco em cinco anos. "At its 41st Session in February 2010, the United Nations Statistical Commission proposed celebrating 20 October 2010 as World Statistics Day (Decision 41/109). Acknowledging that the production of reliable, timely statistics and indicators of countries' progress is indispensable for informed policy decisions and monitoring implementation of the Millennium Development Goals, the General Assembly adopted on 3 June 2010 resolution 64/267, which officially designated 20 October 2010 as the first ever World Statistics Day under the general theme "Celebrating the many achievements of official statistics."

Com todos os objetivos propostos, na sua realização anual os Dias Internacionais estimulam a concretização de novos tijolos no edifício da História e Ciência no mundo e, bem assim, nos mais diversos países. E assim, uma motivação mundial pode gerar e contribuir para o interesse nacional. Estes Dias Mundiais, induzem as concretizações Nacionais e nestas surge, por exemplo, o "Dia da Estatística".

3. "A Ciência em geral e a Estatística em particular é uma nobre atividade, necessária ao corpo e ao espírito, indispensável ao bem-estar e à felicidade. Mas, a ciência é cara. Assim, só os ricos a podem praticar… e os pobres se a praticam ficam mais pobres. Embora exigindo grande esforço e dedicação a solução deve estar em (apesar de tudo) fazer ciência para caminhar na saída daquele dilema. É o que se exige aos estatísticos portugueses congregados em torno de um projeto líder – a SPE! Esta, a Sociedade Portuguesa de Estatística reúne investigadores e fazedores de Ciência". (Cf. Rosado (2005). *Memorial da Sociedade Portuguesa de Estatística,* Edições SPE, p. 139-50).

No entanto, para "fazer Ciência", também se exige e é essencial o contributo nacional. Neste, muitos intervenientes podem surgir como vitais e alguns podem mesmo ser das origens menos esperadas. São exemplo disso os "estrangeiros, estrangeirados e os outros" intrínsecos no "triunvirato editorial" recentemente oferecido na página da SPE (em Publicações (spestatistica.pt) e do qual damos notícia neste boletim) – após as três recentes publicações disponibilizadas na página da SPE – trabalhos da autoria dos nossos colegas Dinis Pestana, Ivette Gomes, Maria Antónia Turkman e Rui Santos.

Com tudo isto se cria a História (também da Estatística). E todos são fundamentais. Os obreiros diretos, os que põem as mãos na massa, são os autores de "pequenos contributos" que unidos formam "a obra". O *Boletim SPE* deve ser também um "pequeno contributo" que, em cada publicação, é acrescentado pela colaboração valiosa e generosa das já muitas dezenas de autores. O Editor é (apenas) o arauto desse possível valor. São aqueles autores que, através do Boletim também ajudam a concretizar (no terreno) os objetivos dos Dias Mundiais.

O Tema Central do próximo Boletim será

Teoria das Filas de Espera

Mensagem do Presidente

Caros Sócios da SPE

A Sociedade Portuguesa de Estatística (SPE) tem como um dos seus principais objetivos valorizar os seus sócios e o seu contributo para o desenvolvimento da Estatística em Portugal. Neste sentido, a atual direção da SPE está empenhada não apenas em promover e reconhecer o papel dos sócios ativos, mas também em preservar o legado dos estatísticos que fundaram a nossa Sociedade, cujo trabalho e dedicação merecem ser perpetuados.

Este compromisso é particularmente evidente na celebração do centenário do nascimento de Bento José Ferreira Murteira, sócio fundador da SPE, distinguido como sócio honorário e laureado com o Prémio Carreira SPE em 2013. Bento Murteira teve uma influência determinante no ensino e na evolução da Estatística em Portugal, com uma longa carreira como docente no Instituto Superior de Economia e Gestão (ISEG).

Para assinalar esta data especial, a SPE e o ISEG, em parceria com o Centro de Estatística e Aplicações (CEAUL) e o Centro de Matemática Aplicada à Previsão e Decisão Económica (CEMAPRE), realizarão um evento no dia 15 de novembro de 2024. Este evento será uma oportunidade para honrar o legado de Bento José Ferreira Murteira e reconhecer a sua influência no desenvolvimento da Estatística em Portugal.

Durante as comemorações, será lançada uma nova edição do livro *Estatística: Inferência e Decisão*, que foi previamente editado em 1988 pela Imprensa Nacional-Casa da Moeda. A recuperação desta obra emblemática visa disponibilizá-la gratuitamente para toda a comunidade académica e científica.

A recuperação de publicações históricas é uma das prioridades da atual direção, que está igualmente empenhada em preservar o legado de outros estatísticos de renome. Neste âmbito, destacamos que foram recentemente disponibilizados no site da SPE três documentos da autoria de Dinis Pestana e Rui Santos, Ivette Gomes, e Maria Antónia Turkman. Estes textos resultam das comunicações apresentadas no 37º Encontro do Seminário Nacional de História da Matemática, realizado em junho de 2024, na Academia das Ciências de Lisboa e no Museu Nacional de História Natural e da Ciência.

Aproveito também para recordar que o XXVII Congresso Nacional da SPE decorrerá no Algarve. Este evento marcará o regresso a esta encantadora região e promete reunir a comunidade estatística, proporcionando mais uma oportunidade para a troca de ideias e conhecimentos. O site oficial do congresso, com informações detalhadas sobre o evento, será divulgado muito brevemente.

Por fim, é com satisfação que recordo a criação das redes sociais da SPE, conforme prometido. Estas plataformas foram desenvolvidas para aproximar os sócios, dinamizar a comunicação e divulgar as atividades da Sociedade. Apelamos a todos os sócios que acompanhem e interajam nas nossas redes, contribuindo para promover e fortalecer a comunidade estatística em Portugal.

Com o contributo e participação de todos, acreditamos que conseguiremos continuar a expandir e consolidar o papel da SPE na promoção da Estatística.

Cordiais saudações

Luís Machado

Notícias

• JOCLAD 2024

Decorreu entre 18 e 20 de abril de 2024, as **XXXI Jornadas de Classificação e Análise de Dados** (JOCLAD 2024), organizadas pela Associação Portuguesa de Classificação e Análise de Dados (CLAD) e a Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria (ESTG — IPLeiria) na cidade de Leiria.

A SPE foi convidada a apresentar uma sessão temática cujo título deste ano foi "Statistical Methods Applied to Health Sciences" moderada por Luís Meira Machado. A sessão contou com a colaboração de três oradores:
1. Vera Afreixo "Stable genome-wide selection procedure: an application to Alzheimer's disease", trabalho conjunto com Ana Helena Tavares;
2. Maria João Polidoro "Statistical modeling methods applied to kidney transplantation outcomes", trabalho conjunto com Ana Pinho; e
3. Luís Meira-Machado "Flexible odds ratio modeling for mortality risk in acute coronary syndrome patients", trabalho conjunto com Marta Azevedo.

MJP

11th European Conference on Quality in Official Statistics (Q2024)

De 4 a 7 de junho de 2024, realizou-se o **11th European Conference on Quality in Official Statistics (Q2024)** organizado pelo Instituto Nacional de Estatística (INE) e o Eurostat, no Centro de Congressos do Estoril.

A SPE participou no evento, através da sua comissão técnica CT225, com uma comunicação oral intitulada "Applications of Statistical Methods and (Inter)national Standardization" para dar a conhecer a importância da Normalização nas estatísticas oficiais.

MJP

• ISO/TC 69 "Applications of statistical methods"

No passado mês de junho, na semana de 17 a 21 de 2024, na cidade de Amarante realizou- se a reunião plenária anual do Comité Técnico *ISO/TC 69 – Applications of statistical methods* da *International Organization for Standardization* (ISO).
A organização e acompanhamento da reunião ficou a cargo da equipa de gestão da Comissão Técnica Nacional, CT 225- Aplicação de métodos estatísticos que integra os especialistas que representam Portugal na ISO/TC 69 e que é presidida pela Sociedade Portuguesa de Estatística (SPE) e secretariada pela CCDR-Norte, IP.
A reunião plenária contemplou 33 fóruns técnicos em formato híbrido, em várias salas que o Município de Amarante disponibilizou. Neste evento técnico estiveram presentes as delegações de especialistas vindos da Alemanha, China, Estados Unidos da América, Índia, Inglaterra, Japão, França e Portugal.
As atividades técnico-científicas no âmbito da Normalização estiveram focadas nas temáticas que ISO/TC 69 desenvolve, nomeadamente:

Acceptance sampling

Measurement methods and results

Applications of statistical and related techniques for the implementation of Six Sigma

Application of statistical and related methodology for new technology and product development

Statistical interpretation of data

Big data analytics

Terminology and symbols

Bayesian methods Coordination Group (New)

Destaques e resultados alcançados

1. Projetos em curso, num total de 20 projetos normas em desenvolvimento, em diferentes fases da aprovação.
2. Aprovação de 4 novos projetos do Grupo de trabalho *Big data analytics*, para os temas "Principles and foundations of Big Data Analytics" (Japão), "Curation, cleansing and wrangling of big and large datasets" (Portugal); "(Business) Transformations methodologies (incl. Six Sigma & Lean) and its integration with Big Data" (UK); "EDA and Data visualization of large and highly dimensional datasets" (China).
3. Nomeação de uma especialista nacional como *Project Leader* para o projecto "Curation, cleansing and wrangling of big and large datasets".
4. Proposta de nomeação de uma especialista nacional para coordenar o *Working Group Terminology and symbols* (ISO/TC 69/WG13), por um período de 3 anos (2025-2017).
5. Decisão da Comissão de introduzir os Métodos Bayesianos, nas várias áreas temáticas que a ISO/TC 69 desenvolve.

Balanço global

O balanço global da reunião plenária foi muito positivo, com 34 decisões tomadas; nove agendas de trabalho foram cumpridas; e o formato presencial das reuniões foi considerado pelos participantes o aspeto mais positivo para a concretização dos trabalhos em agenda.

A reunião plenária da ISO/TC 69 em 2025 terá como anfitrião a China, mas por agora ficam as memórias dos 50 membros que participaram no evento de Amarante.

Os sócios da SPE que queiram participar nas atividades normativas da CT225 ou obter mais informações, podem fazê-lo através do seguinte *link* https://www.spestatistica.pt/seccoes-e-comissoes.

MJP

• Dia Internacional das Mulheres na Estatística e Ciência de Dados

3º Dia Internacional das Mulheres na Estatística e Ciência de Dados

No dia 8 de outubro de 2024, o 3º Dia Internacional Anual da Mulher na Estatística e Ciência de Dados (IDWSDS 2024) foi comemorado através de uma conferência global virtual de 24 horas, organizada pelo *Caucus for Women in Statistics and Data Science* (CWS), em parceria com a Sociedade Portuguesa de Estatística (SPE) e a *American Statistical Association* (ASA). O evento, acessível através do site oficial https://www.idwsds.org, manteve o compromisso de destacar e valorizar as mulheres que atuam nas áreas da Estatística e Ciência de Dados em todo o mundo.

Sob o tema *"Empowering the Next Generation of Women Statisticians and Data Scientists"*, a conferência deste ano apresentou quatro sessões plenárias, mais de 60 sessões, e reuniu 180 palestrantes oriundos de 26 países, representando os cinco continentes. A distribuição geográfica dos palestrantes foi a seguinte: África (3%), América (47%), Ásia (5%), Europa (31%) e Oceania (4%). Destes, 80% eram provenientes do meio académico, 7% estavam vinculados a entidades governamentais, e 13% atuavam na indústria. As palestras abrangeram uma vasta gama de tópicos técnicos em estatística e ciência de dados, partilha de experiências profissionais, conselhos, e reflexões sobre a história da estatística, com contributos de oradoras reconhecidas internacionalmente.

O programa completo do evento está disponível em https://www.idwsds.org/program/.

A sessão de abertura da conferência contou com a participação da Presidente do CWS, Cynthia Bland, a Presidente da ASA, Bonnie Ghosh-Dastidar e a vice-Presidente da SPE, Lisete Sousa.

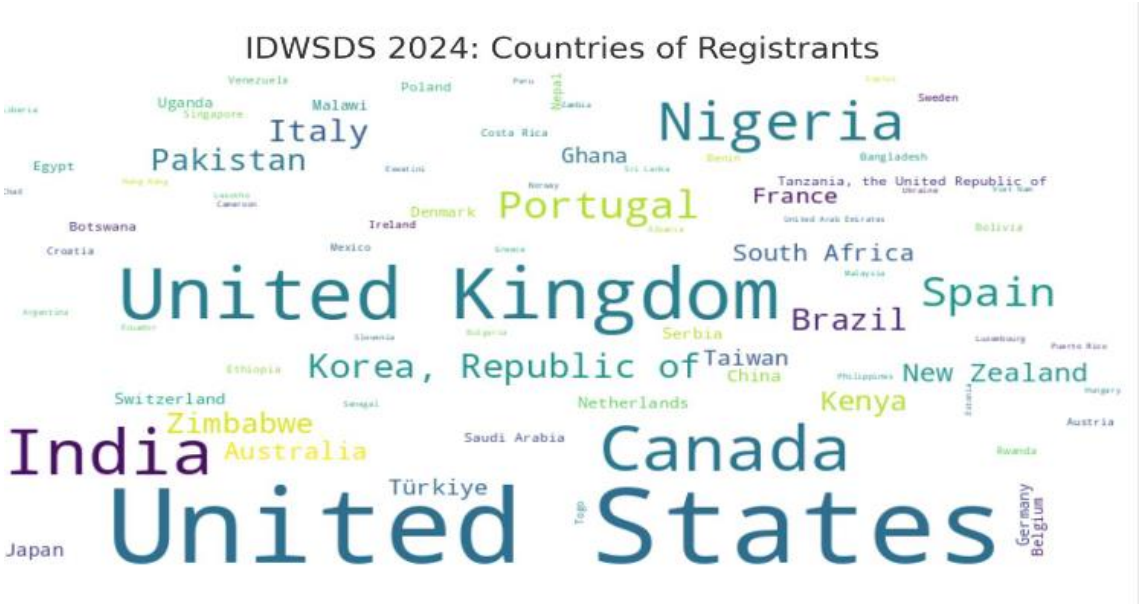
Este ano, a SPE esteve presente com a organização de duas sessões na conferência. A primeira sessão, intitulada "*Next Generation: Showcasing Young Portuguese Talent in Statistics and Data Science*", contou com a participação de Daniela Silva, investigadora, que apresentou a palestra "*Joint Model for Zero-Inflated Data Combining Fishery-Dependent and Fishery-Independent Sources*", Maria Almeida Silva, da Universidade Lusófona, que falou sobre "*Stratified Sampling of Billing Data and Singular Spectrum Analysis for Smart Meter Installation and Flow Time Series Decomposition in Drinking Water Distribution Systems*", e Vanessa Silva, da Universidade do Porto, com a palestra "*Multivariate Time Series Analysis via Multilayer Graphs*".

A segunda sessão foi organizada pela Secção de Biometria da SPE e teve o título "*Next Generation: Showcasing Young Portuguese Talent in Biometry and Data Science*". Nesta sessão, Ana Martins, da Universidade de Aveiro, apresentou uma comunicação intitulada "*Spatiotemporal Models for Time*

Series of Counts", enquanto Patrícia Soares, do Instituto Nacional de Saúde Doutor Ricardo Jorge, falou sobre "*Beyond the Instrumental Inequalities*".

Ambas as sessões destacaram o talento emergente português nas áreas de Estatística, Biometria e Ciência de Dados, reforçando o papel da SPE no apoio ao desenvolvimento de jovens investigadores.

A importância de iniciativas globais que promovam a união e apoio às mulheres nestas áreas foi um dos destaques do evento, que este ano registou a participação de mais de 500 pessoas de 77 países.



O sucesso da conferência reforçou a necessidade de continuar a promover a diversidade e a inclusão num campo onde as mulheres ainda se encontram em menor número.

Gostaríamos de expressar os nossos mais sinceros agradecimentos a todos os participantes, oradores e patrocinadores, cujo apoio foi essencial para o êxito do 3º Dia Internacional da Mulher na Estatística e Ciência de Dados. A dedicação de cada um de vós impulsiona a nossa missão de celebrar e capacitar as mulheres neste campo tão crucial.

Para aqueles que não puderam participar, as gravações da conferência estarão brevemente disponíveis no nosso canal do YouTube: https://youtube.com/@cwstat9787?si=G_eYyR0YAGpmzvNF.

Esperamos continuar esta jornada e poder contar com a vossa valiosa participação no evento do próximo ano, previsto para 14 outubro de 2025. Juntos, podemos continuar a causar um impacto positivo e duradouro no mundo da estatística e da ciência de dados. Fiquem atentos para mais atualizações sobre a conferência do próximo ano.

Obrigada por fazerem parte desta comunidade global!

Em nome da Comissão Organizadora,

Lígia Henriques-Rodrigues (Universidade de Évora)

Vanda Lourenço (Universidade NOVA de Lisboa)

• Novas Publicações no Site da Sociedade Portuguesa de Estatística

Foram disponibilizadas novas publicações no site da Sociedade Portuguesa de Estatística (SPE). Os três documentos resultaram de comunicações apresentadas no **37.º Encontro do Seminário Nacional de História da Matemática**.

Os novos documentos disponibilizados são os seguintes:

Como os Livros Contam uma História da Estatística em Portugal: Obras de Estrangeiros, de Estrangeirados e dos Outros
Ano: 2024
Autores: Dinis Pestana e Rui Santos

Breve Historial da Escola de Extremos e Aplicações em Portugal (PORTSEA)
Ano: 2024
Autora: M. Ivette Gomes

Relembrando Bento Murteira no ano do seu Centenário
Ano: 2024
Autora: Maria Antónia Amaral Turkman

Os documentos podem ser consultados (aqui)

A SPE agradeceu publicamente aos autores das comunicações por partilharem o seu conhecimento e por terem acedido a disponibilizar os respetivos documentos para o usufruto de todos os sócios.

FR

• Secretariado da SPE

A equipa de secretariado da SPE passou a ser liderada por Cristina Santos, que assegurará a continuidade dos trabalhos com toda a experiência adquirida, após um período de transição de funções em colaboração com Telma Matias.
A Direção da SPE expressou o agradecimento a Telma Matias pelo empenho e dedicação demonstrados durante o tempo em que desempenhou a função, contribuindo significativamente para o bom funcionamento da Sociedade Portuguesa de Estatística. A Telma Matias também é devido o testemunho de agradecimento pela excelente colaboração nas tarefas de Edição do Boletim SPE que ao secretariado respeitam.
A comunicação com o secretariado deverá continuar a ser feita, como habitualmente, através do correio eletrónico spe@spestatistica.pt.
Todos os contactos estão disponíveis em https://www.spestatistica.pt.

FR

• Declaração sobre Ética Profissional

A SPE – Sociedade Portuguesa de Estatística é membro institucional do ISI – Instituto Internacional de Estatística. O ISI está envolvido. no estabelecimento de uma declaração sobre ética profissional, ao longo de décadas. A primeira declaração foi apresentada na Assembleia Geral na celebração do Centenário em 1985. O objetivo da declaração era fornecer orientação e não regulamentação.

Com o passar do tempo, o ISI decidiu investigar a necessidade de atualizar a declaração e, consequentemente, foi adotada uma versão substancialmente revista em 2010. Embora o conteúdo da declaração de 2010 permaneça válido, a utilização crescente de uma diversidade de fontes de dados, conjuntos de dados interligados e métodos estatísticos computacionalmente intensivos exigiram algumas atualizações que foram introduzidas em 2023. A versão em inglês da declaração atualizada está disponível (aqui); também em português.

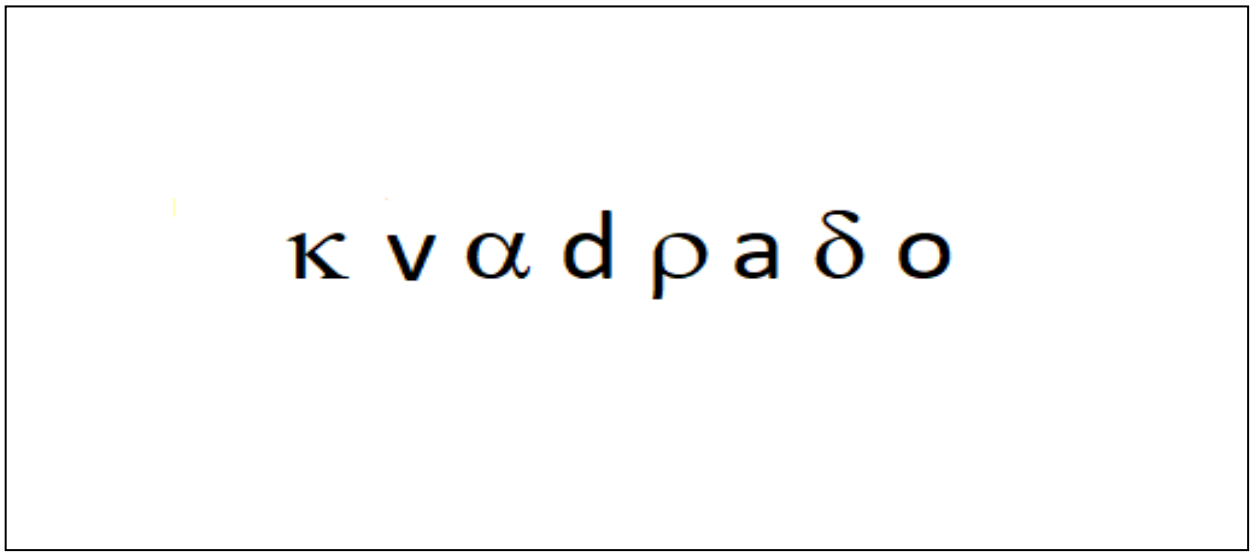
Mais informa o ISI: "O objetivo da declaração é permitir que os julgamentos e decisões éticas individuais sejam informados por valores e princípios partilhados. Reconhece que a operação de um princípio pode impedir a operação de outro princípio. Os estatísticos podem ter de fazer escolhas entre princípios. Não tenta resolvê-los com regras rígidas a serem aplicadas. Em vez disso, fornece um quadro para ajudar os estatísticos a tomar estas decisões, compreendendo e considerando ativamente fatores concorrentes nas circunstâncias que enfrentam".

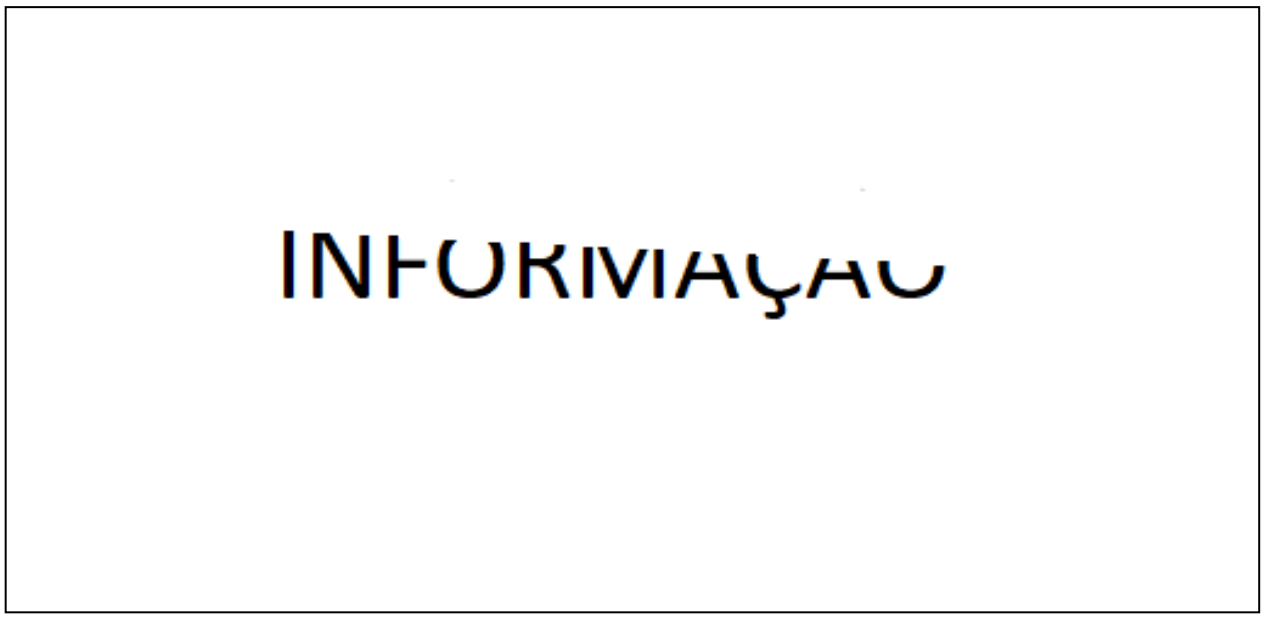
A referida versão, no Preâmbulo, anuncia "ao que vem":

"A Declaração de Ética Profissional do ISI consiste numa declaração de Valores Profissionais Partilhados e num conjunto de Princípios Éticos derivados desses valores. Para os efeitos do presente documento, a definição de quem é um estatístico vai muito além daqueles com diplomas formais na área, incluindo um vasto leque de criadores e utilizadores de dados e ferramentas estatísticas. Os estatísticos trabalham numa variedade de contextos económicos, culturais, jurídicos e políticos, cada um dos quais influencia a ênfase e o foco da investigação estatística. Também trabalham num dos muitos ramos diferentes da sua disciplina, cada um envolvendo as suas próprias técnicas e procedimentos e, possivelmente, a sua própria abordagem ética. Qualquer que seja a sua área de especialização, os princípios éticos dos estatísticos são parte integrante da sua competência profissional e devem fazer parte de uma formação estatística abrangente. Os estatísticos trabalham em diversas áreas como, por exemplo, a economia, a psicologia, a sociologia e a medicina, cujos profissionais possuem convenções éticas que podem influenciar a sua conduta. Mesmo dentro do mesmo contexto e ramo da estatística, os indivíduos podem enfrentar várias situações e condicionalismos em que possam surgir questões éticas. O objetivo desta declaração é permitir que os juízos e decisões éticas individuais do estatístico sejam informados por valores e experiências partilhadas e não por regras rígidas impostas pela profissão (…)"

FR

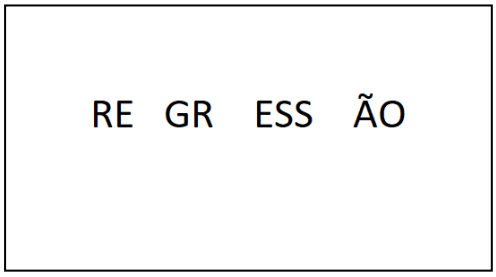
Enigmística de mefqa



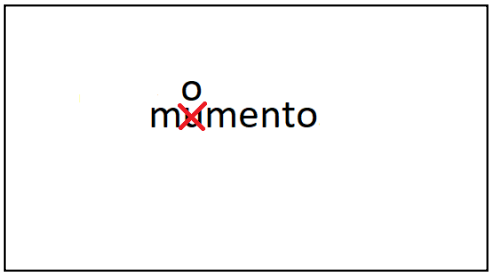


Enigmas 45 e 46

No Boletim SPE primavera de 2024 (p. 14):



Enigma 43: regressão por troços



Enigma 44: momento corrigido

Episódio na História da Estatística

Centenário do Nascimento do Professor Bento Murteira
1924 – 2024

Comemoração

15 de novembro de 2024 – Salão Nobre do ISEG

Em 2024, celebra-se o centenário do nascimento do Professor Bento José Ferreira Murteira.
O Prof. Bento Murteira foi uma personalidade que, profundamente, marcou o florescimento da Probabilidade e Estatística em Portugal, no século passado.
Deixou-nos um legado ímpar que não é demais recordar.
O Prof. Bento Murteira inegavelmente merece um lugar na História da Matemática Aplicada e da Estatística em Portugal.

Neste contexto, a Sociedade Portuguesa de Estatística (SPE) e o Instituto Superior de Economia e Gestão (ISEG), juntamente com o Centro de Estatística e Aplicações (CEAUL) e o Centro de Matemática Aplicada à Previsão e Decisão Económica (CEMAPRE), decidiram assinalar esta data com a realização de um evento, que irá decorrer no Salão Nobre do ISEG, no dia 15 de novembro de 2024, das 16:30 às 19:30 e com o seguinte programa:

Abertura:

Professor João Duque - Presidente do ISEG
Professor Luís Machado - Presidente da SPE

Testemunhos:

Maria Antónia Turkman
Carlos Silva Ribeiro
Nuno Crato
João Pedro Faria
Isabel Proença
Inês Murteira

Coffee break/Convívio final

O Prof Bento Murteira foi um dos 11 membros co-fundadores da *Sociedade Portuguesa de Estatística* em 28 de Novembro de 1980 e foi eleito primeiro *Sócio Honorário SPE* em 1993.
No Boletim SPE, de entre as diversas referências, encontramos:
- No *Boletim SPE outono de 2016*, p. 60, o último livro que publicou no ano de 2015, em co-autoria, com o título Introdução à Estatística.
- No *Boletim SPE outono de 2018*, p. 3-10, por ocasião do seu falecimento.

No evento será lançada uma nova edição do livro *Estatística: Inferência e Decisão*, editado em 1988 pela Imprensa Nacional-Casa da Moeda, e que se encontra esgotado.
Uma edição digital do livro ficará disponível, em local a designar para livre acesso.
Mais informações em: https://www.iseg.ulisboa.pt/event/comemoracao-do-centenario-do-nascimento-do-professor-bento-murteira/

Maria Antónia Turkman
Fernando Rosado

INQUÉRITO ÀS CONDIÇÕES DE VIDA, ORIGENS E TRAJETÓRIAS DA POPULAÇÃO RESIDENTE EM PORTUGAL 2023

Instituto Nacional de Estatística

O Inquérito às Condições de Vida, Origens e Trajetórias da População Residente em Portugal (ICOT), realizado em 2023, visa conhecer a dimensão de cada um dos grupos étnicos com os quais a população residente em Portugal se identifica, e permitir a sua caracterização. O objetivo principal é compreender de que forma as pessoas se autoidentificam, e como relatam e interpretam as suas origens, para compreender e combater a discriminação e desigualdades em vários domínios. Pretende-se desta forma contribuir para que o sistema estatístico português disponha de dados oficiais relativos à origem e pertença étnica da população residente em Portugal, e respetiva caracterização.

Combater o racismo e a discriminação étnica, assim como obter dados e conhecimento acerca desta temática com o intuito de produzir e apoiar a definição de políticas públicas é uma prioridade para a Comissão Europeia, expressa no Plano de ação da União Europeia contra o racismo 2020-2025. Ao nível nacional, foi criado em 2020 o Grupo de Trabalho para a Prevenção e o Combate ao Racismo e à Discriminação (Despacho n.º 309-A/2021, de 8 de janeiro). Adicionalmente, a Resolução da Assembleia da República n.º 11/2021 recomenda "a realização de estudos que conduzam à recolha de informação estatística, através do organismo responsável pela estatística nacional, relativa à discriminação étnico-racial" e a Resolução da Assembleia da República n.º 16/2021 recomenda a elaboração e implementação de uma estratégia nacional de combate ao racismo. A presente operação encontra-se prevista no Plano Nacional de Combate ao Racismo e à Discriminação 2021-2025 (Resolução do Conselho de Ministros n.º 101/2021).

Tendo em vista esse propósito, foi realizado em 2021/2022 o Inquérito piloto às Condições de Vida, Origens e Trajetórias da População Residente, com o principal objetivo de testar o desenho amostral e os modos de recolha que mais se adequavam, os conteúdos e a adesão dos respondentes às temáticas inquiridas. Com base na informação obtida a partir dos resultados do inquérito piloto, o INE desenvolveu uma proposta de questionário final sobre as temáticas da origem, pertença, trajetórias e discriminação implementado à escala nacional.

O inquérito tem uma natureza multidimensional, visando caracterizar tanto quanto possível essa diversidade e possibilitar, em consequência, a exploração analítica entre as diferentes características da população e a vivência de experiências de discriminação em diversos domínios. Efetivamente, pretende-se avaliar as condições de vida nas suas múltiplas expressões, como sejam o acesso e a qualidade do emprego, saúde, educação, habitação, mobilidade, redes de socialização. A pertença do ponto de vista étnico resulta de uma autoclassificação das pessoas e a origem foi observada pela naturalidade do respondente e dos seus ascendentes, até à terceira geração.

O ICOT é um inquérito amostral, cuja informação foi recolhida diretamente junto das unidades de observação – indivíduos dos 18 aos 74 anos de idade que residiam há pelo menos um ano em Portugal (ou cuja intenção de residência era de pelo menos um ano) – mediante um modo de recolha misto, CAPI (*Computer-Assisted Personal Interview)*, CATI (*Computer-Assisted Telephone Interview)*, e CAWI (*Computer-Assisted Web Interview)*, dando oportunidade aos respondentes de utilizarem o modo que mais lhes convém.

O inquérito foi aplicado em todo o território nacional, entre janeiro e agosto de 2023, a uma amostra de 35 035 unidades de alojamento, constituindo a maior amostra de inquéritos às famílias realizados pelo

INE. Foi entrevistada apenas uma pessoa por alojamento, selecionada pelo método do último aniversário no alojamento, tendo sido obtidas 21 608 entrevistas completas.
Os resultados foram calibrados tendo por referência as estimativas anuais da população residente em 31 de dezembro de 2022 (base Censos 2021).

Informação disponível

- Para uma análise mais detalhada da metodologia seguida, sugere-se a leitura do documento metodológico do ICOT, no Portal das Estatísticas Oficiais;
- Os principais resultados podem ser consultados em destaque1 e destaque2, disponíveis no mesmo Portal;
- Poderão também ser consultados os 24 indicadores estatísticos disponíveis na Base de Dados do Portal do INE, selecionando o tema *Condições de Vida e Cidadania* e o subtema *Origens, Trajetórias e Pertença Étnica*;
- A base de microdados anonimizados do ICOT encontra-se disponível para utilização em fins científicos por parte de investigadores, após a sua devida acreditação. Em caso de interesse no acesso a esta Base de microdados poderá consultar aqui o processo de pedido e as condições de acesso.

INQUÉRITO SOBRE SEGURANÇA NO ESPAÇO PÚBLICO E PRIVADO 2022

Instituto Nacional de Estatística

O Inquérito sobre Segurança no Espaço Público e Privado (ISEPP), realizado em 2022, visa contribuir para a consolidação de um sistema de informação estatístico europeu sobre a temática da violência de género e violência doméstica. É uma operação estatística financiada pela Comissão Europeia (CE), e consta do Programa Estatístico Europeu (PEE) para 2021-2027.

Combater a violência de género e a violência doméstica e melhorar o conhecimento sobre esta temática para apoio à definição de medidas de política constitui uma prioridade aos níveis europeus e nacional expressa, designadamente, nos seguintes documentos estratégicos: Compromisso Estratégico para a Igualdade de Género 2016-2019 e, mais recentemente, Estratégia Europeia para a Igualdade de Género 2020-2025; Plano de Ação para a Prevenção e o Combate à Violência Contra as Mulheres e à Violência Doméstica (PAVMVD), inscrito na Estratégia Nacional para a Igualdade e a Não Discriminação (ENIND) — Portugal + Igual.

Adicionalmente, a Convenção de Istambul, de 2011, de que Portugal é signatário desde 2013, no seu art.º 11, introduz a obrigatoriedade de recolha regular de dados sobre violência de género e violência doméstica através de inquéritos à população que abranjam todas as formas de violência referidas na Convenção (física, sexual, psicológica e económica).

Neste contexto, foi constituído um grupo de trabalho no Sistema Estatístico Europeu (SEE), com o Eurostat, no qual Portugal está representado pelo INE, para desenvolver, à escala europeia, um inquérito com enfoque nas questões da violência de género. O objetivo deste grupo de trabalho era desenvolver e testar a metodologia de um inquérito à população para a recolha de estatísticas representativas sobre a prevalência e caracterização da violência de género nos Estados Membros, em consonância com os requisitos definidos na Convenção de Istambul.

Tendo em vista esse propósito, foi realizado em 2019 um inquérito piloto, para testar a metodologia, em termos de modos de entrevista e abrangência (de zonas rurais e urbanas; de homens e mulheres; e de população adulta, sem limite etário superior). Os resultados do inquérito piloto apoiaram a elaboração de um questionário mais completo adotado ao nível europeu na operação estatística principal. O grupo de trabalho do SEE desenvolveu uma proposta de metodologia e de questionário sobre a temática da violência de género e da violência doméstica, implementada à escala europeia, com vista à obtenção de dados ao nível europeu, harmonizados e comparáveis. É neste contexto que se insere o ISEPP, cujos principais conceitos e definições, bem como orientações técnicas e metodológicas de recolha de dados, seguem as recomendações definidas no Manual metodológico desenvolvido pelo Eurostat para esse efeito.

O ISEPP é um inquérito amostral, cuja informação foi recolhida diretamente junto das unidades de observação – homens e mulheres com idade dos 18 aos 74 anos, residentes em unidades de alojamento de residência principal – mediante um modo misto sequencial, que combinou a recolha por preenchimento via web (CAWI), com a recolha por entrevista telefónica (CATI) e presencial (CAPI), para as unidades de alojamento que não responderam por aquela via.

O inquérito foi aplicado em todo o território nacional, entre julho e início de outubro de 2022, a uma amostra de 21 030 unidades de alojamento. Foi entrevistada apenas uma pessoa por alojamento, selecionada pelo método do último aniversário no alojamento. Foram obtidas 11 346 entrevistas completas.

Os resultados foram calibrados tendo por referência as estimativas anuais da população residente em 31 de dezembro de 2022 (base Censos 2021).

Informação disponível

- Para uma análise mais detalhada da metodologia seguida, sugere-se a leitura do documento metodológico do ISEPP 2022, no Portal das Estatísticas Oficiais;
- Os principais resultados podem ser consultados em destaque1 e destaque2, disponíveis no mesmo Portal;
- Poderão também ser consultados os 25 indicadores estatísticos disponíveis na Base de Dados do Portal do INE, selecionando o tema *Justiça* e o subtema *Violência no Espaço Público e Privado*;
- A base de microdados anonimizados do ISEPP encontra-se disponível para utilização em fins científicos por parte de investigadores, após a sua devida acreditação. Em caso de interesse no acesso a esta Base de microdados poderá consultar aqui o processo de pedido e as condições de acesso.

ALEA: 25 ANOS AO SERVIÇO DA LITERACIA ESTATÍSTICA E CIDADANIA

Instituto Nacional de Estatística

Em 2024, o ALEA (Ação Local de Estatística Aplicada) celebra 25 anos de existência, consolidando-se como instrumento de promoção da literacia estatística em Portugal. Criado com o objetivo de aproximar a estatística da comunidade escolar e do público em geral, o ALEA tem difundido conteúdos educativos que incentivam a compreensão e o uso das estatísticas oficiais no quotidiano, promovendo não só o conhecimento, mas também a cidadania ativa.

O projeto ALEA (https://alea.ine.pt) é uma iniciativa conjunta do Agrupamento de Escolas Tomaz Pelayo, de Santo Tirso, e do Instituto Nacional de Estatística, à qual se associou a antiga Direção Regional de Educação do Norte, formalizando uma parceria através de um acordo de colaboração tripartido. Posteriormente, a Direção-Geral dos Estabelecimentos Escolares assumiu o papel que antes era o da Direção Regional de Educação do Norte.

Com uma plataforma digital rica em recursos interativos, assente no triângulo Espaço Lúdico - Estatística - Dados (onde predominam cursos, jogos, desafios envolvendo milhares de alunos e dados estatísticos atualizados), o ALEA tem permitido que alunos e professores explorem e compreendam o papel da estatística nas mais variadas áreas da sociedade. Ao longo dos seus 25 anos, o projeto tem-se reinventado, respondendo às necessidades contemporâneas da educação e à crescente importância da literacia estatística no mundo digital.

Recentemente, o ALEA disponibilizou uma área de **Educação para a Cidadania**, que promove a **literacia estatística** como ferramenta essencial para formar cidadãos críticos e informados. Através de conteúdos educativos e interativos, esta área incentiva o uso da estatística na interpretação de dados, o que é fundamental para a tomada de decisões conscientes em áreas da vida pública e pessoal tais como economia, saúde e políticas públicas. Ao capacitar indivíduos com competências estatísticas, o ALEA contribui para uma sociedade mais participativa e democrática.

Esta área tem sido complementada com as **TarefALEAs,** tarefas com as quais se pretende responder a algumas questões de Cidadania envolvendo temas como clima, envelhecimento e desigualdade utilizando dados estatísticos oficiais produzidos pelo INE, Eurostat, Banco de Portugal, entre outras entidades, e sistematizados pela PORDATA. A ideia é desafiar os estudantes a aplicar conceitos estatísticos em contextos reais e variados, incentivando a compreensão e o raciocínio crítico sobre dados. Essas atividades são voltadas para o desenvolvimento de competências na interpretação e análise de informações estatísticas, contribuindo para a formação de cidadãos mais informados e participativos. A TarefALEA disponibilizada mais recentemente incide na desigualdade salarial entre homens e mulheres, e destina-se a alunos do 10.º ano de escolaridade. Nela se propõe a abordagem do tópico “Dados

quantitativos bivariados" e dos subtópicos "Diagrama de dispersão", "Coeficiente de correlação linear" e "Reta de regressão". Privilegia-se, ainda, a interpretação de gráficos.

Para comemorar os 25 anos do ALEA, está a ser organizado um evento que irá ocorrer no próximo dia 15 de novembro na Escola Tomaz Pelayo em Santo Tirso, com o seguinte programa:

1) Introdução pelos Diretores (ESTP, DGestE, INE)
2) Intervenção da Diretora da Rede de Bibliotecas Escolares
3) Intervenção sobre as TarefALEAs e Cidadania (Emília Oliveira)
4) Intervenção do Prof. Jaime Carvalho e Silva

11ª CONFERÊNCIA EUROPEIA DA QUALIDADE (Q2024)

Instituto Nacional de Estatística

No rescaldo da 11ª Conferência Europeia da Qualidade (Q2024), que teve lugar de 4 a 7 de junho, no Estoril, é com uma enorme satisfação que damos conta da realização bem-sucedida deste evento. Esta Conferência, organizada pelo Instituto Nacional de Estatística I.P. (INE) Portugal e EUROSTAT, acolheu mais de 500 participantes, dos quais naturalmente destacamos a presença expressiva dos colegas do INE, que muito nos orgulha.

Este evento, promovido pelo EUROSTAT e iniciado em 2001, tem-se realizado rotativamente em cada 2 anos nos países que integram o Sistema Estatístico Europeu e com a organização dos respetivos Institutos de Estatística.

A Q2024 realizou-se sob o lema "**As estatísticas oficiais enquanto pilar da democracia**" - resulta do entendimento de que não é suficiente afirmar que Estatísticas Oficiais são um bem público produzidas com independência profissional, imparcialidade e compromisso com a qualidade. Para além deste importante tópico, foi possível abordar domínios tão díspares quanto complementares, tais como: a gestão da qualidade, a confidencialidade, a informação geoespacial, o uso e integração de novas fontes de dados, a inteligência artificial, a relação com os utilizadores, a coordenação e governação dos sistemas estatísticos, entre muitos outros tópicos de relevo.

Conforme referido anteriormente, o INE participou ativamente nas sessões de trabalho, com a apresentação de diversas práticas inovadoras e moderação de inúmeras sessões. Todos os artigos e apresentações encontram-se disponíveis no site da conferência.

No primeiro dia, prévio ao início da Conferência, realizaram-se quatro cursos de formação. Estes cursos foram cuidadosamente concebidos e adotadas abordagens de conhecimentos teóricos e práticos, para a melhor compreensão do papel da qualidade nas estatísticas oficiais. Este dia foi, também, palco de um evento paralelo, um seminário organizado por Portugal, dedicado à cooperação com os parceiros da Comunidade dos Países de Língua Portuguesa, subordinado à temática: "Qualidade nos INE da CPLP: Liderar o Futuro".

Em todo o evento, e num clima de descontração, existiu um grande envolvimento dos participantes, que permitiu a troca de experiências, o debate, o desenvolvimento de *networking* e a partilha de conhecimentos, investigação e desenvolvimentos em matéria de estatísticas oficiais.

São destacados os principais números da Q2024, ilustrados na infografia abaixo e onde estão presentes os participantes do INE no evento.

Para informação mais detalhada sobre o evento, sugerimos a consulta do site oficial em https://q2024.pt. Os momentos mais importantes apresentam-se no vídeo: "Q2024 – best of".

CERTIFICAÇÃO ISO DA GESTÃO DA QUALIDADE E DA SEGURANÇA DA INFORMAÇÃO NO INSTITUTO NACIONAL DE ESTATÍSTICA I.P.

Instituto Nacional de Estatística

Foi com uma enorme satisfação que a equipa do Instituto Nacional de Estatística I.P. (INE) recebeu a notícia da APCER (Associação Portuguesa de Certificação) sobre a certificação do seu Sistema de Gestão Integrado, que abarca conjuntamente i) o Sistema de Gestão da Qualidade pela Norma NP EN ISO 9001, para o âmbito total da Organização - "Produção e divulgação de estatísticas oficiais do Instituto Nacional de Estatística", e ii) o Sistema de Gestão de Segurança de Informação do INE pela Norma NP EN ISO 27001, para o âmbito "no âmbito dos processos das estatísticas de comércio internacional – intra e extra UE", cujo sistema de gestão tem aplicabilidade em todo o INE, e neste caso a manutenção da certificação iniciada em 2019.

A Produção de Estatísticas Oficiais é um processo complexo, de enorme relevância em todas as Nações e com referenciais Europeus e Internacionais no que respeita às suas várias fases, e subprocessos. As estatísticas oficiais que dele resultam constituem um pilar da democracia.

Há muitos anos que o INE tem vindo a trabalhar com as Normas ISO em várias áreas, e a relativa à Gestão da Qualidade esteve desde muito cedo a acompanhar-nos na forma de organização e documentação dos nossos processos. Trabalhar de forma sistematizada a gestão da qualidade e a melhoria contínua, tem sido um desafio constante e presente nos vários exercícios estratégicos e de planeamento das atividades do INE. Obter a certificação pela ISO 9001 nunca foi um objetivo por si, mas antes uma oportunidade para consolidar e melhorar a sistematização dos nossos processos e demonstrar a consistência dos sistemas. A opção por um Sistema de Gestão Integrado, que juntou a Gestão de Segurança da Informação surgiu como um passo lógico, dado que a própria segurança da informação faz parte do processo de produção de estatísticas oficiais.

Este resultado vem assim confirmar o nosso caminho e robustecer o *compliance* com o nosso enquadramento legal aos níveis nacional e europeu, juntando-nos ao grupo dos outros Institutos Nacionais de Estatística dos estados-membros da União Europeia que já obtiveram a mesma certificação.

Estamos certos de que este processo se reveste de grande relevância para o INE como um todo e para as suas equipas, bem como para os seus respondentes, prestadores de informação e utilizadores, na prossecução do seu Valor Compromisso para com a Qualidade.

Nota Editorial

O Tema Central do *Boletim SPE outono 2024*, como habitual, insere-se no principal objetivo desta publicação da *Sociedade Portuguesa de Estatística* especialmente no que concerne a divulgação científica de assuntos generalistas com óbvio benefício da comunidade em geral e da Ciência Estatística em particular.

O Tema Central do *Boletim SPE outono 2024*
Inclui temática "pouco divulgada" e muito específica em determinados assuntos científicos. Esta é uma motivação para um alinhamento editorial especial onde a "ordem na sequência dos textos" pode ser de relevância.

O texto inicial é "definitório" em jeito de introdução ao Tema. E as três "razões" seguintes:

1. Preocupações das regras de confidencialidade no âmbito de um "fornecedor" de dados, que é o departamento de Estatística;
2. Conceitos e Metodologias associadas ao controlo de divulgação estatística, geralmente utilizadas na preparação dos dados, antes de serem disponibilizados;
3. Preocupações e ferramenta de replicação, disponibilizada pelo BPLIM, para quem já está a utilizar os dados.

aconselham a sequência editorial dos restantes textos.

FR

Nota Introdutória

Replicabilidade e Controlo de Confidencialidade

Introdução

Este Boletim pretende dar a conhecer alguns contributos no âmbito da Replicabilidade e do Controlo da Confidencialidade. Os autores, dos trabalhos que se seguem, reúnem aqui um conjunto de conceitos e terminologias que contextualizam o tema em questão.

Entende-se por **Replicabilidade**, o ato de repetir um estudo independentemente do investigador em questão, sem a utilização de dados originais, mas geralmente utilizando os mesmos métodos [1]. A **Confidencialidade** está relacionada com a propriedade dos dados, e normalmente resulta de medidas legislativas, que impedem a divulgação não autorizada [4].

Segundo o **Princípio do Segredo Estatístico**, todos os dados estatísticos individuais recolhidos pelas autoridades estatísticas são de natureza confidencial. Este princípio visa salvaguardar a privacidade dos cidadãos e garantir a confiança no **SEN - Sistema Estatístico Nacional** [2]. O SEN compreende o Conselho Superior de Estatística (CSE), o Instituto Nacional de Estatística, IP (INE, IP), o Banco de Portugal, os Serviços Regionais de Estatística das Regiões Autónomas dos Açores e da Madeira, e as entidades produtoras de estatísticas oficiais por delegação do INE, IP. Com exceção do CSE, as restantes entidades, na qualidade de responsáveis pela produção das estatísticas oficiais, são consideradas autoridades estatísticas. As autoridades estatísticas, no respetivo âmbito de atuação, podem exigir o fornecimento, com carácter obrigatório e gratuito, a todos os serviços ou organismos, pessoas singulares e coletivas, de quaisquer elementos necessários à produção de estatísticas oficiais e estabelecer a recolha de dados que, ainda que não relevantes para a atividade específica das entidades obrigadas ao seu fornecimento, revistam importância estatística [2].

O **Regulamento Geral de Proteção de Dados (RGPD)**, que entrou em vigor em 2018, tem como objetivo principal adequar as leis de privacidade dos dados na Europa controlando o tratamento de dados pessoais por indivíduos, empresas ou organizações.

A **informação estatística** é classificada como **confidencial** quando permite a identificação dos inquiridos ou de qualquer outra pessoa singular ou coletiva, entidade ou sucursal, quer diretamente através do nome, do endereço ou de um código de identificação oficialmente atribuído, quer indiretamente por meio de dedução, revelando, desse modo, informações de ordem individual [7].

Os métodos de **Controlo de Divulgação Estatística** (CDE) compreendem um conjunto de métodos e técnicas estatísticas para controlo e redução do risco de divulgação de informações que identifiquem indivíduos, empresas ou outras organizações, geralmente baseados na restrição da quantidade ou na modificação dos dados divulgados [3].

Segundo a Comissão Europeia, o **Risco de Identificação** é o risco de se identificar e divulgar informação de uma determinada **Unidade Estatística** - unidade elementar de uma população ou universo [4]. Os métodos de CDE estabelecem um compromisso entre o risco de identificação e a perda de informação que, por definição de um nível aceitável de risco, procuram maximizar a utilidade dos dados. Neste contexto, a **Utilidade dos Dados** refere-se à utilidade dos dados anonimizados para análises estatísticas por utilizadores finais, bem como a validade dessas análises quando realizadas nesses mesmos dados [5].

Define-se **Anonimização** como o processo em que as Informações Pessoais Identificáveis (IPI) são modificadas de forma irreversível, de modo que a entidade não possa ser identificada, direta ou indiretamente, pelo responsável pelo tratamento de IPI, seja individualmente ou em colaboração com qualquer outra pessoa [6]. A anonimização é plena quando a aplicação dos métodos de CDE permite eliminar completamente, direta ou indiretamente, o risco de identificação de uma unidade estatística.

Este Boletim descreve os principais desafios, práticas e ferramentas utilizadas na proteção da privacidade dos dados, nomeadamente em resultados estatísticos produzidos e divulgados pelo Departamento de Estatística do Banco de Portugal e em bases de Microdados disponibilizadas pelo Laboratório de Investigação em Microdados do Banco de Portugal (BPLIM).

Referências

[1] "A Simple Explanation for the Replication Crisis in Science", Roger Peng, de 24 de agosto, 2016, http://simplystatistics.org/2016/08/24/replication-crisis/.

[2] "Diário da República", Lei n.º 22/2008, de 13 de maio, https://diariodarepublica.pt/dr/detalhe/lei/22-2008-249237.

[3] "Eurostat CROS - Collaboration in Research and Methodology, Statistical Disclosure Control", https://cros.ec.europa.eu/book-page/statistical-disclosure-control.

[4] "INE, Sistema Integrado de Metainformação – Conceitos", https://smi.ine.pt/.

[5] "World Bank, Statistical Disclosure Control: A Practice Guide", Thijs Benschop, Cathrine Machingauta, Matthew Welch, de 13 de outubro, 2022, https://readthedocs.org/projects/sdcpractice/downloads/pdf/latest/.

[6] "ISO, International Organization for Standardization", 2011SO/IEC 29100:2011, https://www.iso.org/obp/ui/#iso:std:iso-iec:29100:ed-1:v1:en:sec:A.

[7] "Regulamento do Conselho (CE)", Regulamento n.º 951/2009, de 9 de outubro, https://www.bportugal.pt/sites/default/files/anexos/legislacoes//regce951ano2009.PDF.

Rita Sousa

Controlo de disseminação estatística: Desafios da confidencialidade nas estatísticas da Central de Balanços

Ana Bárbara Pinto, apinto@bportugal.pt
Diogo Barbosa, dfbarbosa@bportugal.pt
Lara Mak, lmak@bportugal.pt
Sónia Mota, scmota@bportugal.pt

Departamento de Estatística, Banco de Portugal

1. Princípios gerais para o controlo de disseminação estatística

1.1. Enquadramento legal

A procura por estatísticas mais granulares, aliada à necessidade de obter dados de forma cada vez mais célere, coloca desafios às autoridades estatísticas nacionais. Para além disso, é crucial garantir a qualidade e a integridade dos dados. Encontrar um equilíbrio entre a preservação da confidencialidade e a garantia da qualidade, tempestividade e integridade das estatísticas publicadas é imperativo para construir uma relação de confiança com o público. A confidencialidade estatística, também conhecida como segredo estatístico, é um princípio essencial regido pela Lei do Sistema Estatístico Nacional (SEN) (Lei n.º 22/2008, de 13 de maio) que estabelece a obrigação de proteger a privacidade e a confidencialidade dos dados individuais recolhidos pelos órgãos do SEN. Neste artigo analisamos os princípios gerais para o controlo da disseminação estatística e mostramos, a título de exemplo, como estas medidas são aplicadas na divulgação dos dados das estatísticas das empresas publicadas pelo Banco de Portugal.

O ponto 1 do artigo n.º 13 da Lei Orgânica do Banco de Portugal atribui ao Banco a responsabilidade de recolher e compilar estatísticas monetárias, financeiras, cambiais e da balança de pagamentos, designadamente no âmbito da sua colaboração com o Banco Central Europeu (BCE). Para cumprir esta responsabilidade, o Banco está autorizado a solicitar os dados necessários de entidades públicas e privadas. Como autoridade estatística nacional e membro do Sistema Estatístico Nacional (SEN), o Banco de Portugal segue os princípios e normas estabelecidos pela Lei n.º 22/2008 de 13 de maio, especialmente no que diz respeito à salvaguarda do segredo estatístico (artigo n.º 6).

De acordo com o n.º 2 do artigo 6.º da Lei do SEN:

"Todos os dados estatísticos individuais recolhidos pelas autoridades estatísticas são de natureza confidencial, pelo que:

a) Não podem ser cedidos a quaisquer pessoas ou entidades nem deles ser passada certidão, sem prejuízo do disposto no nº 3 do artigo 18.º;
b) Nenhum serviço ou autoridade pode ordenar ou autorizar o seu exame;
c) Não podem ser divulgados de modo que permitam a identificação direta ou indireta das pessoas singulares e coletivas a que respeitam;
d) Constituem segredo profissional, mesmo após o termo das funções, para todos os funcionários, agentes ou outras pessoas que, a qualquer título, deles tomem conhecimento no exercício ou em razão das suas funções relacionadas com a atividade estatística oficial."

O princípio do segredo estatístico estabelece a obrigação de proteger a privacidade e a confidencialidade dos dados individuais recolhidos pelos órgãos do SEN. Este princípio garante que os

dados de natureza individual recolhidos para fins estatísticos têm natureza confidencial e não podem ser divulgados a terceiros.

Como membro do Sistema Europeu de Bancos Centrais, o Banco de Portugal cumpre o Regulamento do Conselho (CE) n.º 951/2009 de 9 de outubro, que aborda a compilação e proteção de informações estatísticas confidenciais pelo Banco Central Europeu, conforme definido no artigo n.º 8. Adicionalmente, o n.º 12 do artigo 1.º do mesmo Regulamento define o conceito de informação estatística confidencial como sendo "informação estatística que permita a identificação dos inquiridos ou de qualquer outra pessoa singular ou coletiva, entidade ou sucursal, quer diretamente através do nome, do endereço ou de um código de identificação oficialmente atribuído, quer indiretamente por meio de dedução, revelando, desse modo, informações de ordem individual. Para determinar se um inquirido ou qualquer outra pessoa singular ou coletiva, entidade ou sucursal, podem ou não ser identificados, devem considerar-se todos os meios que possam ser razoavelmente utilizados por terceiros para identificar o inquirido em questão ou a outra pessoa singular ou coletiva, entidade ou sucursal".

1.2. Critérios de confidencialidade

Os critérios de confidencialidade são estabelecidos com o objetivo de minimizar a perda de informações úteis, tendo em consideração fatores como a natureza dos dados, o tamanho da população e os objetivos da análise estatística. Em termos da divulgação de informação estatística, é crucial equilibrar cuidadosamente a confidencialidade dos agentes e a utilidade dos dados. Desta forma são seguidas três regras elementares (Figura 1):

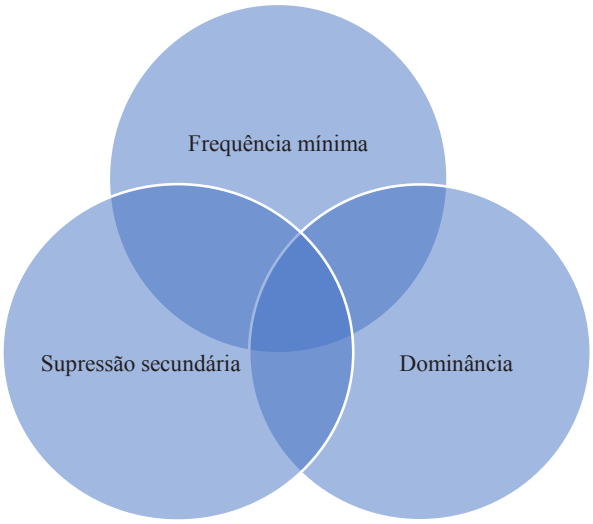


Figura 1. Três principais regras de confidencialidade

1. **Frequência mínima:** conforme a Deliberação n.º 188 do Conselho Superior de Estatística, as entidades autorizadas devem divulgar dados estatísticos de forma a impedir a identificação direta ou indireta das unidades estatísticas. Isso requer a agregação de dados com pelo menos três unidades por variável ou conjunto de variáveis.
2. **Dominância:** o princípio da dominância refere que se uma célula for explicada por um pequeno número de agentes, a mesma não deve ser disseminada. Esta proteção é adotada porque células dominantes podem potencialmente levar à identificação de indivíduos quando combinadas com outras informações.
3. **Supressão Secundária:** a supressão secundária envolve tratar células não-confidenciais como confidenciais para evitar a dedução e identificação de uma célula confidencial, identificada pelas duas regras anteriores, tornando impossível o recálculo dos valores. Esta técnica é utilizada para assegurar que a informação confidencial não possa ser deduzida a partir das células publicadas, mantendo a integridade e a confidencialidade do conjunto de dados.

Estará tudo sujeito ao controlo de disseminação estatístico e à confidencialidade?

Não, existem exceções. O parágrafo 2 do artigo 6.º da Lei do Sistema Estatístico Nacional prevê exceções relativas à confidencialidade estatística, especificando que "Salvo disposição legal em contrário, os dados estatísticos individuais relativos à Administração Pública não estão abrangidos pela confidencialidade estatística.". No que diz respeito à regulamentação internacional, o parágrafo 7 do artigo 8.º do Regulamento (CE) n.º 951/2009 refere que "as informações estatísticas obtidas de fontes acessíveis ao público ao abrigo da legislação nacional não são consideradas confidenciais.".

2. Uma aplicação prática: os Quadros do Setor da Central de Balanços

2.1. Os Quadros do Setor e a sua fonte dos dados

O Banco de Portugal publica um conjunto alargado de informação económico-financeira sobre as empresas em Portugal. Além da informação estatística que divulga no portal BPstat, no domínio das empresas da Central de Balanços, o Banco de Portugal também disponibiliza os Quadros do Setor.

Nos Quadros do Setor, é possível analisar, de forma interativa, um conjunto de indicadores económico-financeiros para os vários setores de atividade económica. Os indicadores estão disponíveis desde 2006 e são atualizados anualmente no site do Banco de Portugal (Informação sobre empresas | Banco de Portugal).

Os Quadros do Setor destacam-se pelo seu elevado nível de detalhe. A classificação das empresas por setor de atividade económica é feita ao nível mais detalhado possível (5 dígitos, que corresponde à subclasse), de acordo com a CAE-Rev.3. Para além disso, a informação também é divulgada por classe de dimensão das empresas. O critério utilizado na classificação das empresas por dimensão corresponde ao da Recomendação da Comissão Europeia, de 6 de maio de 2003, relativa à definição de micro, pequenas e médias empresas.

Os indicadores económico-financeiros são compilados, desde 2006, a partir dos dados individuais das empresas reportados no Anexo A da Informação Empresarial Simplificada (IES). A IES é um reporte anual que agrega obrigações declarativas de natureza contabilística, fiscal e estatística para cinco entidades: Ministério da Justiça, Ministério das Finanças, Banco de Portugal, Instituto Nacional de Estatística e Direção Geral das Atividades Económicas. Antes da implementação da IES, os indicadores económico-financeiros provinham do inquérito anual da Central de Balanços, realizado pelo Banco de Portugal desde 1983 e descontinuado em 2006 (dados até 2005) decorrente da implementação da IES.

2.2. A confidencialidade nos Quadros do Setor

Para a elaboração dos Quadros de Sector, recorre-se primordialmente a informação proveniente da IES. Adicionalmente, esta informação é combinada com outras fontes de dados não públicas. Por este motivo, o processo de construção dos Quadros de Sector está sujeito a um rigoroso controlo de disseminação estatística.

Os Quadros do Setor deverão ser divulgados para todas as subclasses da CAE-Rev.3 e classes de dimensão que contenham empresas em atividade desde que cumpridos os requisitos de confidencialidade estabelecidos.

Quando não é possível publicar os dados ao nível da subclasse, consideram-se níveis superiores de agregação: classe (quatro dígitos), grupo (três dígitos), divisão (dois dígitos) e secção (uma letra). Para a classe de dimensão existem seis níveis: "Grandes empresas", "Médias empresas", "Pequenas empresas", "Microempresas", "Micro, Pequenas e Médias Empresas" e "Todas as dimensões".

Para preservar a confidencialidade dos dados individuais das empresas da Central de Balanços e garantir a relevância económica dos indicadores económico-financeiros, não serão divulgados Quadros do Setor para os agregados onde se verifique alguma das seguintes condições:

- Menos de três empresas;

- O volume de negócios de uma empresa contribui para mais de 85% do volume de negócios do agregado[1];
- O volume de negócios e o ativo total são inferiores a 1.000€ e não existem pessoas ao serviço (considera-se que o agregado não apresenta sinais de atividade relevante).

A implementação direta das regras anteriores determina, em primeiro lugar, que um conjunto de agregados não seja divulgado. Em segundo lugar, a divulgação de outros agregados também pode ser suprimida após a implementação de várias decisões condicionais. O objetivo é, portanto, evitar que um agregado em falta seja identificado através de outros agregados publicados. Com base nas condições anteriores, o setor de atividade e a classe de dimensão dos Quadros do Setor são publicados com o máximo detalhe possível.

Independentemente da divulgação de outros indicadores, a publicação dos quartis relativos à distribuição dos rácios económico-financeiros de um agregado também está sujeita às seguintes regras:

- Os quartis da distribuição só são divulgados se o número de empresas do agregado for igual ou superior a 11;
- Se um agregado tiver entre 6 e 10 empresas, apenas é apresentado o 2.º quartil;
- Para agregados com menos de 6 empresas, não é fornecida qualquer informação de quartis.

3. Conclusão

Num contexto em que o acesso a dados de maior granularidade é crucial para a tomada de decisão informada, proteger a confidencialidade dos dados é também essencial para manter a integridade e a confiança do público. Portanto, é imperativo estabelecer e implementar princípios e métodos robustos de confidencialidade. A divulgação de estatísticas por parte do Banco de Portugal baseia-se em três princípios fundamentais: (i) a frequência mínima, garantindo que os dados estatísticos são divulgados apenas quando há um número mínimo de unidades por variável ou conjunto de variáveis base; (ii) a dominância, que proíbe a divulgação de células se as mesmas forem explicadas por um pequeno número de agentes; (iii) e a supressão secundária, que trata células não confidenciais como confidenciais para evitar a divulgação inadvertida de informação confidencial.

Referências

1. Suplemento n.º 2 ao Boletim Estatístico, Banco de Portugal, 2013.
2. Estudos da Central de Balanços – Quadros do setor e quadros da empresa e do setor nº 36, Banco de Portugal 2019.
3. D'Aguiar, Luís, Menezes, Paula. *Impact and benefits of micro-databases' integration on the statistics of the Banco de Portugal*. Proceedings of the 59th World Statistics Congress Hong Kong, 2013.
4. Duncan, George, Elliot, Mark, Salazar-González, Juan-José *Statistics for Social and Behavioral Sciences*. Springer, New York, New York 2011.
5. Lei do Sistema Estatístico Nacional (SEN) (Lei n.º 22/2008, de 13 de maio)
6. Lei Orgânica do Banco de Portugal (Lei n.º 5/98 de 31 de janeiro, com as alterações introduzidas pelos Decretos-Leis n.º 118/2001, de 17 de abril, n.º 50/2004, de 10 de março, e n.º 39/2007, de 20 de fevereiro, Decreto-Lei n.º 31-A/2012, de 10 de fevereiro, Decreto-Lei n.º 142/2013, de 18 de outubro, Lei n.º 23-A/2015, de 26 de março e Lei n.º 39/2015, de 25 de maio).

[1] O critério da Dominância foi recentemente revisitado, em conformidade com as práticas internacionais, com o intuito de reforçar a proteção da informação. Neste contexto, o critério passará a abranger as duas principais empresas contribuidoras.

Métodos de Controlo de Divulgação Estatística

Jorge Morais, *jogerogerio98@hotmail.com*
Softinsa

Rita Sousa, *rcsousa@bportugal.pt*
BPLIM, Departamento de Estudos Económicos, Banco de Portugal
Centro de Matemática e Aplicações, FCT-UNL

Susana Faria, *sfaria@math.uminho.pt*
CMAT, Departamento de Matemática, Universidade do Minho

1. Introdução e Terminologia

A procura pelo acesso a dados tem vindo a crescer muito nos últimos anos. Nesse sentido, o compromisso entre a utilidade da informação disponibilizada e a proteção da confidencialidade é cada vez mais importante. As técnicas de Controlo de Divulgação Estatística (CDE) sugerem métodos para modificar os dados de forma a serem publicados sem revelarem informações confidenciais que possam estar associadas a indivíduos específicos (Benschop *et al*., 2021; Rao *et al*., 2012). Um método ideal de CDE é aquele que minimiza o risco de identificação e maximiza a utilidade dos dados. Quanto menor o risco de identificação, maior a perda de informação e, por consequência, menor a utilidade dos dados para os utilizadores. Quando a modificação dos dados é grande, o risco de identificação diminui, mas a diferença entre a base de dados original e a base de dados modificada aumenta.

Neste estudo pretende-se apresentar as técnicas que melhor equilibrem o *trade-off* entre o risco de identificação e a utilidade dos dados. Apresentam-se ainda várias medidas de utilidade dos dados e de risco de identificação para avaliar o desempenho dos diferentes métodos. Descrevem-se e comparam-se diferentes abordagens de perturbação baseadas principalmente em modelos lineares e não lineares. Por último, apresentam-se várias ferramentas para aplicação dos métodos CDE.

Considere um conjunto de dados $W = \{X, S\}$ de tamanho N, onde X é um conjunto de P variáveis sensíveis ou confidenciais, e S é um conjunto de Q variáveis não confidenciais. Seja Y o conjunto de variáveis X modificadas e/ou perturbadas. Seja Z o conjunto de variáveis-chave que pertencem a S. As variáveis-chave são o conjunto de variáveis que podem indiretamente permitir a identificação de uma observação individual, por si só ou por referência cruzada com outras variáveis de bases de microdados externas, como as disponibilizadas *online*.

2. Risco de Identificação

Supondo que os utilizadores acedem a informação contida em S, o risco de identificação corresponde à probabilidade de um utilizador prever os valores de X por intermédio da função de densidade condicional $f(X|S)$. Quando o conjunto de variáveis modificadas Y é divulgado, se este não fornecer informação adicional sobre X, então o risco de identificação é ótimo, ou seja: $f(X|S, Y) = f(X|S)$. Neste contexto, é importante avaliar o risco de identificação do conjunto de dados original ($W = \{X, S\}$), bem como do conjunto de dados modificado ($W^* = \{Y, S\}$) para uma escolha adequada do método de CDE.

De seguida, apresentam-se algumas medidas para calcular o risco de identificação em variáveis categóricas e numéricas e mostra-se como calcular os riscos individual e global.

O risco individual é a probabilidade de identificação de uma observação individual, sendo o risco global a proporção de observações que podem ser identificadas por um utilizador.

Risco Individual e Risco Global

É possível calcular o risco individual de uma observação i através da frequência como $r_i = 1/F_k$, onde F_k representa a frequência populacional de uma combinação k de variáveis-chave, que corresponde ao número total de indivíduos que possuem a mesma combinação k de variáveis-chave na população[1]. O risco global pode ser calculado pela média aritmética dos riscos de todas as observações na população: $R = \frac{1}{N}\sum_{i=1}^{N} r_i$. Assim, o risco global pode ser visto como a proporção de todas as observações que podem ser identificadas por um utilizador. Os riscos individuais e globais permitem estimar a probabilidade de identificação, mas podem ser complementados com outras medidas de risco, conforme o tipo de variável.

Medidas de Risco para Variáveis Categóricas

O risco de identificação para variáveis categóricas baseia-se na contagem do número de observações que possuem combinações raras de variáveis chave. De seguida, apresentam-se algumas medidas para calcular o risco de identificação em variáveis categóricas.

***K-anonymity* e *L-diversity*.** Um conjunto de dados satisfaz a condição *K-anonymity* se cada combinação k de variáveis chave é comum a pelo menos K observações, ou seja, $F_k \geq K$, $\forall k \in \{1, ..., C\}$, em que C é o número total de combinações possíveis para as variáveis-chave escolhidas. Neste caso, o risco de identificação é medido pelo número de observações que não satisfazem a condição *K-anonymity* para um determinado valor K,

[1] Quando o conjunto de dados é uma amostra, F_k é estimado a partir da frequência amostral f_k.

$$\sum_{i=1}^{N} I_i(F_k < K), \text{ with } I_i = \begin{cases} 1 & \text{if } F_k < K \\ 0 & \text{if } F_k \geq K \end{cases}$$

onde N é o número total de observações. O valor de K pode ser qualquer número inteiro positivo e varia dependendo do nível de anonimização exigido. Por exemplo, uma observação não respeita o critério de 2-*anonymity* quando tem uma combinação única de variáveis-chave. Para 3-*anonymity* a condição não é respeitada se tivermos apenas 2 observações com a mesma combinação de variáveis chave, e assim por diante.

A principal desvantagem do método *K-anonymity* é não ser suficientemente rigoroso. Informações confidenciais podem ser reveladas mesmo se a condição *K-anonymity* for satisfeita. Isso pode acontecer em conjuntos de dados que contêm dados não identificáveis, mas com informação confidencial (Benschop et al., 2021). Para resolver esse problema, considera-se a definição de *L-diversity*.

Um grupo de observações que partilham a mesma combinação de variáveis-chave é considerado *L-diverse* se existirem pelo menos L valores distintos para a variável sensível, onde L varia entre 1 e o número de níveis da variável sensível. O nível pretendido de *L-diversity* depende do número de valores possíveis que a variável sensível pode assumir. Esta medida permite controlar a variabilidade de variáveis sensíveis dentro de cada combinação de variáveis-chave, de forma a garantir que os valores sensíveis não são revelados.

Medidas de Risco para Variáveis Numéricas

As medidas de risco para variáveis numéricas são geralmente calculadas depois da aplicação de um método de CDE e determinam a distância entre os dados modificados e os dados originais, em muito similares aos métodos de perda de informação. De seguida, apresentam-se as principais medidas de risco para variáveis numéricas.

Record Linkage. De acordo com Templ *et al.* (2014), o *Record Linkage* é um método que avalia o número correto de ligações entre os valores divulgados e os valores originais, a partir do cálculo das distribuições. É um processo bastante intensivo computacionalmente e funciona da seguinte maneira (Morais, 2022): considere-se x_{ip} a observação i da variável confidencial X_p da base de dados original e y_{ip} a mesma observação na base de dados modificada. Para cada y_{ip} calcula-se a distância relativamente a todas as observações de X_p; consideram-se apenas a primeira e a segunda observação mais próxima de y_{ip}, como x_{1p} e x_{2p}, respetivamente. Se x_{1p} ou x_{2p} for a observação original utilizada para gerar y_{ip} nos dados modificados, então diz-se que a observação y_{ip} está "ligada" à observação original. Repete-se o processo para todas as observações e por fim, o risco de identificação global é estabelecido como a percentagem de observações y_{ip} "ligadas" às observações originais x_{ip}, com $i = 1, ..., N$ e $p = 1, ..., P$.

Medida Intervalar. A aplicação desta medida envolve a criação de intervalos em torno de cada valor perturbado, verificando-se então se o valor original pertence ou não ao intervalo definido. Valores que estão dentro da faixa são considerados muito próximos e, portanto, são necessárias mais perturbações para garantir que a identificação não ocorra. Os intervalos são obtidos como $[y_{ip} - \sigma_p \cdot \alpha; y_{ip} + \sigma_p \cdot \alpha]$, onde y_{ip} é o valor perturbado da observação i da variável confidencial p, σ_p é o desvio padrão da variável confidencial p e α é um parâmetro escalar definido pelo responsável pelo conjunto de dados. O risco de identificação é calculado como a proporção dos valores originais que pertencem ao intervalo definido. Para a maioria das situações, é um método satisfatório. No entanto, este método não é tão eficaz na presença de *outliers*, uma vez que estas observações apresentam um risco maior de identificação (Benschop *et al.,* 2021).

Contagem de *Outliers*. As observações *outliers* são relevantes na medição do risco de identificação em variáveis numéricas, já que quando existem valores originais *outliers*, a aplicação de um método CDE também pode gerar um valor modificado *outlier* e portanto, com alto risco de identificação (Morais, 2022).

De acordo com Templ (2017), na prática, a identificação das observações *outliers* é realizada a partir da identificação de valores que são inferiores ou superiores a um determinado percentil. A contagem de *outliers* estima o risco de identificação a partir de três passos: **1)** calcular a distância de *Mahalanobis* robusta entre indivíduos, obtendo-se uma distância multivariada para cada observação *i*; **2)** estimar os intervalos para cada observação *i* dos dados originais. A amplitude do intervalo depende das distâncias calculadas em 1) e de um parâmetro escalar: quanto maior for a distância de *Mahalanobis* robusta, maior será a amplitude dos intervalos; **3)** verificar se os valores perturbados pertencem aos intervalos calculados em 2) e se o valor pertencer ao intervalo, então, a observação está em risco de identificação.

3. Utilidade *vs* Perda de Informação

Após a aplicação de um método CDE, recomenda-se a avaliação da segurança e utilidade dos dados. A relação entre o risco de identificação e a utilidade dos dados é caracterizado por dois extremos: risco zero de identificação, não há divulgação de informação individual, mas perde-se a informação, sendo os dados inúteis para os utilizadores; risco máximo de identificação, os dados são divulgados sem qualquer modificação e a perda de informação é nula.

Idealmente, as características dos dados divulgados deveriam ser idênticas às dos dados originais, e qualquer análise realizada nos dados modificados deveria conduzir o utilizador a resultados semelhantes aos obtidos nos dados originais (Rao *et al*., 2012). Portanto, a utilidade ideal define-se como $f(Y; S) = f(X; S)$, onde $f(\ ;\ )$ é a função de densidade conjunta e X, Y e S são as variáveis sensíveis, modificadas e confidenciais, respetivamente.

Medidas de Utilidade

De seguida, apresentam-se algumas medidas úteis para avaliar a perda de informação no processo de CDE.

Contagem valores em falta. Alguns métodos de CDE substituem os valores originais por valores em falta, o que pode levar a uma redução significativa na utilidade dos dados modificados (Templ, 2017). Assim, uma medida importante de perda de informação é a comparação do número total de valores em falta entre as variáveis originais *X* e as variáveis modificadas *Y*.

Esta medida pode ser descrita como:

$$M = \sum_{i=1}^{N} \sum_{p=1}^{P} r_{ip},$$

em que N e P correspondem ao número de observações e ao número de variáveis sensíveis ou confidenciais X, respetivamente; $r_{ip} = 1$ quando a observação i da variável p está em falta nos dados modificados, mas não nos dados originais, caso contrário, $r_{ip} = 0$. Como estamos a avaliar os valores em falta, quanto maior o valor da medida M, maior será a perda de informação nos dados modificados (Templ, 2017).

***Information Loss* 1 (*IL*1).** A medida *Infomation Loss* 1 (IL1) calcula a variação média entre os dados originais e os dados modificados (Benschop *et al*., 2021):

$$IL1 = \frac{100}{PN} \sum_{i=1}^{N} \sum_{p=1}^{P} \frac{|x_{ip} - y_{ip}|}{|x_{ip}|},$$

Em que x_{ip} e y_{ip} são as observações i da variável p para as variáveis original e modificada, respetivamente. Se a observação x_{ip} for nula, é substituída no denominador da fração pelo valor modificado y_{ip}. No caso de ambos os valores serem nulos, a p-ésima variável é excluída do cálculo para a i-ésima observação. Em $IL1$, o efeito da variação absoluta depende da distância entre x_{ip} e 0. A medida $IL1$ é maior para variáveis próximas de 0. Para contornar esse problema, foi criada a medida *Information Loss* 1s (IL1s) que pode ser interpretada como a distância escalar entre os dados originais e os dados modificados (Templ, 2017), que se define da seguinte forma:

$$IL1s = \frac{1}{PN} \sum_{p=1}^{P} \sum_{i=1}^{N} \frac{|x_{ip} - y_{ip}|}{\sqrt{2} S_p},$$

Em que S_p é o desvio padrão da p-ésima variável nos dados originais. Quanto menor for o valor desta distância, maior será a utilidade dos dados no processo de perturbação.

Outras Medidas. Outra possibilidade para avaliar a utilidade dos dados é através da comparação dos valores próprios entre as variáveis originais e as variáveis modificadas (Benschop *et al*., 2021). Nesta medida, são calculadas as diferenças médias absolutas entre os valores próprios das matrizes de covariância das variáveis originais e modificadas (Torra e Domingo-Ferrer, 2001).

O erro quadrático médio (EQM) e o erro médio absoluto (EMA) são também duas medidas que são frequentemente utilizadas para avaliar as diferenças entre os dados originais e os dados modificados.

4. Métodos Perturbativos e Não Perturbativos

Os métodos de Controlo de Divulgação Estatística (CDE) têm como objetivo principal modificar os dados estatísticos para que possam ser divulgados sem relevar informações que permitam a identificação de uma observação a partir da base de dados perturbada. A maior dificuldade neste processo é alcançar essa modificação de forma eficaz, ou seja, publicar uma base de dados segura do ponto de vista da identificação e útil do ponto de vista da análise por parte dos utilizadores.

Após a identificação das variáveis chave e sensíveis, os métodos de CDE devem ser aplicados às variáveis que possam conduzir à identificação de uma observação ou mais observações ou que contenham informações confidenciais.

Existem três tipos de métodos CDE:

• **Métodos Não Perturbativos:** estes métodos não modificam os dados, mas realizam supressões ou reduções parciais, como por exemplo os métodos de Recodificação e Supressão Local;

• **Métodos Perturbativos:** estes métodos alteram os valores das variáveis sujeitas a perturbação, por exemplo, Método de Pós-Randomização, *Shuffling*, Adição de Ruído e Micro-agregação;

• **Geração de Dados Sintéticos:** são técnicas que geram dados sintéticos, resultando geralmente numa base de dados de menor dimensão, preservando algumas estatísticas ou relações entre variáveis presentes na base de dados original. Esta técnica geralmente reduz o risco de identificação e aumenta a utilidade dos dados.

Para além desta classificação, também é possível dividir os métodos em probabilísticos e não probabilísticos:

• **Métodos Probabilísticos:** baseiam-se em mecanismos probabilísticos ou em mecanismos geradores de números aleatórios;

• **Métodos Não Probabilísticos:** baseiam-se num certo algoritmo e produzem os mesmos resultados se aplicados repetidamente na mesma base de dados.

Estes métodos diferem consoante a natureza das variáveis.

Métodos Não Perturbativos

Na Tabela 1 estão apresentados alguns dos principais métodos não perturbativos de CDE, tipo de aplicação e a sua classificação como probabilístico ou não probabilístico.

Tabela 1: Métodos Não Perturbativos

Método	Variáveis Categóricas	Variáveis Contínuas	Probabilístico	Não Probabilístico
Recodificação	X	X		X
Supressão Local	X			X

Recodificação. Recodificação é um método determinístico que pode ser aplicado tanto a variáveis contínuas como a variáveis categóricas. Utiliza-se este método para reduzir o número de categorias distintas de uma variável categórica ou o número de valores possíveis de uma variável contínua. Para variáveis categóricas, recodificação combina várias categorias, aumentando assim as frequências de cada combinação de variáveis chave enquanto diminui o detalhe nos dados. Para variáveis numéricas, este método corresponde a discretizar a variável, ou seja, transformar a variável contínua numa variável categórica, onde as categorias correspondem a intervalos de valores possíveis para a variável em estudo (Templ, 2017). Recodificação é aplicado a todas as observações da base de dados e não apenas às observações em risco de identificação (Benschop, 2021).

Supressão Local. Quando existem combinações únicas de variáveis-chave no conjunto de dados, ou seja, $F_k = 1$, recomenda-se aplicar este método, garantindo que todas as observações cumpram uma dada condição de *K-anonimity*. Uma aplicação deste método corresponde a substituir as combinações únicas de variáveis-chave por valores em falta (Templ, 2017). Esta técnica também pode ser aplicada apenas às observações com risco individual superior a um limite estabelecido. Desta forma, reduz-se o risco de identificação sem perder muita informação durante o processo.

Métodos Perturbativos

Na Tabela 2 estão representados os principais métodos perturbativos de CDE, tipo de aplicação e a sua classificação como probabilístico ou não probabilístico.

Tabela 2: Métodos Perturbativos

Método	Variáveis Categóricas	Variáveis Contínuas	Probabilístico	Não Probabilístico
Modelos lineares		X	X	
Modelos não lineares		X	X	
Microagregação		X	X	
PRAM	X		X	
Generalização	X		X	
Data Shuffling		X	X	

Variáveis Categóricas

Método de Pós-Randomização (PRAM). Este método, geralmente aplicado a variáveis categóricas, também se pode aplicar a variáveis numéricas quando convertidas em categóricas. É normalmente utilizado para variáveis não confidenciais e variáveis não-chave. É o método mais eficaz para a proteção dos dados no que diz respeito à utilidade dos dados perturbados. O método consiste na troca das categorias de uma variável categórica com base numa matriz de transição pré-definida, *T*, que especifica a probabilidade de uma categoria ser substituída por outra na mesma variável. Os valores diagonais da matriz mostram a probabilidade de cada categoria permanecer inalterada. O PRAM protege os dados originais aplicando uma perturbação; no entanto, como o mecanismo de probabilidade (matriz *T*) é conhecido, as características dos dados originais podem ser estimadas a partir dos dados perturbados (Templ, 2017).

Generalização. Este método é aplicado a variáveis categóricas para combinar várias categorias em novas categorias, mantendo a estrutura hierárquica da variável. Consiste em substituir os valores de variáveis chave por valores de outra observação, minimizando assim a perda de informação. Em cada interação, dois critérios são usados: selecionar combinações de variáveis chave com menores frequências e encontrar a combinação mais semelhante para reduzir a perda de informação. O algoritmo repete a substituição até atingir a condição de *K-anonymity*.

Os diversos testes realizados com este método perturbativo indicam que as variáveis perturbadas apresentam maior similaridade do que quando se aplica o método PRAM (Navarro-Arribas e Torra, 2015).

Variáveis Numéricas

Os métodos perturbativos aplicados a variáveis numéricas podem ser baseados em Modelos Lineares ou Modelos Não Lineares. Os métodos baseados em Modelos Lineares incluem o Modelo de Ruído Aditivo

Independente, o Modelo de Ruído Aditivo Correlacionado e o Modelo Aditivo Geral Exato para Perturbação de Dados (EGADP). Por outro lado, os métodos baseados em Modelos Não Lineares incluem o Modelo de Ruído Multiplicativo, *Data Shuffling*, e Microagregação.

Métodos de Perturbação baseados em Modelos Lineares

A perturbação de um conjunto de dados através de Modelos Lineares é geralmente baseada na adição ou subtração de ruído aleatório aos valores originais. Estes métodos protegem a correspondência exata dos dados com possíveis ficheiros externos.

Modelo de Ruído Aditivo Independente. Este método pode ser descrito da seguinte forma (Rao et al., 2012):

Seja $Y_p = X_p + \epsilon_p$, onde X_p e Y_p são os vetores de observações da variável p da base de dados original e perturbada, respetivamente; $\epsilon_p \sim N(0, \sigma)$ e $\sigma = \alpha * \sigma_p$, sendo σ_p o desvio padrão da variável original X_p e α é uma constante positiva. A aplicação deste modelo preserva a média e a variância de cada variável, no entanto, as covariâncias e correlações entre variáveis são alteradas (Templ, 2017).

Modelo de Ruído Aditivo Correlacionado. Este modelo apresenta vantagens em relação ao modelo independente porque, neste caso, a matriz de covariância do ruído é proporcional à matriz de covariância original, ou seja, $\epsilon_p \sim N(0,\Sigma_\epsilon = c\Sigma)$, onde Σ é a matriz de covariância de X_p e c é uma constante positiva. Assim, as covariâncias e correlações entre variáveis não se alteram (Benschop, 2021). Este modelo apresenta um risco de identificação inferior ao modelo de Ruído Aditivo Independente (Rao et al., 2012).

Modelo Aditivo Geral Exato para Perturbação de Dados (EGADP). Para obter resultados mais eficazes do ponto de vista do risco de identificação, foi sugerido o modelo EGADP (Rao et al., 2012). Esse resultado é alcançado se garantirmos que os termos de ruído gerados são ortogonais às variáveis confidenciais originais. Assim, o EGADP proporciona uma utilidade dos dados superior a todos os outros modelos de perturbação linear e um risco de identificação que é o mais baixo entre todos os modelos de perturbação linear.

Métodos de Perturbação baseados em Modelos Não Lineares

Os métodos de perturbação baseados em modelos não lineares são métodos capazes de incorporar modelos multiplicativos ou de combinar mais do que um processo de perturbação.

Modelo de Ruído Multiplicativo. Uma das maiores desvantagens dos modelos de ruído aditivo é que, para valores próximos de zero, causam grande perturbação e, para valores mais afastados de zero, causam menor perturbação. Para resolver este problema, foi criado o modelo de ruído multiplicativo.

Seja Z a matriz de perturbação, com uma média igual a 1 e desvio padrão $\sigma_Z > 0$. A matriz Y dos dados perturbados é obtida como $Y = X \bullet Z$, onde • se refere ao produto de *Hadamard*, definido como $y_{ip} = x_{ip} \cdot z_{ip}$, com $i = 1, ..., N$, $p = 1, ..., P$ (Benschop, 2021).

***Data Shuffling*.** Neste método, os valores das variáveis confidenciais são baralhados entre as variáveis. Este procedimento é bastante eficaz, pois resulta em riscos de identificação baixos e fornece a máxima utilidade na informação da base de microdados perturbada, visto que preserva todas as relações existentes entre as variáveis originais (Muralidhar e Sarathy, 2006).

Microagregação. Este método começa por dividir as observações em partições e depois calcula uma medida estatística de cada partição (normalmente a média), e cada valor é substituído pela medida estatística calculada da partição à qual pertence (Templ, 2017). A eficácia deste método depende da escolha das partições, ou seja, obtêm-se resultados eficazes quando as observações agrupadas são homogéneas, já que a medida estatística calculada não difere significativamente do valor original pelo qual será substituída. Este processo apresenta várias alternativas sobre como construir as partições e substituir os valores, tais como: Classificação Individual, Microagregação baseada na Análise de Componentes Principais (PCA), Distância Máxima ao Valor Médio (MDMV) ou Distância de *Mahalanobis*.

Comparação dos Métodos

Os métodos não perturbativos aplicam-se essencialmente a variáveis chave, geralmente categóricas, para garantir uma frequência mínima das suas combinações. No entanto, estes métodos podem não garantir, por si só, a anonimização[2] dos dados, sendo muitas vezes necessário recorrer a outras metodologias como é o caso dos métodos perturbativos.

Entre os modelos de perturbação baseados em modelos lineares, o EGADP proporciona o risco de identificação mais baixo. Usando este método, todas as inferências baseadas em modelos lineares são mantidas exatamente nos dados perturbados. No entanto, tem a desvantagem de que as distribuições marginais das variáveis perturbadas podem ser diferentes das variáveis originais (Rao et al., 2012).

Quanto aos modelos de perturbação baseados em modelos não lineares, o *Data Shuffling* oferece a vantagem adicional de que os valores confidenciais originais não são modificados. Assim, para cada variável, a distribuição marginal dos dados perturbados é idêntica à distribuição original. No que diz respeito ao equilíbrio entre risco de identificação e a utilidade dos dados, o *Data Shuffling* é o mais eficaz entre todos os modelos não lineares, conforme indicado em Rao et al., (2012).

[2] Processo no qual as Informações Pessoais Identificáveis (IPI) são modificadas de forma irreversível fazendo com que uma entidade não consiga ser identificada tanto de forma direta quanto indireta, pelo responsável pelo tratamento de IPI por si só ou em colaboração com qualquer outra pessoa (ISO, 2011).

A Microagregação apresenta um desempenho inferior quando comparada tanto com o EGADP ou com o *Data Shuffling*, uma vez que aumenta o risco de divulgação. Em termos de utilidade dos dados, não preserva a validade inferencial como o EGADP, nem preserva as distribuições marginais e as relações monótonas como o *Data Shuffling* (Rao et al., 2012).

Após comparar os métodos mais importantes, verifica-se que a escolha recai principalmente nos modelos *Data Shuffling* e EGADP, dependendo do tipo de base de dados. Em geral, é aconselhável aplicar o *Data Shuffling* quando o conjunto de dados é relativamente grande e contém relações não lineares importantes. Por outro lado, o EGADP é recomendado para conjuntos de dados de menor dimensão (Rao et al., 2012). Contudo, Morais (2022) e Morais *et al*. (2024) mostram que em determinadas aplicações os modelos com adição de ruído podem superar os anteriores em termos de resultados.

A escolha do método de SDC a utilizar pode depender das distribuições dos dados e os resultados mais eficazes podem ser obtidos usando outros métodos. Na prática, a escolha do melhor método depende se o objetivo é minimizar o risco de identificação ou maximizar a utilidade no conjunto de dados perturbados. No final, os dados perturbados podem ser o resultado da aplicação de um ou mais métodos de SDC (Templ, 2017).

5. Ferramentas para CDE

Existem várias ferramentas ou linguagens disponíveis para a aplicação dos métodos CDE, por exemplo, µ-Argus, C^{++} e R.

µ-argus

Através de um programa financiado pela União Europeia, µ-argus foi desenvolvido com o objetivo de perturbar bases de microdados. Em 1995, foi lançada a primeira versão da ferramenta pelo Departamento de Métodos Estatísticos na Holanda. Atualmente, continua a ser aprimorado por várias instituições estatísticas. Originalmente desenvolvido em Visual Basic até à versão 4.2, agora é escrito em Java e está disponível gratuitamente no site do CASC (https://research.cbs.nl/casc/mu.htm).

Linguagem C^{++}

Estudos realizados pelo *International Household Survey Network* (IHSN) resultaram no desenvolvimento de uma linguagem de programação para a perturbação de microdados. O IHSN desenvolveu um código em C^{++} que possibilita a perturbação e anonimização de bases de microdados com o objetivo de apoiar a divulgação segura de dados confidenciais. Este código desenvolvido por IHSN está disponível gratuitamente para todos os utilizadores.

Linguagem R

As opções anteriormente apresentadas são mais limitadas em comparação com o *package sdcMicro* do ambiente R (Templ *et al*., 2015). O *package sdcMicro* não só permite a aplicação de um maior número de métodos, como também fornece uma linguagem simples e eficaz para um utilizador aplicar os diferentes métodos. A primeira versão do *package sdcMicro* incluía apenas alguns métodos de CDE. Ao longo dos anos, este *package* foi sendo atualizado e atualmente disponibiliza a maioria dos métodos de CDE, tornando-se uma das ferramentas mais completas para proteção de bases de dados. Este *package* inclui e mantém atualizado o código desenvolvido pelo IHSN em C++.

Além disso, o ambiente R fornece ainda o *package sdcTable* que contém diversas funções para a perturbação de dados tabulares (Meindl, 2023).

6. Conclusão

A implementação de métodos de Controlo de Divulgação Estatística requer uma abordagem equilibrada entre a proteção da privacidade dos dados e a preservação da sua utilidade. A utilização de medidas adequadas para avaliar a perda de informação e o risco de identificação é essencial para garantir a eficácia destes métodos e a segurança dos dados divulgados. O equilíbrio entre o risco de identificação e a utilidade dos dados é essencial para garantir que os dados modificados sejam úteis para o utilizador sem comprometer a sua confidencialidade.

Cada método possui as suas vantagens e desvantagens, e a escolha do mais adequado dependerá das necessidades específicas do contexto de divulgação estatística e da importância atribuída à proteção da privacidade e à preservação da utilidade dos dados.

Referências

Aggarwal, C., Yu, P. S., 2008. Privacy Preserving Data Mining: Models and Algorithms. Springer New York, NY. DOI: https://doi.org/10.1007/978-0-387-70992-5

Benschop, T., Machingauta, C., Welch, M., 2021. Statistical Disclosure Control: A Practice Guide. In: The World Bank.

ISO - International Organization for Standardization, 2011. ISO/IEC 29100:2011. https://www.iso.org/standard/45123.html.

Meindl, B., 2023. sdcTable: Methods for Statistical Disclosure Control in Tabular Data. R package version 0.32.6, https://github.com/sdcTools/sdcTable.

Morais, J., 2022. Comparação de Métodos Perturbativos: Utilidade e perda de informação em bases de microdados. Dissertação de Mestrado, Universidade do Minho.

Morais, J., Sousa, R., Faria, S., 2024. Perturbation Methods: an application to real dataset. High-quality and Timely Statistics: New Methods and Applications. Conference Proceedings, Springer.

Muralidhar, K., Sarathy, R., 2006. Data shuffling - A new masking approach for numerical data. Management Science, 52(5), pp. 658-670.

Navarro-Arribas G., Torra V., 2015. Advanced Research in Data Privacy. Vol. 1. Springer.

Rao, C.R., Chakraborty, R., Sen, P. K., 2012. Handbook of Statistics, vol. 28: Bioiformatics in Human Health and Heredity. 1st. North Holland IFIP, pp. 511-533.

Templ, M., 2017. Statistical Disclosure Control for Microdata: Methods and Applications in R. Vol. 1. Springer International Publishing.

Templ, M., Kowarik, A., Meindl, B., 2015. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. Journal of Statistical Software, 67 (4), pp. 1–36. doi:10.18637/jss.v067.i04

Templ, M., Meindl, B., Kowarik, A., Chen, S., 2014. Introduction to Statistical Disclosure Control (SDC). Project: Relative to the testing of SDC algorithms and provision of practical SDC, data analysis OG.

Tendick, P., Matloff, N., 1994. "A modified random perturbation method for database security". In: ACM Transactions Database System 19, pp. 47-63.

Torra, V., Domingo-Ferrer, J., 2001. Disclosure Protection Methods and Information Loss for Microdata. In P. Doyle, J. Lane, J. Theeuwes, & Z. L., Theory and Practical Applications for Statistical, pp. 91-110.

Aplicação de Replicação

Gustavo Iglesias, giglesias@bportugal.pt
BPLIM, Departamento de Estudos Económicos, Banco de Portugal

Paulo Guimarães, pfguimaraes@bportugal.pt
BPLIM, Departamento de Estudos Económicos, Banco de Portugal
Faculdade de Economia do Porto

Introdução

A reproducibilidade é um pilar fundamental da ciência. Refere-se à capacidade de obter resultados consistentes ao repetir experiências ou estudos seguindo os mesmos métodos e condições originais. Esta característica é crucial para validar descobertas científicas e garantir que as mesmas não são produtos de acaso ou de erro experimental.

A reproducibilidade fortalece a confiança nas conclusões científicas e permite que outros investigadores criem estudos baseados em trabalhos anteriores, promovendo o avanço do conhecimento. No entanto, a comunidade científica enfrenta desafios significantes neste âmbito. Problemas como a falta de acesso aos dados utilizados contribuem para a não reproducibilidade da investigação.

Por forma a melhorar a reproducibilidade na investigação, diversas iniciativas têm sido implementadas. O acesso aberto a dados e código, a transparência nos métodos e a revisão por pares são algumas das práticas recomendadas. Atualmente, muitas das revistas científicas mais conceituadas requerem o acesso ao código e dados - um pacote de replicação - para aceitarem uma publicação.

A utilização de dados confidenciais para fins de investigação coloca desafios únicos à reproducibilidade devido à necessidade de proteção da privacidade e confidencialidade dos indivíduos ou entidades envolvidas. O acesso a estes dados é naturalmente condicionado, o que dificulta a verificação de reproducibilidade. Neste sentido, é importante que exista transparência sobre o modo como os centros de investigação lidam com microdados confidenciais. Só assim será possível credibilizar perante terceiros, por exemplo, editores de revistas e outros investigadores, a investigação levada a cabo com microdados confidenciais.

O Laboratório de Microdados do Banco de Portugal (BPLIM)

Dada a importância crescente da utilização dos microdados para fins científicos, o Banco de Portugal criou em 2016 o Laboratório de Investigação em Microdados do Banco de Portugal (BPLIM) com o

objetivo de permitir que investigadores externos acedam aos microdados do banco (para mais informação sobre as condições de acesso ver o site do BPLIM. Como algumas dessas bases de dados são confidenciais o Banco elaborou um conjunto de orientações sobre o acesso e manipulação de microdados confidenciais. Essas orientações estipulam que o acesso apenas é permitido a investigadores devidamente credenciados e para projetos que tenham mérito científico. Nesse caso, o Banco terá a responsabilidade de disponibilizar o acesso aos microdados na infraestrutura do Banco com o *software* adequado para investigação, dotado de salvaguardas que impeçam o seu uso abusivo.

Tratando-se de bases de microdados em que existam preocupações de confidencialidade o Banco entendeu que se deveria disponibilizar aos investigadores externos uma base de dados "modificados" que apenas serviria o propósito de criação do código de análise dos dados. Caberia ao BPLIM executar esse código (scripts) sobre os dados originais e disponibilizar, após controlo de confidencialidade, os resultados aos investigadores externos. Na prática, trata-se de um exercício de replicabilidade que o BPLIM tem de efetuar, utilizando o código desenvolvido pelos investigadores externos e os microdados confidenciais. Para que este processo funcione bem torna-se fundamental que os investigadores adotem as melhores práticas de reprodutibilidade. Para ajudar ao processo, o BPLIM desenvolveu uma aplicação, a **Aplicação de Replicação**. Embora esta aplicação seja de uso obrigatório sempre que o investigador trabalha com dados "modificados" ela pode também ser útil para investigadores que não trabalham com dados modificados pois cria de forma automática um pacote de replicação. De seguida, detalhamos o processo de replicação de microdados confidenciais no BPLIM, dando particular enfoque ao uso da **Aplicação de Replicação**.

Processo de Replicação

Os dados “modificados” servem para os investigadores externos testarem e prepararem os seus scripts. Uma vez concluída a análise, deverão solicitar à equipa do BPLIM que replique a análise nos dados originais. Este processo de replicação tem duas fases. A primeira fase é da responsabilidade do investigador, que deve confirmar se o processo corre perfeitamente numa única execução. Idealmente, isso deve ser feito usando a **Aplicação de Replicação** do BPLIM. Se esta fase for bem sucedida, os investigadores deverão comunicar este resultado ao BPLIM por e-mail e solicitar que o código seja executado nos dados originais. Depois de verificar que o código foi executado com sucesso e que nenhum ficheiro foi alterado, a equipa do BPLIM executará os scripts do investigador nos dados originais.

Aplicação de Replicação

Dentro da área de trabalho do projeto de cada investigador existe um ficheiro “.desktop”, que lança a aplicação desenvolvida pela equipa do BPLIM para fins de replicação. Para ilustrar como a aplicação

funciona, usaremos um projeto simulado baseado em Stata, **pxxx_BPLIM**[1]. A estrutura de ficheiros dos projetos é a seguinte:

```
/bplimext/projects/pxxx_BPLIM/
│
├── initial_dataset/
│   ├── modified/
│   │   ├── PM110_BBS_P_MBNK_JAN2000DEC2020_JUN21_ASSET_V01.dta
│   │   └── ...
│   ├── intermediate/
│   │   └── ...
│   └── ...
│
├── work_area/
│   ├── ReplicationApp.desktop
│   └── Submissions/
│       ├──  master.do
│       └──  do_files/
│           ├──  data_creation.do
│           ├──  dissertation.do
│           ├──  dissertation_interest.do
│           ├──  dissertation_extra.do
│           ├──  dissertation_means.do
│           └──  dissertation_until2017.do
│
├── tools/
│
└── results/
```

O diretório initial_dataset é onde a equipa do BPLIM coloca os dados para os investigadores, que possuem apenas permissão de leitura. Dentro desta pasta, podem existir dois diretórios, modified e intermediate, o primeiro para dados modificados[2] e o último para dados intermédios criados pelo investigador[3].

Se o investigador clicar no ficheiro ReplicationApp.desktop, a aplicação é iniciada, exibindo uma caixa de diálogo (Cf. Figura 1).

O primeiro passo é selecionar o caminho principal para a replicação, ou seja, o diretório onde o investigador colocou o código para executar a análise. Este diretório pode incluir outros diretórios e subdiretórios. Note-se que cada diretório e subdiretório no caminho selecionado será copiado para a área de replicação. Se clicarmos no botão *Browse*, uma caixa de diálogo será aberta para permitir ao utilizador selecionar o caminho (Cf. Figura 2).

---

[1] Embora este seja um projeto simulado, na verdade estamos a usar dados e código real de uma aluna que usou dados BPLIM na sua dissertação de mestrado. Neste projeto a aluna trabalhou apenas com dados modificados.

[2] Para uma descrição mais detalhada sobre como trabalhar com dados modificados no BPLIM, consulte o Guia para Investigadores.

[3] Para processos demasiado longos, cuja função principal é criar dados intermédios, é possível aos investigadores solicitarem a movimentação desses ficheiros de dados para a sub-pasta intermediate.

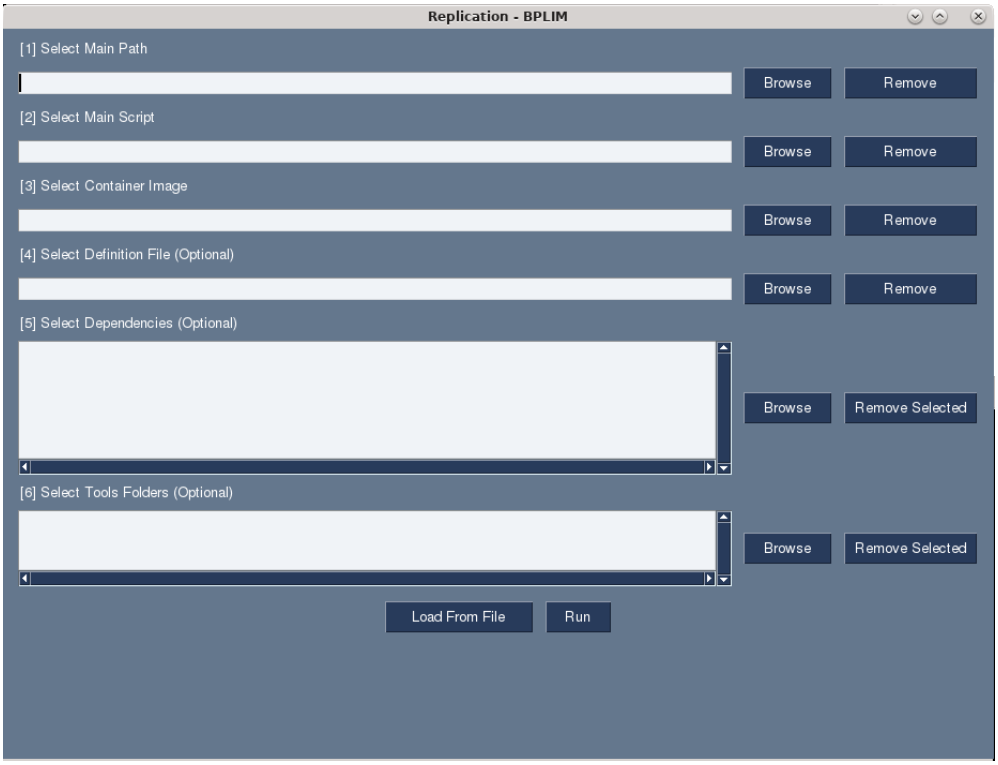


Figura 1

No nosso exemplo, o caminho é "/bplimext/projects/pxxx_BPLIM/work_area/Submissions":

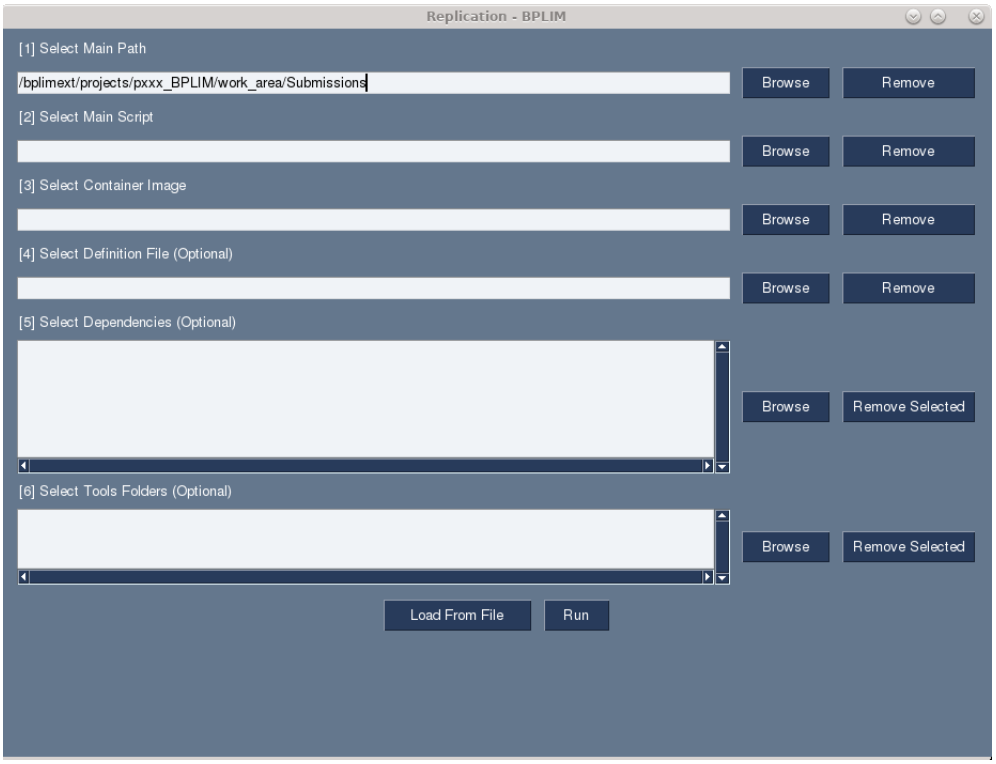
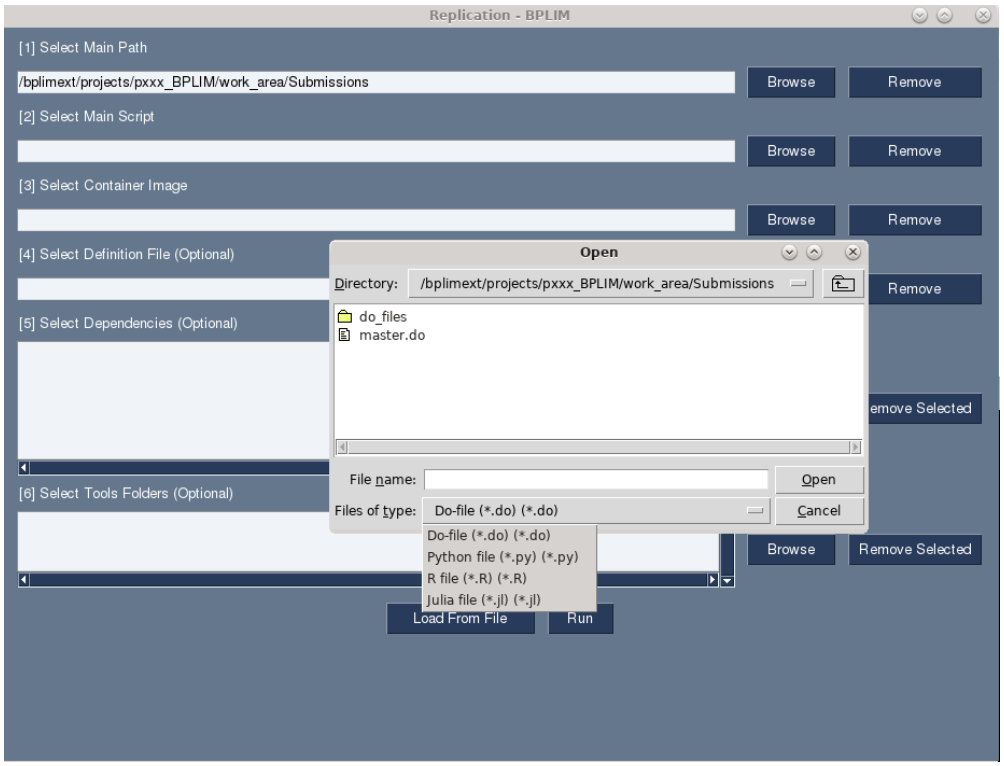
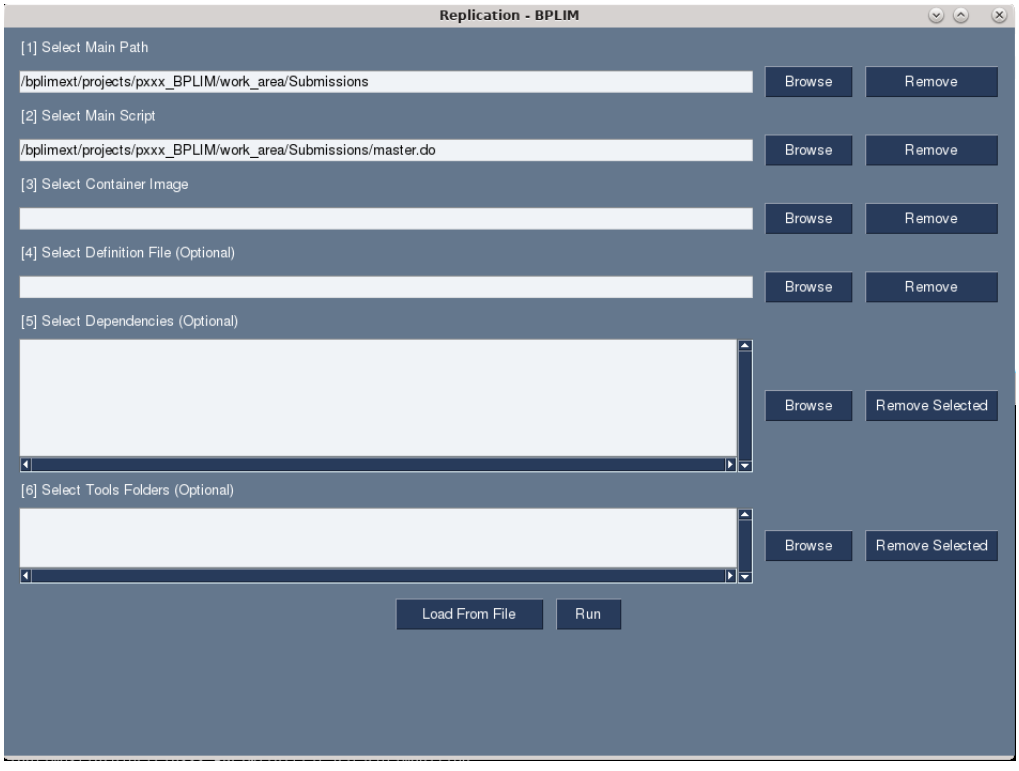


Figura 2

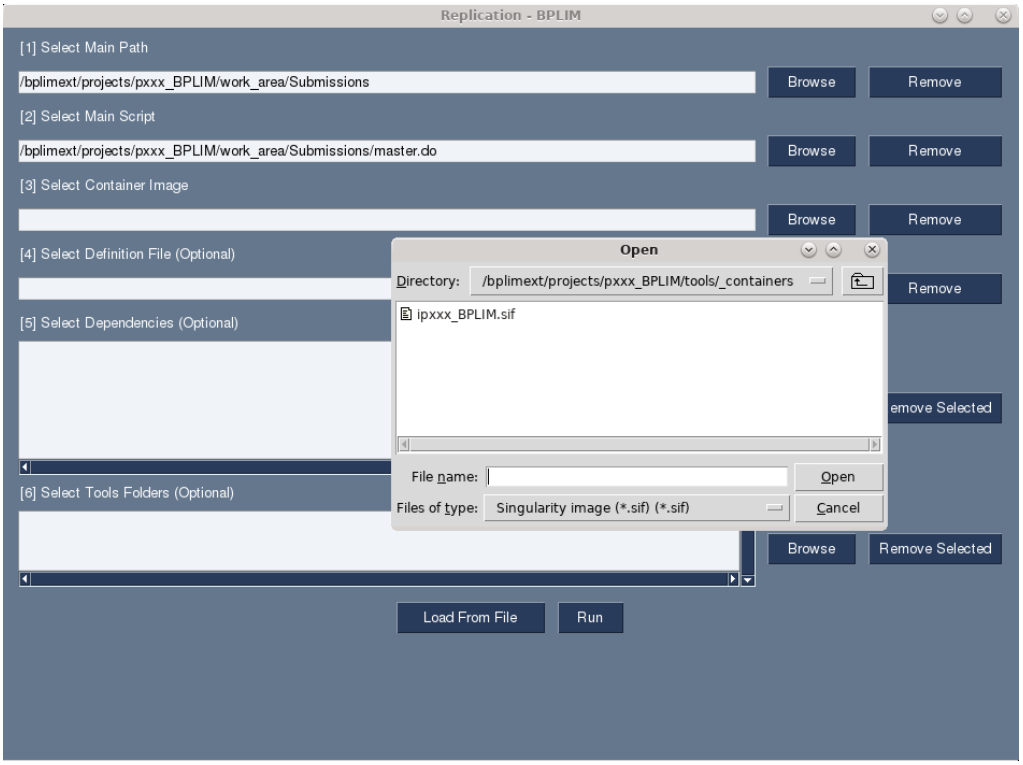
Após selecionar o caminho principal, temos de selecionar o script principal. Este é o ponto de entrada da replicação. É o ficheiro que executa a replicação e chama as dependências caso estas existam. O ficheiro deve estar no caminho principal (não necessariamente no primeiro nível) selecionado na etapa

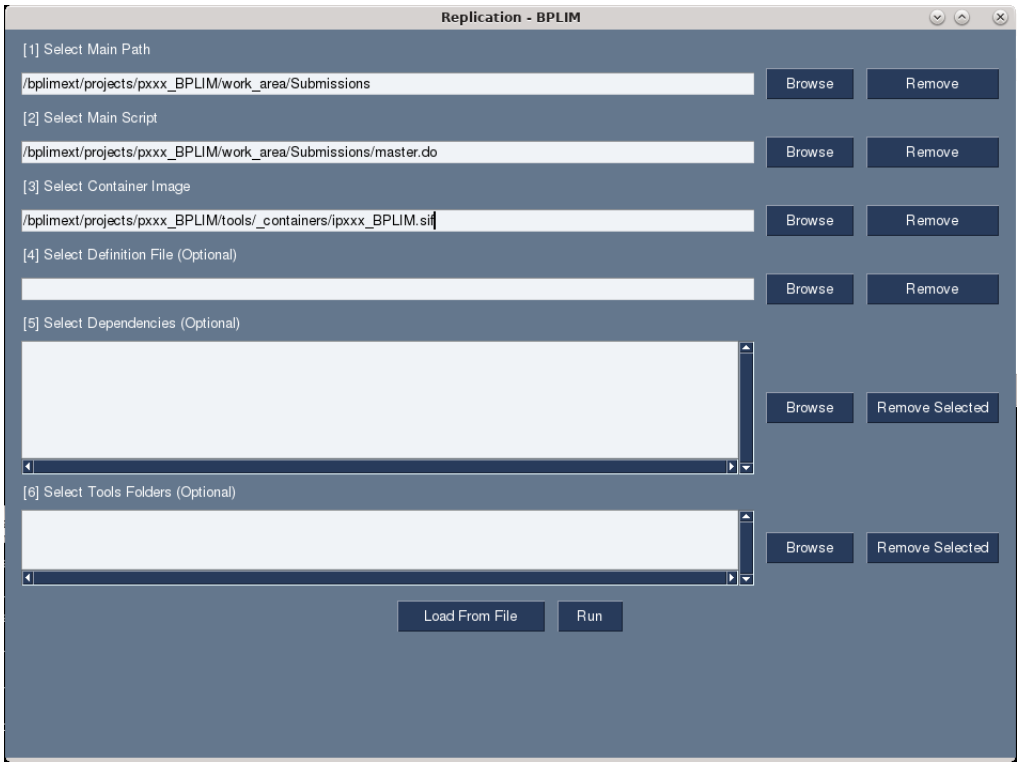
anterior, caso contrário a replicação não será executada. Somente extensões de ficheiros Stata, Python, R e Julia são permitidas neste campo:



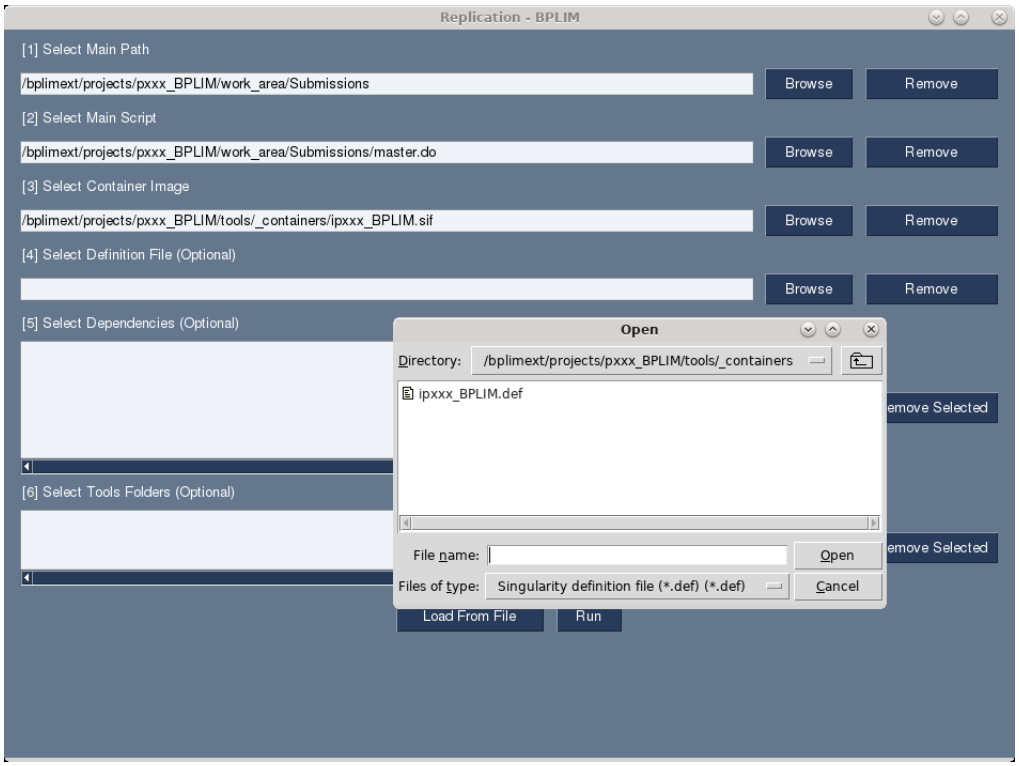


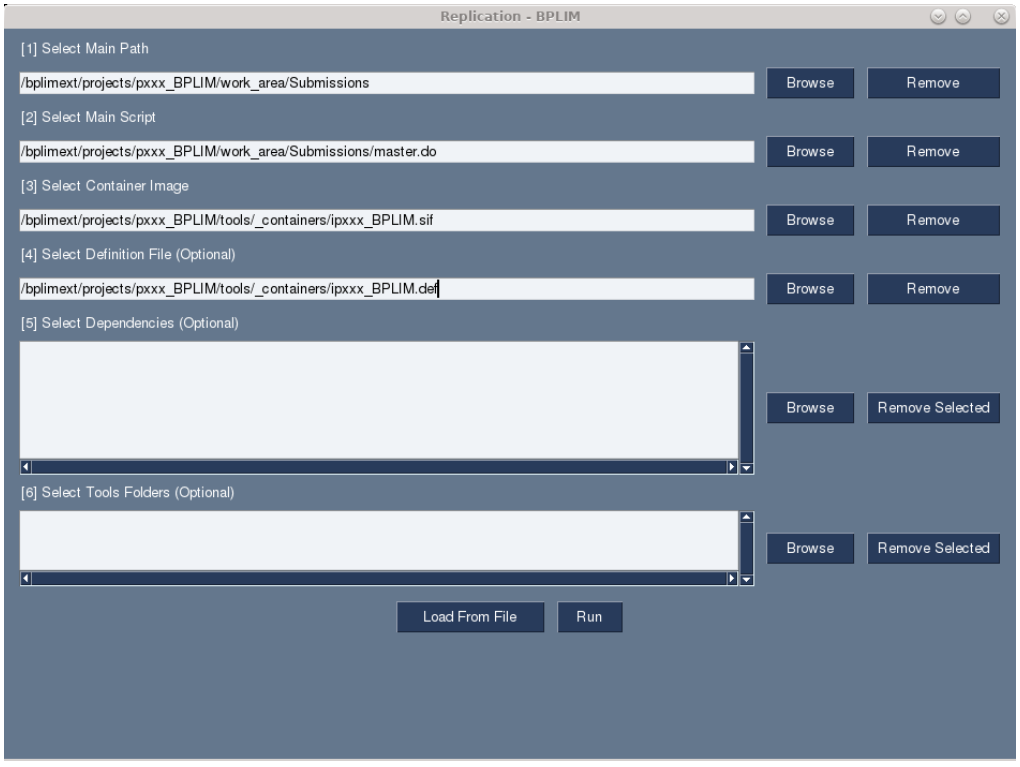
No terceiro campo devemos selecionar a imagem do container utilizada para executar a replicação. Os containers são uma ótima maneira de controlar o ambiente computacional por forma a garantir (não completamente) que a investigação é reprodutível. Para mais informações sobre containers e sobre a sua utilização no BPLIM, siga este link. No contexto da aplicação, basta o utilizador selecionar a Imagem Singularity (ficheiro *.sif), colocada pela equipa do BPLIM em *"…/tools/_containers/"*:



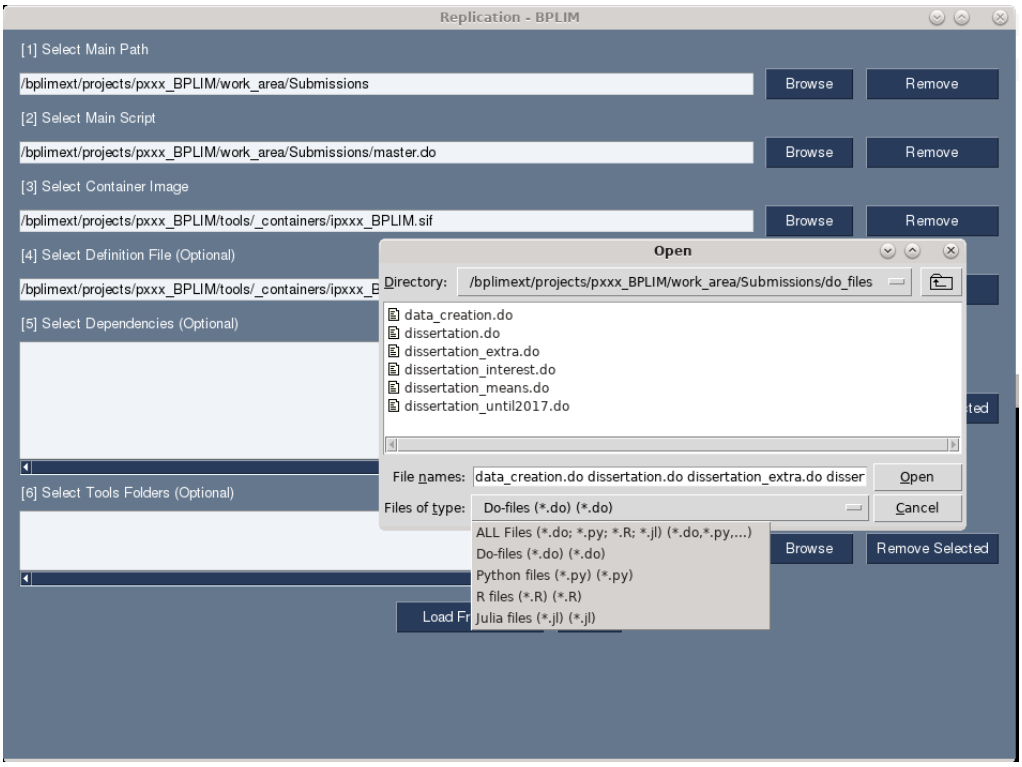


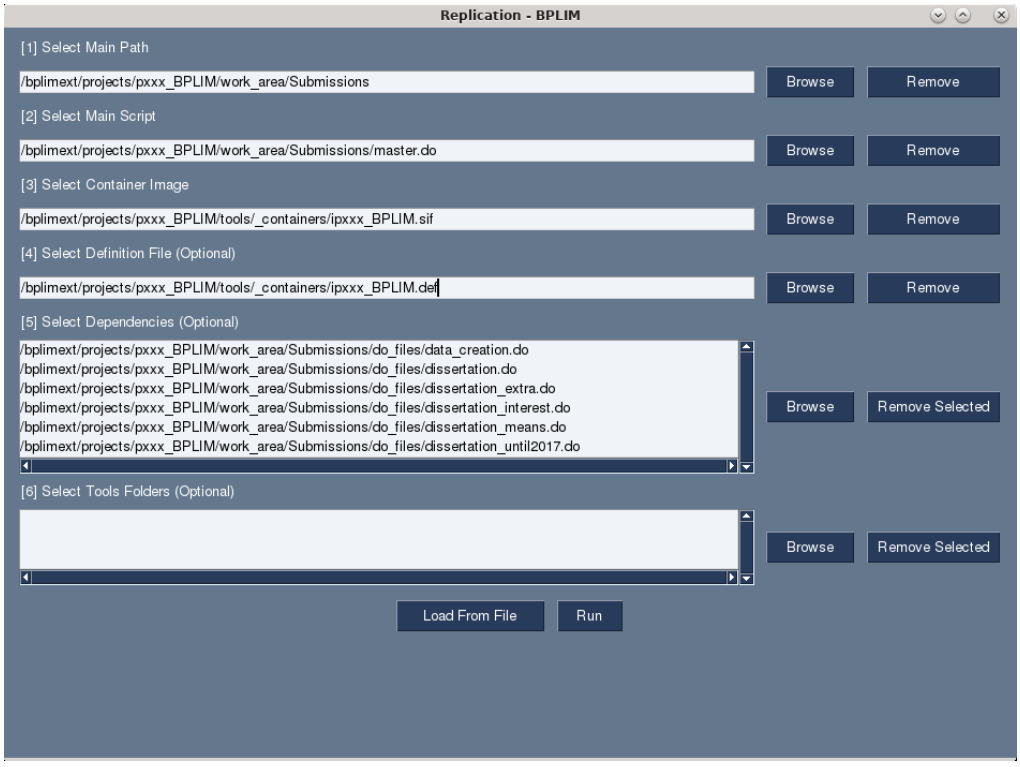
A seguir podemos selecionar o ficheiro de definição (ficheiro *.def) usado para criar o container. Este campo é opcional, pois é possível utilizar imagens externas de fontes fidedignas sem ficheiro de definição. Embora não seja obrigatório, é recomendável fornecer um ficheiro de definição, que será copiado para a área de replicação. Como não copiamos a imagem para a área de replicação por motivos de espaço, o ficheiro de definição é uma forma de garantir que o ambiente computacional do investigador possa ser recriado no futuro para reproduzir a análise.





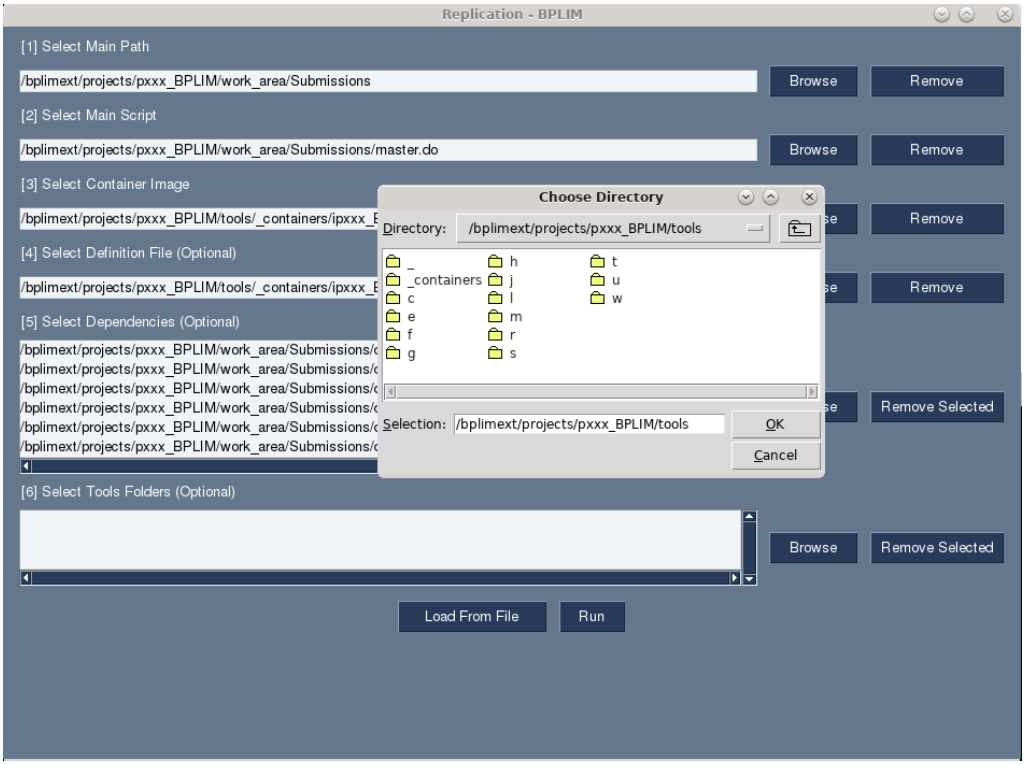
No quinto campo indicamos as dependências que serão utilizadas na replicação. As dependências não são mais do que outros scripts chamados pelo script principal e que podem ser colocados em diretórios ou subdiretórios para fins de organização. Este campo não é obrigatório porque o investigador tem liberdade para executar toda a análise utilizando apenas o script principal. No nosso exemplo, especificamos seis ficheiros neste campo:

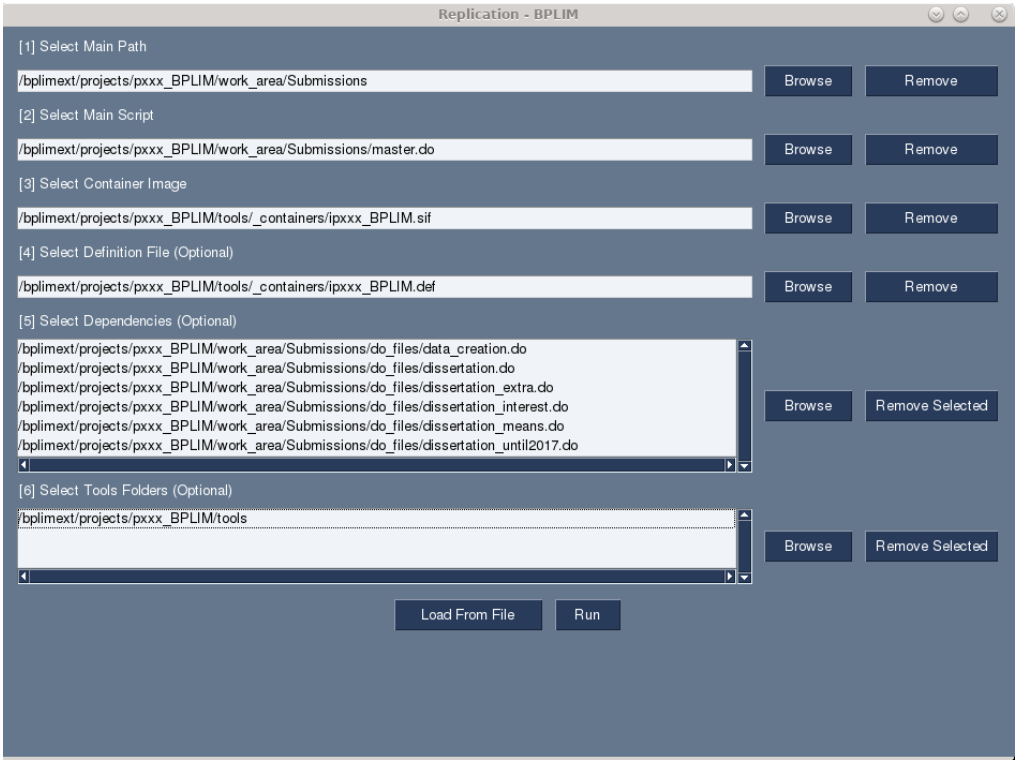




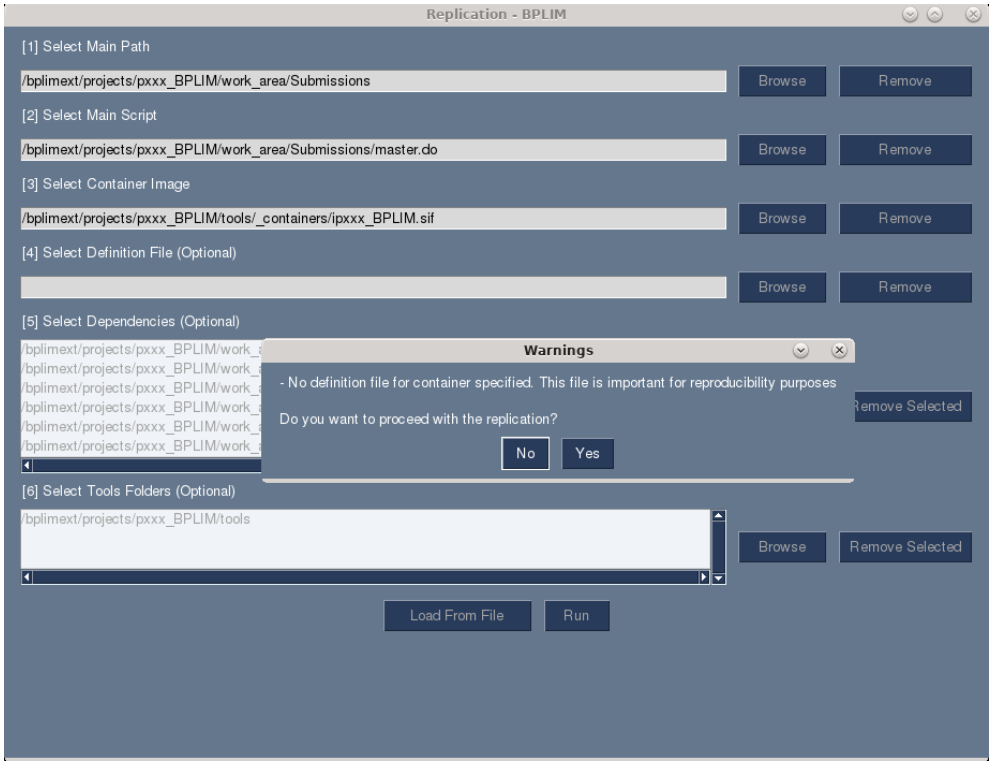
Finalmente, podemos selecionar diretórios para ferramentas. Por ferramentas entendemos programas, módulos ou pacotes utilizados pelo investigador na replicação. A aplicação cria um ficheiro de configuração que apontará para esses diretórios para ter as ferramentas disponíveis no ambiente de execução. O utilizador pode selecionar um ou mais diretórios. Dois casos são possíveis: se o diretório especificado estiver no caminho principal selecionado no primeiro campo, todo o diretório será copiado para a área de replicação, embora o seu tamanho não possa exceder 10MB; caso contrário, o ficheiro de configuração simplesmente apontará para esse caminho, não copiando seu conteúdo para a área de

replicação. No exemplo em questão, apenas o último se aplica, uma vez que os ados (comandos de Stata) necessários foram todos colocados fora do caminho de replicação base:

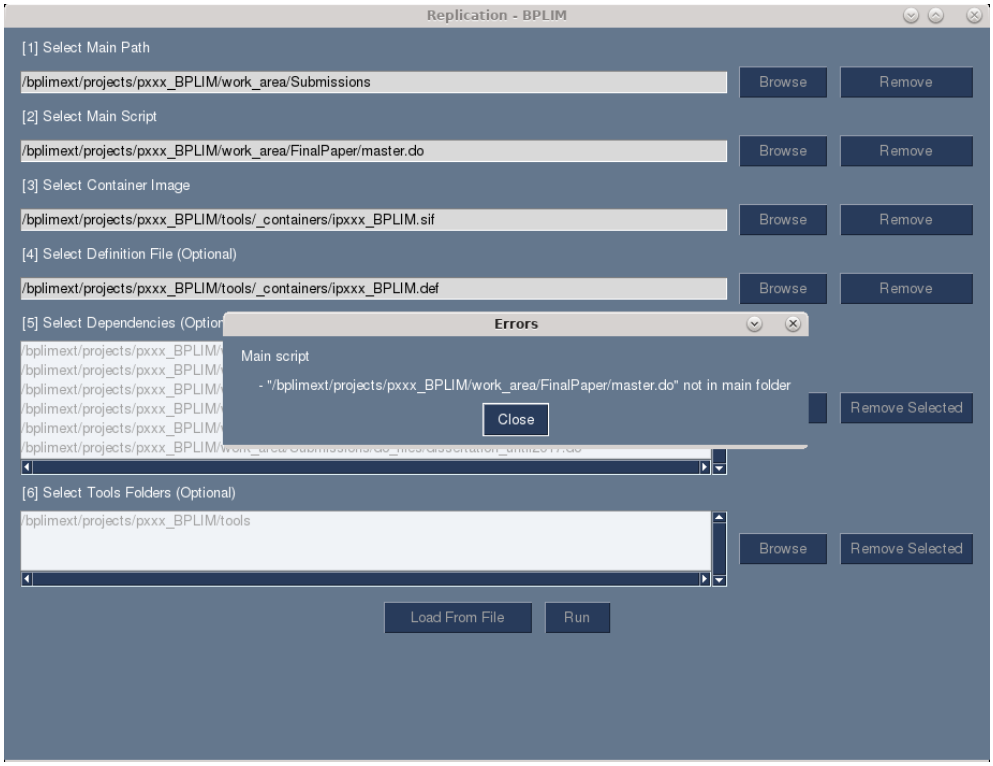




Com os campos preenchidos corretamente, clicamos no botão Run para executar a replicação. Se não houver erros e avisos, a replicação deverá começar imediatamente. Porém, caso existam *warnings*, a aplicação exibirá uma caixa de diálogo perguntando ao utilizador se deseja prosseguir com a replicação. Por exemplo, se não especificássemos um ficheiro de definição, receberíamos um aviso:

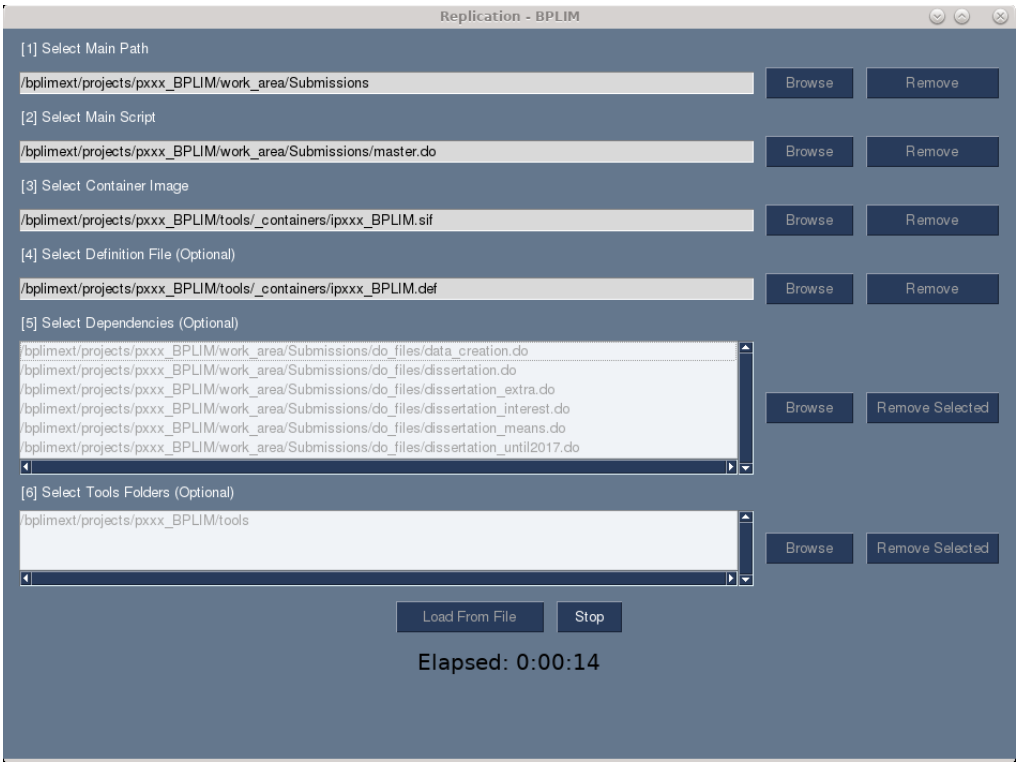


Podemos prosseguir com a replicação clicando em Yes. A situação é diferente se a aplicação encontrar algum erro. Neste caso, o utilizador não tem permissão para prosseguir com a replicação e deve alterar o campo que está a causar o erro. Por exemplo, imagine que especificamos um script principal que não está no caminho principal de replicação (primeiro campo). Isso resultará no erro seguinte:

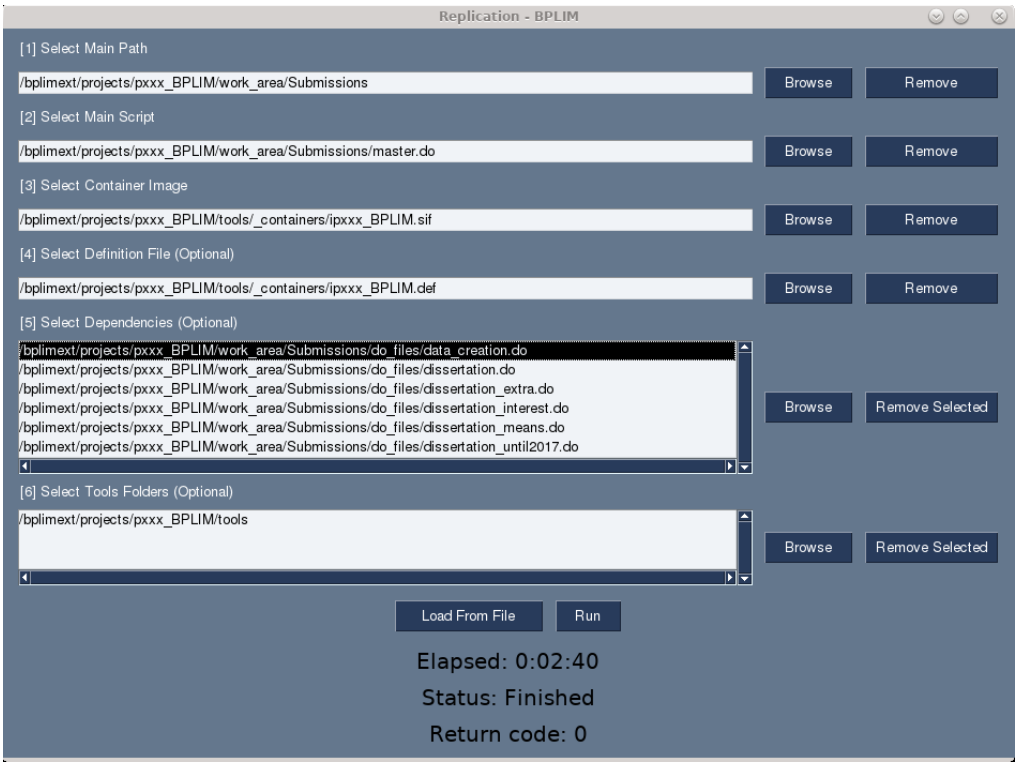


O utilizador deverá indicar um ficheiro válido no campo do script principal para executar a replicação.

Se todos os campos forem preenchidos corretamente, a replicação será iniciada e a aplicação mostrará um contador, exibindo o tempo decorrido desde o início da replicação:



O utilizador deverá ver o seguinte após uma replicação bem-sucedida:



O campo Status tem dois resultados possíveis: Finished - quando a replicação termina, com erro ou sem erros - ou Interrupted - quando o utilizador clica no botão Stop durante uma execução, terminando efetivamente a replicação. O campo Return code é exibido quando o resultado Status é Finished e também tem dois resultados possíveis: 0 - a replicação terminou sem erros - ou 1 - a replicação terminou, mas com erros. No nosso exemplo, a replicação terminou sem erros.

Mas qual é a vantagem de executar a replicação através da aplicação em vez de, por exemplo, executá-la diretamente no Stata? Além de ser um produto BPLIM que nos ajuda a tornar mais expedito o processo

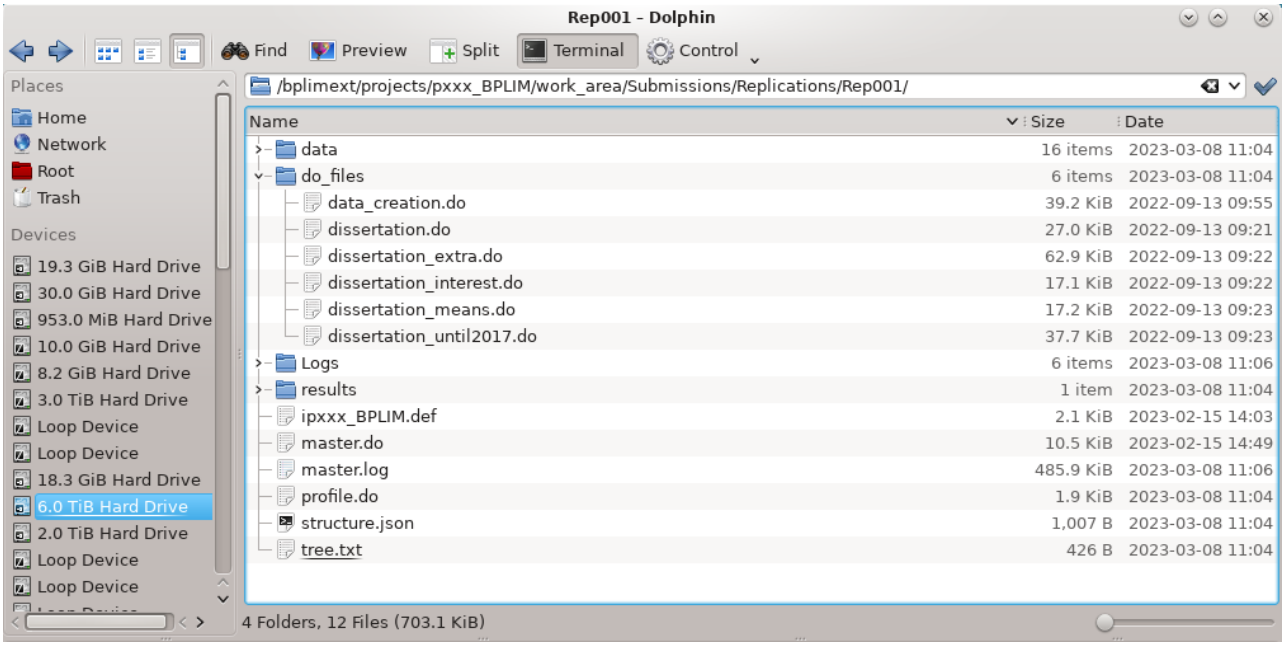
de replicação, a aplicação apresenta vantagens para os investigadores. Para ver essas vantagens, devemos inspecionar os resultados da aplicação. Quando clicamos no botão Run, a aplicação passa por uma série de etapas antes de executar o código.

Primeiro, caso não exista, o diretório Replications é criado no caminho principal da replicação - no nosso caso, "/bplimext/projects/pxxx_BPLIM/work_area/Submissions". No diretório Replications, o diretório Rep001 será criado automaticamente. Note que o número incluído no nome do diretório será atualizado automaticamente se já houver outras replicações dentro da pasta Replications.

Se for executado pela primeira vez, a aplicação cria a pasta

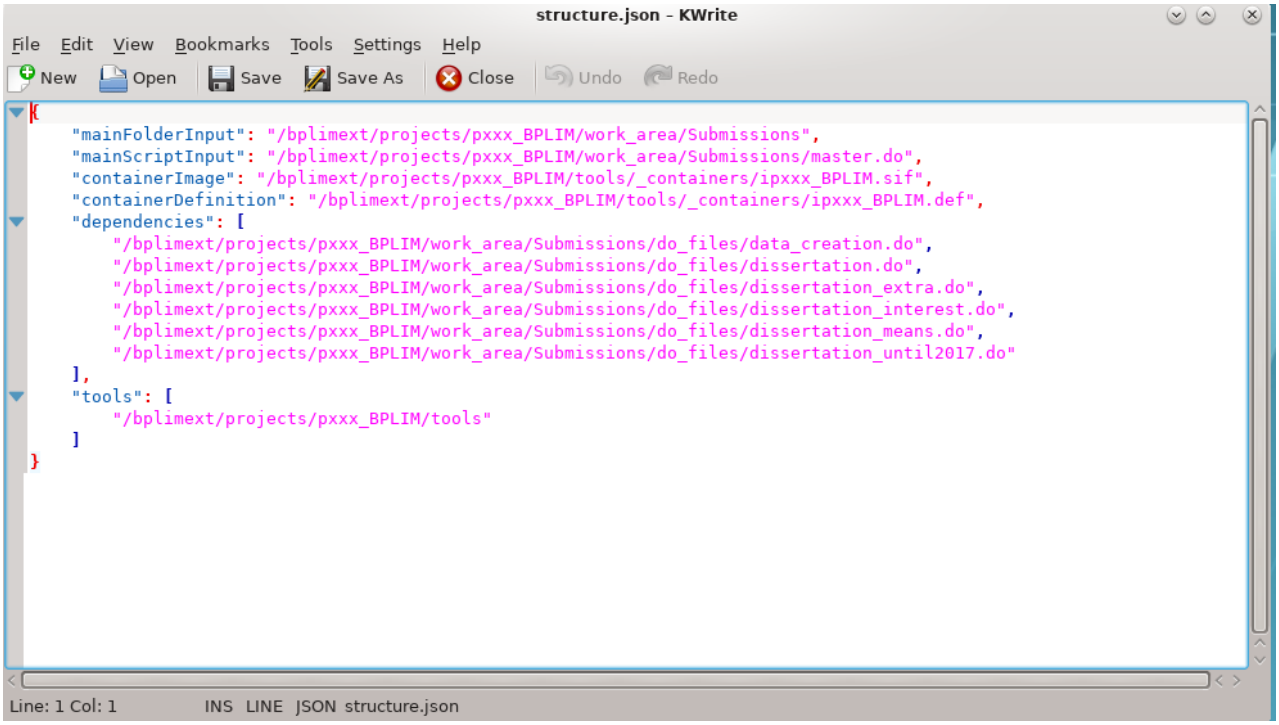
"/bplimext/projects/pxxx_BPLIM/work_area/Submissions/Replications/Rep001".

A estrutura de ficheiros do nosso exemplo é exibida na imagem abaixo (após o término da replicação):

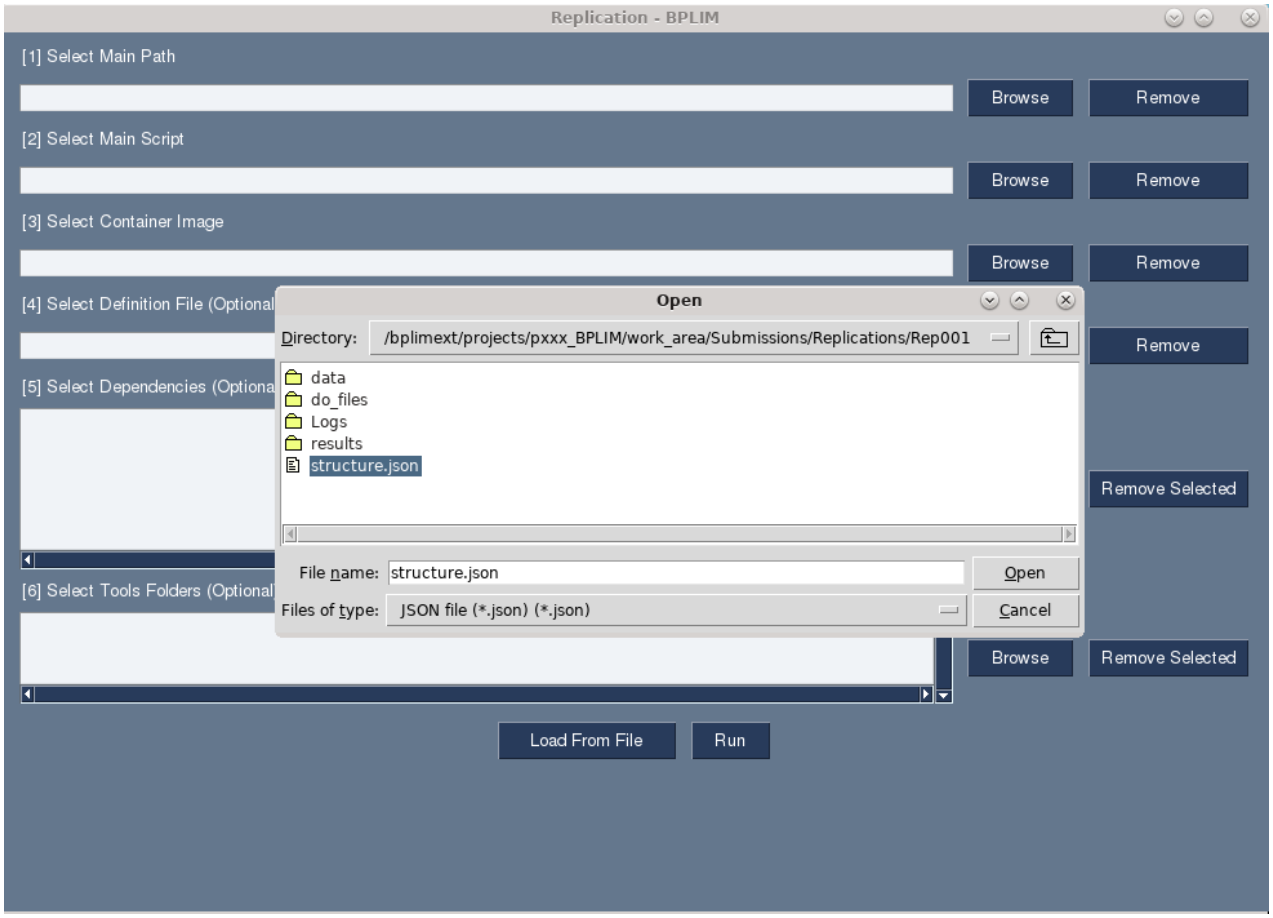


Os diretórios data, Logs e results (conforme especificado pelo investigador), bem como os ficheiros que eles contêm, são criados pelo código durante a replicação. Os demais diretórios e ficheiros foram copiados ou criados pela aplicação. Podemos ver que as dependências, o script principal e o ficheiro de definição foram copiados para a área de replicação. Toda a estrutura do caminho principal (primeiro campo da replicação) é copiada para a área de replicação. Mas alguns ficheiros adicionais também foram criados, como structure.json, que contém informação sobre cada campo da aplicação para uma replicação específica. O ficheiro serve não apenas para documentar a replicação, apresentando os caminhos originais para ficheiros e diretórios, mas também é útil em replicações futuras. Imagine que a segunda execução do utilizador usa os mesmos ficheiros da primeira ou dependências adicionais, por exemplo.

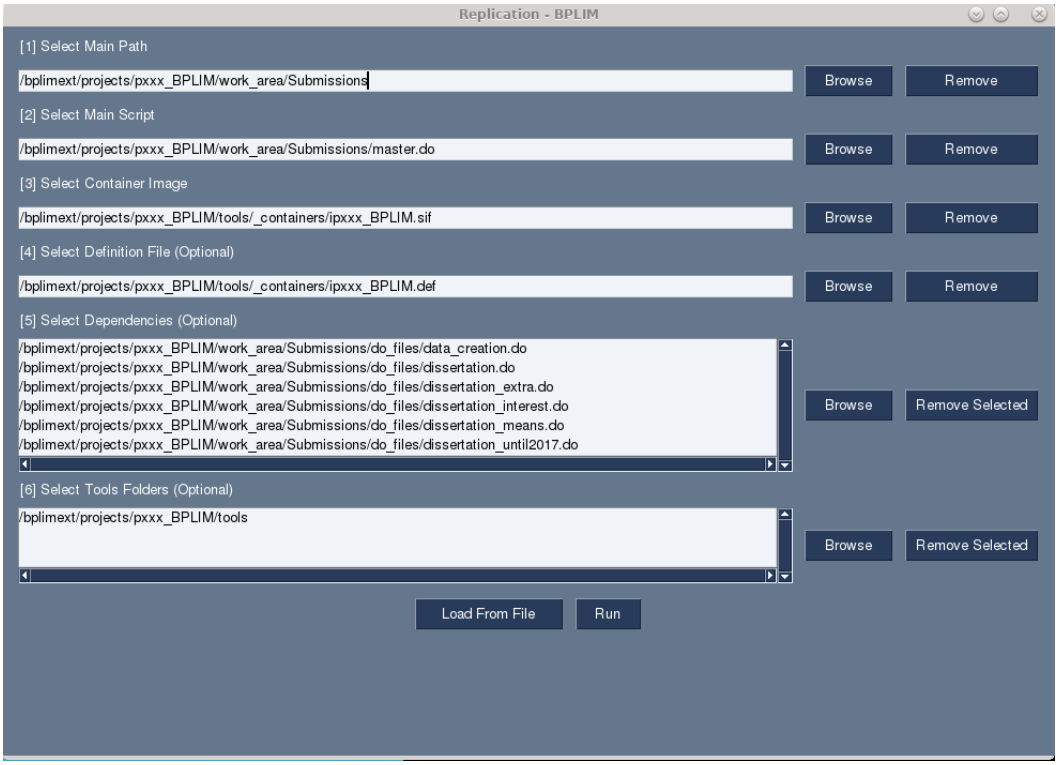
Nesse caso, ao iniciar a aplicação, o utilizador não precisará de preencher novamente todos os campos.



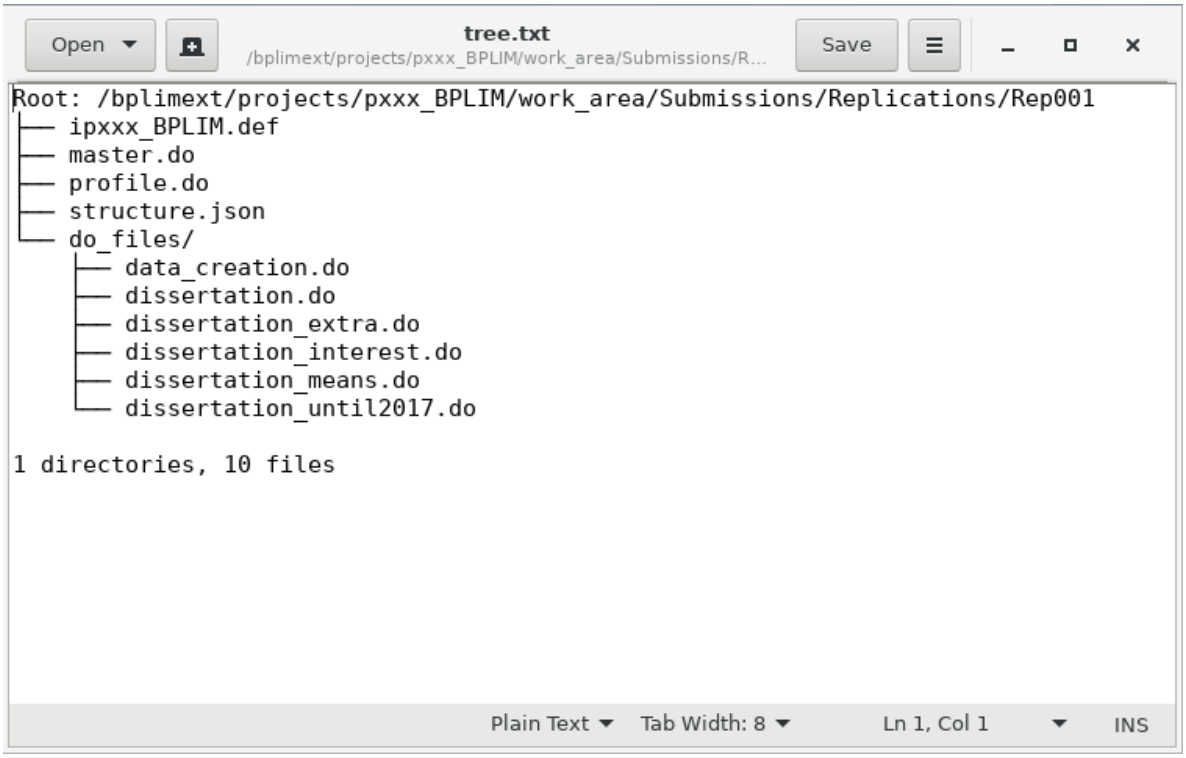
O botão Load from file abre uma caixa de diálogo para selecionar um ficheiro JSON:



Após selecionar o ficheiro structure.json, clique em “Open” e todos os campos serão preenchidos automaticamente com as informações da primeira replicação:



Outro ficheiro criado pela aplicação é tree.txt, que contém a estrutura em árvore da área de replicação antes de executar a replicação:



Finalmente, o último ficheiro criado pela aplicação é o ficheiro de configuração para executar esta replicação específica. Como estamos a usar Stata, o nosso ficheiro de configuração chama-se profile.do e é colocado no mesmo diretório do script principal. Ao colocar o ficheiro na mesma pasta que o script principal, o Stata executa o profile.do antes de executar o ficheiro principal. Desta forma, o ficheiro profile.do é usado como ficheiro de configuração, definindo globais, paths, etc. Como poderá ser notado,

os caminhos para os globais path_rep, path_source* e comando adopath++ são baseados no projeto e nos campos de entrada da aplicação.

Os investigadores devem usar os globais definidos no ficheiro profile.do nos seus scripts ao enviar uma replicação. Os diretórios data e results também serão copiados para a área de replicação. Portanto, o investigador deve certificar-se que erros na criação da estrutura de pastas são devidamente capturados. Durante a replicação o investigador pode usar caminhos absolutos ou relativos. Se estiver a usar caminhos absolutos, é imperativo usar os globais definidos no ficheiro profile.do e no script principal.

Quer o investigador utilize caminhos relativos ou absolutos, a equipa do BPLIM só precisa de alterar o ficheiro profile.do para executar a replicação com outras configurações. No caso de caminhos absolutos, os mesmos devem sempre ser especificados em termos dos globais previamente definidos, pelo menos path_rep.

Caso o investigador utilize R, a aplicação cria um ficheiro de configuração chamada config.R, no mesmo diretório do script principal.

Conclusão

O processo de acesso a dados confidenciais no BPLIM exige que os investigadores externos trabalhem nos servidores do Banco usando dados "modificados". Esses dados "modificados" servem apenas para preparar os scripts de análise. A equipa do BPLIM usa esses scripts para replicar o trabalho dos investigadores utilizando os dados originais em vez dos dados "modificados". Este processo tem a vantagem de obrigar o investigador a tornar a sua investigação reprodutível. O BPLIM desenvolveu uma aplicação que ajuda neste processo e que permite a criação automática de um pacote de replicação. Este pacote de replicação contém toda a informação necessária para reproduzir o trabalho dos investigadores, à exceção dos dados confidenciais. Pode por isso ser divulgado publicamente, e satisfaz os requisitos exigidos pelas revistas científicas que exigem reprodutibilidade da investigação.

Ciência Estatística

• Artigos em Revistas

Papança, F. (2024) - O Curso Vestibular da Academia Militar de Portugal nas Áreas da Matemática e da Estatística. *EasyChair* Preprint 13494.

• Livros

Título: *Stochastic Processes: Theory, Examples & Exercises*
Autor: Manuel Cabral Morais
Ano: 2024
Editora: IST Press.
ISBN: 978-989-9208-00-1 (500 páginas).
Mais informação: https://istpress.tecnico.ulisboa.pt/produto/7503/

Título: *Como os Livros Contam uma História da Estatística em Portugal: Obras de Estrangeiros, de Estrangeirados, e dos Outros*
Autores: Dinis Pestana e Rui Santos
Ano: 2024
Disponível em https://www.spestatistica.pt/publicacoes/categoria/outras-publicacoes

Título: *Breve Historial da Escola de Extremos e Aplicações em Portugal (PORTSEA)*
Autora: M. Ivette Gomes
Ano: 2024
Disponível em https://www.spestatistica.pt/publicacoes/categoria/outras-publicacoes

Título: *Relembrando Bento Murteira no ano do seu Centenário*
Autora: Maria Antónia Amaral Turkman
Ano: 2024
Disponível em https://www.spestatistica.pt/publicacoes/categoria/outras-publicacoes

Título: *Introdução à Estatística* (4ª Edição)
Autores: Bento Murteira, Carlos Pimenta, Filomena Pimenta, Carlos Silva Ribeiro e João Andrade e Silva
Ano: 2023
Editora: Escolar Editora

• Tese de Mestrado

Título: *Previsão de indicadores de gestão*
Autora: Tiago Fernandes Martins, *tiago_martins1998@hotmail.com*
Orientadoras: Sandra Ramos e Maria do Carmo Miranda Guedes

Doutoramento

Título: *Uncertainty and urgency: Tackling ME/CFS research and assessing the risk of SARS-CoV-2 Omicron subvariants in vaccinated populations*
Autor: João Malato, *jmalato@medicina.ulisboa.pt*
Orientadores: Nuno Sepúlveda e Luís Graça

A minha tese de doutoramento insere-se no ramo das Ciências Biomédicas e foca-se em estratificação na encefalomielite miálgica/síndrome de fadiga crónica (EM/SFC) e risco de infeção SARS-CoV-2 no contexto de populações altamente vacinadas.

EM/SFC é uma doença complexa e semelhante à COVID longa. A inexistência de biomarcadores para a identificação da EM/SFC faz com que o seu diagnóstico se baseie unicamente em critérios de avaliação de sintomas. No entanto, a heterogeneidade dos sintomas (e subjetividade no seu reporte) podem levar ao diagnóstico incorreto de doentes de outras condições com um perfil sintomatológico semelhante. Segundo este pressuposto, realizámos estudos de simulação para avaliar o impacto do diagnóstico incorreto na diluição de possíveis fatores relacionados com a doença—à semelhança de contextos de má classificação em estatística. Demonstrámos que estudos em EM/SFC sofrem de um poder estatístico insuficiente para detetar de forma consistente (e replicável) potenciais associações de interesse. Para lidar com esta incerteza, sugerimos focar subgrupos de doentes com mecanismos patológicos subjacentes, numa tentativa de agrupar por fatores desencadeadores da doença. Através de escalonamento multidimensional e análise de classes latentes sugerimos a estratificação de doentes de acordo com a severidade dos seus sintomas, relacionando estes resultados com a evidência de alterações do sistema imune causada por infeções. Embora ainda seja necessária mais investigação, a estratificação de doentes fornece informações promissoras sobre as diferentes vias que podem levar a esta doença heterogénea.

Esta tese foi elaborada no decorrer da pandemia de COVID-19. Em 2022 a subvariante Ómicron BA.5 substituiu variantes anteriores (BA.1 e BA.2), o que levantou sérias preocupações sobre a eficácia das vacinas ainda em desenvolvimento. Nessa altura, tive a oportunidade de analisar dados oficiais da pandemia em Portugal e estimar os riscos de infeção pela BA.5 na população vacinada. Através de uma regressão de Poisson com um estimador sanduíche para variância do erro comparámos o risco de infeção em indivíduos sem infeção prévia documentada (imunidade vacinal) relativamente aos infetados durante períodos de dominância de variantes anteriores (imunidade híbrida). Concluímos que as populações com imunidade híbrida estavam 2 a 4 vezes mais protegidas contra infeção pela nova variante, mantendo-se eficazmente protegidas até 8 meses após a primeira infeção, considerando o decaimento da imunidade dependente do tempo. Estes resultados estabeleceram que uma primeira infeção reduzia o risco de infeções por BA.5 em populações altamente vacinadas (como o caso de Portugal), e mostraram que as vacinas adaptadas a BA.1 poderiam ainda ser disponibilizadas e usadas eficazmente até que opções mais específicas estivessem disponíveis. Além disso, a prova da estabilidade da proteção da imunidade híbrida permitiu às autoridades sanitárias planear doses de reforço subsequentes.

Para mais informações e discussão estatística, contactem-me! Este trabalho foi financiado por uma Bolsa Mista da Fundação para a Ciência e Tecnologia (referência SFRH/BD/149758/2019).

João Torrado Malato

Edições SPE - Mini Cursos

Título: *Análise Estatística de Dados Financeiros*
Autores: C. Amado, C. Nunes, A. Sardinha
Ano: 2019.

Título: *Uma introdução à Meta-Análise*
Autora: Maria de Fátima Brilhante
Ano: 2017.

Título: *Estatística Bayesiana Computacional – uma introdução*
Autores: M. Antónia Amaral Turkman e Carlos Daniel Paulino
Ano: 2015.

Título: *Análise de Valores Extremos: Uma Introdução*
Autoras: M. Ivette Gomes, M. Isabel Fraga Alves e Claudia Neves
Ano: 2013.

Título: *Modelos com Equações Estruturais*
Autora: Maria de Fátima Salgueiro
Ano: 2012.

Título: *Análise de Dados Longitudinais*
Autoras: Maria Salomé Cabral e Maria Helena Gonçalves
Ano: 2011

Título: *Uma Introdução à Estimação Não-Paramétrica da Densidade*
Autor: Carlos Tenreiro
Ano: 2010

Título: *Análise de Sobrevivência*
Autoras: Cristina Rocha e Ana Luísa Papoila
Ano: 2009

Título: *Análise de Dados Espaciais*
Autoras: M. Lucília de Carvalho e Isabel C. Natário
Ano: 2008

Título: *Introdução aos Métodos Estatísticos Robustos*
Autores: Ana M. Pires e João A. Branco
Ano: 2007

Título: *Outliers em Dados Estatísticos*
Autor: Fernando Rosado
Ano: 2006

Título: *Introdução às Equações Diferenciais Estocásticas e Aplicações*
Autor: Carlos Braumann
Ano: 2005

Título: *Uma Introdução à Análise de Clusters*
Autor: João A. Branco
Ano: 2004

Título: *Séries Temporais – Modelações lineares e não lineares*
Autoras: Esmeralda Gonçalves e Nazaré Mendes Lopes
Ano: 2003 (2ª Edição em 2008)

Título: *Modelos Heterocedásticos. Aplicações com o software Eviews*
Autor: Daniel Muller
Ano: 2002

Título: *Inferência sobre Localização e Escala*
Autores: Fátima Brilhante, Dinis Pestana, José Rocha e Sílvio Velosa
Ano: 2001

Título: *Controlo Estatístico de Qualidade*
Autoras: Maria Ivette Gomes; Fernanda Figueiredo e Maria Isabel Barão
Ano: 1999 (2.ª edição, revista e aumentada, 2010)

Título: *Tópicos de Sondagens*
Autor: Paulo Gomes
Ano: 1998

Retrospetiva do Boletim SPE - Tema Central

Disponíveis em *https://www.spestatistica.pt/publicacoes/categoria/boletim-da-spe*

Primavera de 2024 – *Rising Stars*
Outono de 2023 – Educação (e) Estatística
Primavera de 2023 – PORTSEA – um mar de Extremos em Portugal
Outono de 2022 – Prémios na Sociedade Portuguesa de Estatística
Primavera de 2022 – Liderança Estatística
Outono de 2021 – *Machine Learning* e Inteligência Artificial
Primavera de 2021 – Especial Covid: a Estatística ao serviço da sociedade
Outono de 2020 – 40 anos SPE: De onde viemos? Onde estamos? Para onde vamos?
Primavera de 2020 – INE–85 anos de estatísticas a servir o país
Outono de 2019 – Estatística nas Ciências da Saúde
Primavera de 2019 – Séries Temporais de Valor Inteiro
Outono de 2018 – Equações diferenciais estocásticas e algumas aplicações
Primavera de 2018 – Estatística Multivariada – perspetiva no século XXI
Outono de 2017 – O Tema Central da Estatística - um novo olhar
Primavera de 2017 – Incerteza em Engenharia
Outono de 2016 – O Tema Central da Estatística
Primavera de 2016 – Séries Temporais e suas aplicações
Outono de 2015 – Estatística em Genética
Primavera de 2015 – Estatística no Desporto
Outono de 2014 – Estatística no Ensino Básico e Secundário
Primavera de 2014 – (Um) Ano Internacional da Estatística
Outono de 2013 – A "Escola Bayesiana" em Portugal
Primavera de 2013 – Estatística não - paramétrica
Outono de 2012 – Métodos Estatísticos em Medicina
Primavera de 2012 – Estatística no Ensino Superior Politécnico
Outono de 2011 – Análise de Sobrevivência
Primavera de 2011 – Sondagens e Censos
Outono de 2010 – Estatística Espacial
Primavera de 2010 – Data Mining - Prospecção (Estatística) de Dados?
Outono de 2009 – Modelos Econométricos
Primavera de 2009 – Investigação (em) Estatística
Outono de 2008 – Processos Estocásticos
Primavera de 2008 – ALEA - Um sítio do nosso mundo
Outono de 2007 – Bioestatística
Primavera de 2007 - A "Escola de Extremos" em Portugal
Outono de 2006 – Ensino e Aprendizagem da Estatística

O MUNDO DA ESTATÍSTICA

ORGANIZAÇÃO PARTICIPANTE

FENStatS

Federation of European National Statistical Societies

Índice

Este boletim tem o apoio de:

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR