REGEPE ENTREPRENEURSHIP AND SMALL BUSINESS JOURNAL

EDITORA ANEGEPE

ISSN: 2965-1506 v.12, n.2, maio / ago., 2023 https://regepe.org.br/

Editorial

Pesquisa quantitativa em empreendedorismo e o apoio do software R para análise de dados

Daniel do Prado Pagotto[a]* e Cândido Borges[b]

[a] *Universidade de Brasília (UnB), Brasília, DF, Brasil*

[b] *Universidade Federal de Goias (UFG), Goiânia, GO, Brasil*

Detalhes Editoriais

Histórico do Artigo
Disponível online: *02 de maio de 2023*

Classificação JEL: *C00, L26*

ID do Artigo: *2257*

Editor Chefe[1] ou Adjunto[2]:
[1] *Dr. Edmundo Inácio Júnior*
Universidade Estadual de Campinas, UNICAMP

Editor Associado Responsável:
Dra. Rose Mary Almeida Lopes
ANEGEPE

Editor Executivo[1] ou Assistente[2]:
[2] *M. BA. João Paulo Moreira Silva*

Revisão Ortográfica e Gramatical:
Dra. Mônica Império Costa
Palavra Seleta Revisão Textual

Item relacionado (hasTranslation):
https://doi.org/10.14211/regepe.esbj.e2384

Citar como:
Pagotto, D. do P., & Borges, C. (2023). Pesquisa quantitativa em empreendedorismo e o apoio do software R para análise de dados. *REGEPE Entrepreneurship and Small Business Journal*, 12(2), e2257. https://doi.org/10.14211/regepe.esbj.e2257

***Autor de contato:**

Daniel do Prado Pagotto
danielppagotto@gmail.com

Resumo

Objetivo do estudo: o presente texto visa apresentar um panorama sobre pesquisa quantitativa em empreendedorismo no Brasil, bem como descrever possibilidades para o avanço desta abordagem. **Metodologia e abordagem**: o artigo consiste em uma publicação conduzida a partir de levantamentos bibliográficos na literatura científica de empreendedorismo e discussões teóricas. **Principais Resultados**: maior parte das pesquisas nacionais em empreendedorismo são de natureza qualitativa**.** Apesar da relevância desta abordagem, acredita-se que a pesquisa quantitativa possui múltiplas potencialidades, sobretudo associada ao uso de dados oriundos de fontes secundárias. **Principais Contribuições teóricas e metodológicas**: apresentamos bases de dados públicas que podem ser empregadas por pesquisadores de empreendedorismo para avançar na teoria. Algumas estratégias de uso destas bases são exemplificadas por meio de um breve tutorial em linguagem R. Finalmente, debatemos acerca de estratégias para robustecer pesquisas quantitativas da área, bem como trazemos uma agenda de pesquisa. **Relevância/Originalidade**: são apresentados conteúdos que ainda são pouco explorados na literatura nacional, como o uso de dados secundários e machine learning. **Contribuições sociais e gerenciais**: algumas das bases apresentadas no estudo são de fonte governamental e podem ser utilizadas para fundamentar a construção de políticas públicas para o empreendedorismo. Ademais, os preceitos sobre pesquisa quantitativa apresentados neste editorial podem apoiar gestores que atuam com análises de dados na formulação de estudos mais robustos, independente da área de atuação, seja prático ou acadêmico.

Palavras-chave: Métodos quantitativos. Software R. Dados secundários.

Quantitative research in entrepreneurship using the R software for data analysis

Abstract

Objective of the study: this editorial aims to present an overview of Brazilian quantitative research in entrepreneurship, as well as describing possibilities for advancing this methodological approach. **Methodology and approach:** the article consists of an editorial publication, built from bibliographic research of entrepreneurship literature and theoretical reflections. **Main Results:** Most national entrepreneurship research follows a qualitative approach. Despite its relevance, quantitative research also has multiple potentialities, especially associated with the use of data originating from secondary sources. **Main theoretical and methodological contributions:** We present public databases that can be used by entrepreneurship researchers to advance theory. Some strategies for using these bases are exemplified through a brief tutorial in R language. We further debate about strategies to strengthen quantitative research in the area. Finally, we bring a research agenda. **Relevance/Originality:** contents that are still little explored in the national literature are presented, such as the use of secondary data and machine learning. **Social and managerial contributions:** some of the databases presented in the study come from government sources and can be used to support the construction of public policies for entrepreneurship. In addition, the precepts on quantitative research presented in this editorial can support managers who work with data analysis to perform more robust studies, regardless of the area, whether practical or academic.

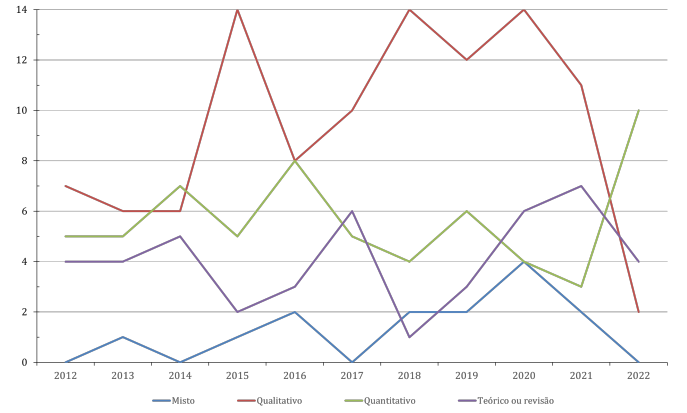
Keywords: Quantitative methods. R software. Secondary data.

INTRODUÇÃO

No Brasil, a abordagem qualitativa é o método mais utilizado nas pesquisas em empreendedorismo. Em levantamento realizado entre os anos de 2012 e o último volume de 2022 da Revista de Empreendedorismo e Gestão de Pequenas Empresas (REGEPE), observa-se que foi publicado um número superior de trabalhos com essa abordagem quando comparado com a quantitativa - 104 da primeira e 62 da segunda. Ademais, conforme apresentado na Figura 1, percebe-se que, apenas no último ano, houve uma reversão na prevalência de pesquisas qualitativas sobre quantitativas.

Figura 1

Evolução de publicações por abordagem



Nota: Elaborado pelos autores.

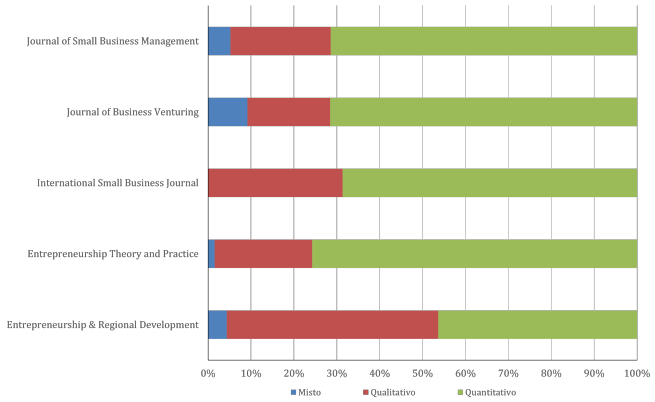
Diferentes revisões ou bibliometrias publicadas no Brasil nos últimos anos chegaram à mesma conclusão. Nassif et al. (2010) encontraram uma predominância de estudos que adotaram métodos qualitativos em uma revisão sobre o perfil das publicações do Encontro de Estudos sobre Empreendedorismo e Gestão de Pequenas Empresas (EGEPE) e o Encontro da Associação Nacional de Pós-Graduação e Pesquisa em Administração (Enanpad) entre 2000 e 2008. Em uma amostra de 219 artigos de natureza teórica-empírica, 60,7% foram qualitativos. Oliveira et al. (2018) analisaram publicações sobre empreendedorismo entre 2000 e 2014 em seis periódicos de administração e também constataram uma prevalência de publicações com metodologias qualitativas. De um total de 54 estudos empíricos, 51,9% foram qualitativos, 11,1% foram mistos e 37% foram quantitativos. Em um universo de 179 artigos publicados entre 2004 e 2020, Ferreira et al. (2020) identificaram que 44% das publicações eram qualitativas, 27% quantitativas e 25% teóricas.

Esta característica da pesquisa nacional difere do panorama internacional, onde os estudos quantitativos predominam. McDonald et al. (2015) realizaram um levantamento em seis dos principais periódicos internacionais de empreendedorismo com horizonte temporal de 1985 a 2013 e constataram que, em uma amostra de 3749 trabalhos, a maioria possui abordagem quantitativa (55%). Avaliando as publicações de 2020 e 2021 das mesmas revistas do levantamento de McDonald et al. (2015) constata-se que, com exceção da Entrepreneurship & Regional Development, as demais publicam com maior frequência estudos quantitativos. Em um universo de 362 artigos empíricos analisados, 69,06% foram de abordagem quantitativa (cf. Figura 2).

Ter um número superior de pesquisas com abordagem qualitativa no Brasil não é necessariamente um problema. Nas ciências sociais aplicadas, a pesquisa qualitativa tem um papel fundamental no desenvolvimento do conhecimento cientifico (Cristi, 2018), e no campo do empreendedorismo não é diferente (Gil & Silva, 2015; Neergaard & Ulhoi, 2007). O problema é o fato de a abordagem quantitativa nas pesquisas em empreendedorismo em território nacional manter um histórico de baixa utilização. A baixa presença de pesquisas quantitativas pode deixar de extrair vantagens que esse tipo de abordagem pode proporcionar, como: possibilidade de abranger amostras representativas de casos; testar e validar teorias construídas e exploradas inicialmente por métodos qualitativos; generalização mediante desenhos amostrais e técnicas de análise adequadas (Cooper & Schindler, 2014). Não avançar em pesquisas que utilizam a abordagem quantitativa pode representar um limite para o desenvolvimento do campo do empreendedorismo no Brasil.

Figura 2

Proporção de estudos por tipo e revistas



Nota: Elaborado pelos autores.

A explicação para esse número de publicações de natureza quantitativa ser baixo possui caráter multifacetado. No entanto, um possível motivo pode estar relacionado à baixa aplicação de dados secundários. De fato, considerando o levantamento feito na REGEPE, citado acima, nas pesquisas quantitativas ou de métodos mistos, mais do que o dobro dos estudos utilizou dados primários – 49 foram com dados primários contra 27 com dados secundários. Tal achado foi congruente ao encontrado por Oliveira Junior et al. (2018), que identificaram um percentual de 79,6% de estudos que recorrem a levantamentos primários, dentro de um universo de 54 artigos empíricos nacionais sobre empreendedorismo. A utilização de dados secundários, por evitar os custos e o tempo, inerentes à coleta de dados primários, e por permitir a realização de estudos com um maior número de observações, podem a alavancar a quantidade e a qualidade de pesquisa quantitativas (Hox and Boeije, 2005).

A maior utilização de fonte de dados secundários para pesquisas é potencializadao com o conhecimento de bases disponíveis e com o domínio de ferramentas e metodologias para acessá-las e utilizá-las. Nesse sentido, o presente artigo apresenta duas contribuições para o desenvolvimento das pesquisas em empreendedorismo com abordagem quantitativa que utilizem dados secundários: primeiro, são apresentadas bases de dados nacionais e internacionais, com acesso livre, que podem ser utilizadas para pesquisas em empreendedorismo; segundo, é apresentado um tutorial de uso do Software R para dados em empreendedorismo.

FONTES DE DADOS SECUNDÁRIOS PARA PESQUISAS EM EMPREENDEDORISMO

Como em muitas áreas das ciências sociais, a obtenção de dados para a pesquisa em empreendedorismo é um desafio. Por um lado, há uma escassez de fontes secundárias sobre etapas iniciais do processo de criação de negócios. Por outro lado, sob a perspectiva de levantamentos primários, acessar empreendedores

é trabalhoso, pois são indivíduos ocupados e cujos negócios estão constantemente sofrendo alterações, o que dificulta a captura de todos os fenômenos (Maula & Stam, 2020).

No entanto, os últimos anos têm registrado um número crescente de dados oriundos de diversas fontes, como registros de bancos de dados, técnicas de web-scrapping e vídeos (Maula & Stam, 2020; Obschonka and Audretsch, 2020). As duas últimas fontes – web-scrapping e vídeos - apresentam um grande potencial para pesquisas qualitativas e quantitativas, pois empregam técnicas de extração, interpretação e análise de dados não estruturados (ex.: textos, imagens, vídeos). Já a primeira fonte se caracteriza pela natureza estruturada dos dados. Cada uma será detalhada a seguir.

Os dados de natureza estruturada são as fontes secundárias mais utilizadas na pesquisa em empreendedorismo. Nesse universo se encontram bases governamentais, institucionais, registros de empresas e levantamentos dedicados prioritariamente ao estudo do fenômeno empreendedor (ex.: Panel Study of Entrepreneurial Dynamics - PSED e Global Entrepreneurship Monitor -GEM). Benatti et al. (2021), por exemplo, utilizaram dados do Microempreendedor Individual (MEI), extraídos a partir do repositório do Data Sebrae[1], para avaliar o efeito dessa categoria de empreendimento sobre o desenvolvimento econômico em municípios paulistas. Audretsch et al. (2021) conjugaram diversas fontes de dados – dentre elas o GEM – para avaliar o efeito de variáveis institucionais sobre o empreendedorismo de oportunidade e necessidade a nível país. Algumas dessas bases serão detalhadas na seção 3.

A exploração das fontes não estruturadas – textos e vídeos – é pouco presente nas pesquisas em empreendedorismo, mas sua utilização é crescente devido ao avanço de ferramentas de análises de dados, pelo poder de processamento de computadores pessoais e as tecnologias de processamento em nuvem. O web-scrapping é uma técnica de coleta e extração de dados oriunda de páginas da internet (Prüfer & Prüfer, 2020). Obschonka et al. (2017), por exemplo, analisaram traços de personalidade de empreendedores e gestores "superstar" a partir de dados das publicações dos usuários na rede social Twitter. Ainda usando a rede social, Pagotto, Barbosa, et al. (2022) analisaram os sentimentos associados a tweets de empreendedores nos primeiros períodos da pandemia da Covid-19. Experiências prévias também incluem extração e análise de textos da grande mídia como os jornais The New York Times e Financial Times para avaliar as diferenças do teor de publicações sobre empreendedores e empresários (Suarez et al., 2020), assim como análise de áudio e vídeo de plataformas de financiamento coletivo para prever o sucesso das campanhas de levantamento de recursos (Kaminski & Hopp, 2020).

Considerando as duas fontes listadas, investigações que utilizam dados estruturados fazem parte da realidade nas pesquisas em empreendedorismo há algum tempo. Por outro lado, a análise de dados de natureza não estruturada já é algo factível, uma vez que já existem programas com interfaces intuitivas que realizam processamento de dados não estruturados, bem como bibliotecas em linguagens de programação como R e Python. Isso faz com que a análise de dados não estruturados seja compreendida como uma proposta inovadora para se mensurar e compreender fenômenos na área de empreendedorismo (Maula & Stam, 2020; von Bloh et al., 2020).

A partir dessa apresentação, percebe-se que a análise de dados não estruturados demonstra um caminho promissor devido a um conjunto de fatores, como: o aumento da disponibilidade de dados, amplificação do poder de processamento de máquinas domésticas e remotas, os softwares tradicionalmente utilizados na pesquisa qualitativa com funcionalidades mais avançadas sobre análises textuais (ex.: Nvivo, Atlas.ti) e pacotes em linguagens de programação dedicados a este fim. Ferramentas e softwares de análise de dados que anteriormente eram empregados primordialmente em pesquisas quantitativas, podem ser também agregados a abordagens qualitativas – sobretudo a partir de dados não estruturados – contribuindo, ocasionalmente, para um estreitamento entre ambas as perspectivas e, assim, fortalecendo a investigação na área. Todavia, a despeito do caráter inovador dos dados não estruturados, ainda há espaço para investigação ao se utilizar dados estruturados oriundos de registros secundários. A próxima seção traz algumas bases para pesquisa em empreendedorismo.

BASES DE DADOS ESTRUTURADOS EM EMPREENDEDORISMO

O objetivo dessa seção é apresentar algumas bases de dados secundários em empreendedorismo, demonstrando a relevância, potencialidade e trazendo alguns exemplos de aplicação delas em estudos científicos da área. A Tabela 1 indica algumas das bases utilizadas para a pesquisa em empreendedorismo. Nos parágrafos seguintes serão brevemente descritas as bases do PSED, do GEM e algumas bases nacionais de organizações governamentais.

Tabela 1

Bases de dados para pesquisa em empreendedorismo

Base de Dados	Perfil da Amostra
Panel Study of Entrepreneurial Dynamics	Empreendedores em estágio nascente dos EUA. Variações são encontradas em outros países, como Australia, Suécia
Global Entrepreneurship Monitor	Dados por país sobre condições para se empreender e atitudes frente à atividade empreendedora
Global Accelerator Learning Initiative	Empresas que passaram por programas de aceleradoras de empresas
Receita Federal Brasileira	Contém o registro de cadastro nacional de pessoa jurídica (CNPJ) das empresas e seus sócios
Bases IBGE do Sistema Integrado de Pesquisas Domiciliares	Levantamentos sobre características da população brasileira, incluindo desagregações por grupos como empregadores e trabalhadores por conta própria. Exemplos de bases: Pesquisa Nacional por Amostra de Domicílios Contínua (PNADc), Pesquisa Nacional de Saúde (PNS), Pesquisa de Informações Básicas Municipais (MUNIC), Censo Agro

Nota: Elaborado pelos autores.

O PSED foi um levantamento capitaneado por Paul Reynolds com o objetivo de coletar dados de uma amostra representativa de empreendedores americanos em formato de painel. Duas edições do PSED foram conduzidas nos EUA, sendo que a versão mais recente, o PSED 2, contou com o monitoramento de empreendimentos nascentes em ondas de entrevistas, aplicadas entre os anos de 2006 e 2011. O grande diferencial do PSED é o seu aspecto longitudinal e o fato de mapear o processo empreendedor e suas diferentes atividades, como a identificação da oportunidade, o registro legal da empresa, a primeira venda e o alcance do ponto de equilíbrio financeiro (Reynolds & Curtin, 2008). Os dados e material de apoio do PSED são disponíveis para livre acesso na página http://www.psed.isr.umich.edu/psed.

Iniciativas semelhantes ao PSED foram realizados em outros países, como na Austrália, China e Suécia, o que permitiu, inclusive, a criação de uma base única harmonizada com observações de todos esses levantamentos (Arenius et al., 2017; Reynolds et al., 2016; Warhuus et al., 2021). A base do PSED ou dos levantamentos derivados dele possuem grande potencial para mais estudos por alguns motivos: 1) os pesquisadores estimulam a investigação do empreendedorismo a partir de um recorte longitudinal, uma vez que a criação de empresas se trata de um processo dinâmico (Maula & Stam, 2020); 2) as bases possuem uma vasta diversidade

de dados, que versam, entre outros, sobre características dos empreendedores, do processo empreendedor, da empresa nascente, financiamento, estratégias de negócio, capital social, suporte da comunidade, motivações. Justamente devido a essa riqueza de dados, a base se difundiu nos estudos de empreendedorismo em temas como empreendedorismo familiar (Dyer et al., 2013), capital social (Semrau & Hopp, 2016), empreendedorismo por mulheres (Kwapsiz & Hechavarria, 2018), previsão da emergência e desistência de negócios nascentes (Koumbarakis & Volery, 2022), dentre outros.

O GEM é um levantamento que teve início em 1999 em múltiplos países e possui como objetivo monitorar aspectos sobre atitudes e comportamentos empreendedores da população, bem como a percepção sobre condições contextuais para empreender. Essas duas dimensões de análise do GEM se traduzem em dois levantamentos publicados anualmente: 1) o Adult Population Survey - APS, que investiga questões relacionadas à percepção da população adulta sobre visualização de oportunidades de negócios na sua localidade, percepção sobre as capacidades para se iniciar um negócio, taxa de empreendedorismo inicial, dentre outros; e o 2) National Expert Survey, levantamento direcionado a especialistas para capturar a percepção acerca de variáveis do contexto empreendedor, como financiamento para empreendedorismo, suporte governamental, tributação e burocracia, dentre outros. Cabe destacar que o GEM também publica relatórios e estudos dedicados a temas específicos, como empreendedorismo social, empreendedorismo familiar, empreendedorismo por mulheres.

Assim como o PSED, as bases do GEM também são amplamente utilizadas em estudos de empreendedorismo, sendo frequentemente conjugada a outros levantamentos, o que expande ainda mais a investigação do empreendedorismo associado a outros fenômenos. Dois exemplos desta composição de bases: Hechavarría and Ingram (2019) associaram o GEM a bases do Banco Mundial para investigar o efeito do ecossistema sobre a prevalência de empreendedorismo de ambos os sexos; Audretsch et al. (2021) combinaram várias bases – Worldwide Governance Indicators - WGI, gastos governamentais do Fundo Monetário Internacional e GEM – para avaliar o efeito de instituições nacionais sobre a taxa de empreendedorismo por necessidade e oportunidade. Um exercício de união do GEM a outra base será apresentado aqui.

Apesar de não serem exclusivamente direcionadas para a pesquisa em empreendedorismo, algumas bases de natureza governamental do Brasil constituem um grande potencial para pesquisadores do país. A seguir serão apresentadas algumas delas, que são disponíveis publicamente e que poderiam compor estudos quantitativos com dados secundários.

O Instituto Brasil de Geografia e Estatística - IBGE realiza levantamentos como a Pesquisa Nacional por Amostra de Domicílio - PNADc, a Pesquisa Nacional de Saúde - PNS, o Censo Agropecuário, e a Pesquisa de Informações Básicas Municipais - MUNIC. A Receita Federal Brasileira - RFB publica dados sobre Cadastro Nacional de Pessoas Jurídicas - CNPJ. O Ministério da Saúde, por meio do Departamento de Informática do Sistema Único de Saúde - DATASUS, consolida compulsoriamente alguns agravos que acometem a população por meio do Sistema de Informações de Agravos e Notificações - SINAN, assim como traz variáveis sobre morbidade e mortalidade, presentes, respectivamente, no Sistema de Informações Hospitalares - SIH e no Sistema de Informações sobre Mortalidade - SIM. O Ministério da Educação congrega bases sobre escolaridade a nível municipal. Recentemente, a Escola Nacional de Administração Pública - ENAP, em parceria com a Endeavor, publicou o último Índice de Cidades Empreendedoras (ICE - 2020) que contempla dados sobre os 100 maiores municípios brasileiros. Algumas dessas bases possuem dados desagregados a nível município (ex.: MUNIC e ICE), outras estão a nível indivíduo (ex.: SINAN).

Em se tratando de bases a nível indivíduo, uma observação deve ser considerada antes de apresentar exemplos de aplicação delas: frequentemente empreendedores podem ser identificados como empregadores ou trabalhadores por conta própria em alguns destes levantamentos. Muitas vezes o primeiro é associado aos empreendedores por oportunidade enquanto os segundos estão ligados ao empreendedorismo por necessidade (Naudé, 2010). Entretanto, tal relação deve ser feita com cuidado, visto que tanto em um grupo quanto em outro encontramos negócios iniciados por oportunidade ou por necessidade. O uso ou não de trabalhadores por conta própria como equivalentes de empreendedores é objeto de debate na literatura e requer avanços teóricos e metodológicos que possibilitem o aprimoramento de análises.

Alguns estudos prévios foram realizados utilizando tais bases, como a aplicação do SINAN para identificar o perfil dos agravos que acomete empreendedores (Barbosa & Borges, 2021), o uso de dados da RFB para mapear o empreendedorismo por mulheres no estado de Goiás (Pagotto et al., 2020), o emprego de múltiplas bases nacionais para avaliar a associação de fatores socioeconômicos e a proporção da MEI em municípios mineiros (Morais et al., 2022), o uso da PNADc para investigar as características dos trabalhadores por conta própria (Rossi, 2018) a informalidade (Santiago & Vasconcelos, 2017) e a relação entre empreendedorismo e crescimento econômico (Barros & Pereira, 2008). Além destes exemplos que demonstram o uso das bases indicadas nos parágrafos precedentes, encontramos autores brasileiros que utilizam outras fontes de dados secundários na realização de pesquisas publicadas em revistas de grande impacto, como é o exemplo de Fischer et al. (2018), que utilizaram dados da FAPESP e do CNPq em estudo sobre empreendedorismo acadêmico.

Diante das potencialidades listadas acima, as próximas duas seções serão dedicadas à introdução ao software R, contemplando uma breve apresentação sobre o programa e, em sequência, a aplicação de uma análise exploratória resultante da conjugação de duas bases de dados, o GEM e o WGI.

O SOFTWARE R

Os pacotes estatísticos sempre foram aliados da pesquisa quantitativa em empreendedorismo. O software R está entre as ferramentas que têm alcançado popularidade nos últimos anos. O R é um ambiente e linguagem de programação dedicado primordialmente a análises estatísticas (Hornik, 2020). Diferente de softwares tradicionalmente utilizados nas ciências sociais aplicadas, como SPSS e Stata, o R é gratuito. Além disso, por ser uma ferramenta cuja interface possui formato de linguagem de programação e devido às funções incorporadas nas centenas de pacotes que podem ser instalados, carrega maior versatilidade em termos de funcionalidades. Ademais, as análises podem ser reproduzíveis, caso se tenha um script construído, o que contribui inclusive para a maior transparência da pesquisa, condição cada vez mais valorizada pela comunidade científica de empreendedorismo (Anderson et al., 2019; Maula & Stam, 2020).

Ao serem adicionados ao R, os pacotes carregam funções[2] com diferentes finalidades, como leitura e tratamento de dados, visualização e análises quantitativas. Tais pacotes na maior parte das vezes são criados e aprimorados por usuários do R, fazendo com que a própria comunidade contribua para o avanço constante da ferramenta. A Tabela 2 apresenta a relação de alguns pacotes do R e funcionalidades que contemplam. Cabe destacar que é uma lista longe de estar esgotada. O rol completo de pacotes, bem como as respectivas documentações podem ser acessados no site do RProject[3]. O R por si só pode ser usado por pesquisadores. No entanto, a linguagem é usualmente manipulada por meio do software RStudio®, que é um ambiente de desenvolvimento integrado, que possui uma interface mais intuitiva e que fornece melhor experiência de uso do R.

Tabela 2

Pacotes da Linguagem R

Pacotes	Funcionaliaddes
Leitura: readxl, vroom, foreign	O readxl permite leitura de arquivos em do MS Excel. O vroom carrega arquivos com maior volume de dados rapidamente. O pacote foreign possui funções que permitem a leitura de arquivos em formato de outros programas, como SPSS e Stata
Tratamento: dplyr, tidyr, lubridate	O pacote dplyr reúne um conjunto essencial de funções que permite filtrar, selecionar, agrupar, sumarizar e juntar duas ou mais bases de dados. O tidyr inclui funções para redimensionar sua base de dados, procedimento necessário para algumas análises visuais e estatísticas. O lubridate é dedicado a tratamentos que envolvem dados no formato de data/horário
Visualização: ggplot2, plotly, leaflet, DT, Shiny	O ggplot2 é a base para criação gráficos. O plotly permite a criação de gráficos interativos. O leaflet possui funções dedicadas à criação de mapas. O DT consegue gerar tabelas interativas. O Shiny permite a criação de aplicações web, como dashboards interativos
Análises quantitativas: survey, stats, tidymodels, psych, laavan	O pacote stats contempla muitas técnicas estatísticas, como testes para comparação de um ou mais grupos (teste-T, Wilcoxon, ANOVA) e regressões. O pacote survey geralmente é aplicado bases decorrentes de pesquisas de amostragem complexa (ex.: PNADc). O tidymodels é um metapacote que conjuga outros vários pacotes dedicados a executarem o workflow de algoritmos de machine learning. O psych possui implementações dedicadas à análise fatorial, por exemplo. Por fim, o laavan possui funções destinadas à execução de modelagem de equações estruturais
Análises textuais: Tidytext, wordcloud, syuzhet, rtweet, bibliometrix	O pacote tidytext possui um conjunto amplo de funções para tratamento de textos. O pacote wordcloud permite gerar nuvens de palavras. O syuzhet permite gerar análises de sentimento que classifica sentenças em valências positivas e negativas, assim como alguns sentimentos definidos por um dicionário léxicos. O rtweet é um pacote de suporte à extração de postagens do Twitter. Por fim, o bibliometrix é um pacote que fornece suporte para a condução de análises bibliométricas

Nota: Elaborado pelos autores.

UM BREVE TUTORIAL DE USO DO R PARA DADOS EM EMPREENDEDORISMO

Para esse tutorial, buscou-se realizar algumas operações de leitura, tratamento de dados e análise exploratória dos dados. Todavia, compreende-se que alguns dos ensinamentos aqui presentes – como o uso de joins – serão úteis para abrir horizontes de possibilidades à pesquisa em empreendedorismo ao permitir conjugar múltiplas bases de dados. Conforme apresentado na seção 3, estudos que usam o GEM frequentemente o conjugam a outras bases.

Para aplicar os conhecimentos é necessário que o R e RStudio® estejam instalados no computador ou o acesso à ferramenta RStudio Cloud[4]. Além disso, é importante acessar as planilhas que contém as bases aplicadas neste estudo de caso, bem como o dicionário de dados a partir do anexo deste documento. Os seguintes procedimentos serão realizados:

1. *Carregar pacotes necessários para o tratamento e análise dos dados;*
2. *Ler dados a partir de planilhas e formato comma separated values (.csv);*
3. *Inspecionar os dados;*
4. *Juntar duas bases de dados;*
5. *Realizar análise exploratória dos dados e visualizar dados*

Primeiramente, vamos carregar os pacotes e ler as bases que serão usadas no exemplo (Quadro 1). Para isso, serão utilizados duas bases que já foram conjugadas em um estudo prévio (Audretsch et al., 2021): o GEM, componente APS em formato agregado e o WGI. Conforme apresentado anteriormente, o base do GEM APS agregada contempla resultado de um levantamento de percepções sobre comportamentos e atitudes empreendedoras por país.

Quadro 1

```
# Instalando os pacotes que serão utilizados. Depois de instalar os pacotes, não há
# necessidade de executar esses códigos novamente
install.packages("readr") # data reading
install.packages("dplyr") # data processing
install.packages("ggplot2") # data visualization
install.packages("skimr") # descriptive data analysis
install.packages("GGally") # visual exploratory analysis
install.packages("ggrepel") # visual support


# Carregando pacotes que serão usados
library(readr) # data reading
library(dplyr) # data processing
library(ggplot2) # data visualization
library(skimr) # descriptive data analysis
library(GGally) # visual exploratory analysis
library(ggrepel) # visual support


# Lendo os bases de dados por meio da função read_csv e guardando nos objetos
# wgi e gem_aps
wgi <- read_csv("https://raw.githubusercontent.com/empreend/
    empreendedorismo/main/wgi.csv")
gem_aps <- read_delim("https://raw.githubusercontent.com/empreend/
    empreendedorismo/main/gem_2019_aps.csv", delim = ";")
```

Já a base do WGI traz dados sobre a qualidade de governança dos países, o que envolve a percepção sobre corrupção, aplicabilidade de leis, estabilidade política, dentre outros elementos. Para esta análise, foram selecionadas as seguintes variáveis de cada base, conforme Tabela 3.

Tabela 3

Grupo de variáveis

Variável	Base	Descrição
Economy	GEM, WGI	País
Continent	GEM	Continente
Entrepreneurship as a good career choice	GEM	Percentual da população entre 18 e 64 anos que concordam com a afirmativa: "no seu país, a maioria das pessoas considera que iniciar um empreendimento é um bom caminho de carreira"
Total early-stage Entrepreneurial Activity (TEA) Rate	GEM	Percentual da população entre 18 e 64 anos que é empreendedor nascente ou gerencia um novo negócio
Rule of Law	WGI	Percepção do grau em que agentes possuem confiança e cumprem as regras da sociedade, bem como a qualidade da execução de contratos, direitos de propriedade, polícia, judiciário
Regulatory quality	WGI	Percepção da habilidade do governo em formular e implementar políticas sólidas e regulação que permitam a promoção do desenvolvimento do setor privado
Political Stability	WGI	Percepção sobe a probabilidade de instabilidade ou tomada do poder por medidas inconstitucionais, violência, incluindo condições motivadas politicamente e terrorismo
Voice Accountability	WGI	Percepção sobre o quanto os indivíduos do país podem participar da seção dos governantes, ter liberdade de expressão/reunião e mídia livre

Nota: Dicionário de dados do GEM e WGI.

Não é escopo do tutorial aprofundar em aspectos de estatística inferencial ou machine learning, o que demandaria maior aprofundamento teórico para proposição de um modelo, bem como o nivelamento de conhecimentos em métodos quantitativos e testes de pressupostos de modelos estatísticos.

Note que, caso esteja usando o RStudio®, as bases estarão carregadas na aba Environment, usualmente localizada no canto superior direito do programa. Agora vamos inspecionar as variáveis por meio da função glimpse() do pacote dplyr. Observa-se que o resultado da função glimpse mostra que a base possui nove colunas (variáveis) e 202 linhas (observações). Em frente a cada variável, são apresentadas as primeiras observações de cada uma (cf. Quadro 2).

Quadro 2

```
# função glimpse serve para inspecionar a base, incluindo o número de
# observações, variáveis e tipos de variáveis

glimpse(wgi)
## Rows: 202
## Columns: 9
## $ country                <chr> "Yemen, Rep.", "Syrian Arab Republic"~
## $ code                   <chr> "YEM", "SYR", "AFG", "LBY", "IRQ", "S~
## $ corruption             <dbl> 0.8185391, 0.8114473, 1.0989244, 0. ~
## $ rule_of_law            <dbl> 0.7266536, 0.4239366, 0.7864730, 0.65~
## $ regulatory_quality     <dbl> 0.8360702, 0.7420965, 1.3794446, 0.15~
## $ gov_effectiveness      <dbl> 0.22057843, 0.78872073, 1.03612506, 0~
## $ political_stability    <dbl> -0.26829433, -0.22799635, -0.14940667~
## $ voice_accountability   <dbl> 0.7339933, 0.5201248, 1.5119677, 1.04~

glimpse(gem_aps)
## Rows: 50
## Columns: 18
## $ cod_pais                  <dbl> 374, 61, 375, 55, 101, 56, 86, ~
## $ economy                   <chr> "Armenia", "Australia", "Belaru~
## $ continent                 <chr> "Asia", "Oceania", "Europa", "A~
## $ abrev                     <chr> "ARM", "AUS", "BLR", "BRA", "C~
## $ year                      <dbl> 2019, 2019, 2019, 2019, 2019, 2~
## $ perceived_opportunities   <dbl> 53.9, 45.7, 29.5, 46.4, 67.1, 4~
## $ perceived_capabitilities  <dbl> 70.0, 56.0, 42.3, 62.0, 56.8, 7~
## $ fear_failure              <dbl> 48.2, 47.4, 38.0, 35.6, 47.2, 5~
## $ entrepreneurial_intentions <dbl> 32.2, 13.0, 6.6, 30.2, 11.9, 57~
## $ tea                       <dbl> 21.0, 10.5, 5.8, 23.3, 18.2, 36~
## $ established_ownership     <dbl> 7.84, 6.53, 2.72, 16.16, 7.44, ~
## $ entrepren_employee_Act    <dbl> 0.6, 8.3, 0.5, 0.6, 5.4, 3.6, 0~
## $ female_male_tea           <dbl> 0.6, 0.7, 0.8, 1.0, 0.7, 0.8, 0~
## $ high_job_creation_expect  <dbl> 30.5, 24.6, 28.2, 8.9, 21.2, 36~
## $ business_service_sector   <dbl> 7.6, 26.3, 10.2, 7.6, 12.2, 19.~
## $ high_status_success_entrp <dbl> 73.4, 74.0, 69.9, 72.3, 79.9, 7~
## $ entrepr_good_career_choice <dbl> 87.2, 56.4, 70.3, 75.3, 69.2, 7~
```

O objetivo agora é juntar as bases. Para isso é importante que ambas possuam alguma variável correspondente. A partir do dicionário de dados e da inspeção inicial usando a função glimpse(), percebe-se que as variáveis code e abrev são equivalentes nas bases wgi e gem_aps, respectivamente. A partir disso, aplicaremos a função left_join. No código abaixo, estamos informando o R (cf. Quadro 3) para juntar as bases gem_aps e wgi a partir das colunas abrev e code. Após isso, o resultado será armazenado em um objeto de nome gem_wgid.

Quadro 3

```
gem_wgid <- gem_aps %>%
        left_join(wgi, by = c("abrev" = "code"))
```

Em sequência serão selecionadas apenas as variáveis interessantes para o estudo por meio da função select do pacote dplyr (cf. Quadro 4).

Quadro 4

```
gem_wgid <- gem_wgid %>%
        select(economy, continent, entrepr_good_career_choice, tea, rule_of_law,
    regulatory_quality, political_stability, voice_accountability)
```

Agora será aplicada uma análise descritiva da base, usando a função skim do pacote skimr (cf. Quadro 5).

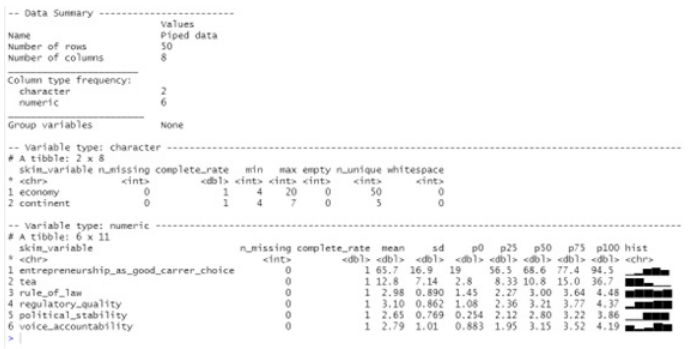
Quadro 5

```
gem_wgid %>%
  skim()
```

O resultado desta função traz um conjunto de medidas, como média, desvio-padrão, percentis e um histograma (cf. Figura 3).

Figura 3

Resultados da função skim()



Nota: Elaborado pelos autores.

Por fim, vamos aplicar a função ggpairs() do pacote GGally para visualizar uma matriz de correlação das variáveis (cf. Quadro 6). Como as variáveis de identificação do país (economy) e continente (continent) são categóricas, optou-se por retirá-las da análise aplicando a função select(-nome da variável). Os resultados estão dispostos na Figura 4.

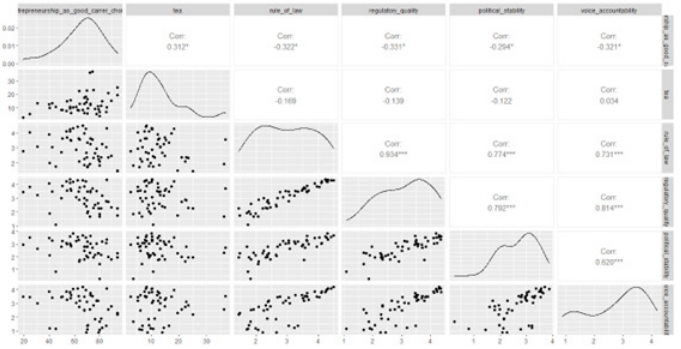
Quadro 6

```
gem_wgid %>%
  select(-economy,-continent) %>%
  ggpairs()
```

É possível identificar que a variável entrepreneurship as good career choice apresentou uma correlação negativa e significativa com as variáveis institucionais. Estas, por sua vez, foram altamente correlacionadas entre si, o que é de se esperar considerando os fenômenos que mensuram. Novamente, o presente estudo de caso se limita a aplicar funções da linguagem R para analisar dados e não pretende adentrar em aspectos teóricos.

Figura 4

Resultado da função ggpairs()



Nota: Elaborado pelos autores.

Por fim, vamos explorar um pouco mais a relação de duas variáveis - political stability e entrepreneurship as good career choice (cf. Quadro 7). Para isso, foi usada a função de visualização de dados ggplot. No primeiro parêntese, dentro do argumento aes, as coordenadas x e y são vinculadas às variáveis estabilidade política e empreendedorismo como boa escolha de carreira, respectivamente.

Quadro 7

```
gem_wgid %>%
  ggplot(aes(x = political_stability, y = entrepreneurship_as_good_career_choice))
    + geom_point(aes(col = continent, size = 1.5)) +
  geom_smooth(method = "lm", se = FALSE) +
  geom_text_repel(aes(label = economy)) +
  facet_grid(~continent) + theme_minimal() + ylab("Entrepreneurship as a good
    career of choice") + xlab("Political Stability")
```

Em seguida, temos que informar qual o formato de disposição dos dados: pontos (geom_point()) ou reta que descreve a relação (geom_smooth()). Em ambas as funções, podemos adicionar parâmetros (ex.: colorir os pontos conforme os continentes e aumentar o tamanho dos pontos para melhor visualização). A função seguinte – geom_text_repel() – adiciona textos a cada ponto com base na variável economy, enquanto o facet_grid divide os dados em vários painéis, com base na variável continentes. Finalmente, a função theme_minimal() adiciona um design minimalista. As funções xlab() e ylab() mudam os títulos dos eixos, conforme o texto que optamos. O resultado pode ser visto na Figura 5.

O presente tutorial pode ser acessado também em formato de vídeo no canal do Youtube® do Laboratório de Pesquisa em Empreendedorismo e Inovação da Universidade Federal de Goiás (LAPEI – UFG). No ano de 2021, o LAPEI-UFG – em parceria com a Associação Nacional de Estudos em Empreendedorismo e Gestão de Pequenas Empresas (ANEGEPE) e a Divisão Inovação, Tecnologia e Empreendedorismo da Associação Nacional de Pós-Graduação e Pesquisa em Administração (ITE-ANPAD) – promoveu um curso de R aplicado às pesquisas em empreendedorismo. O curso contou com três módulos e foi transmitido de modo síncrono. O curso obteve 161 inscrições, com participantes de diversas instituições de ensino e pesquisa de todo o Brasil. Até outubro de 2022, as gravações contavam com mais de 1500 visualizações no Youtube®. Ao final do treinamento, uma avaliação de reação foi aplicada; os módulos foram classificados entre "satisfatório" (35%) e "muito satisfatório" (65%). Como pontos fortes, os participantes destacaram a didática e a qualidade dos materiais disponibilizado. As oportunidades de melhoria envolveram a divisão de módulos mais curtos, com mais encontros, e em horários fora do turno comercial.

Figura 5

Relação entre variáveis Entrepreneurship as good career of choice e political stability

Nota: Elaborado pelos autores.

ABORDAGENS ANALÍTICAS PARA A PESQUISA EM EMPREENDEDORISMO

A análise dos dados deve ser condicionada à pergunta de pesquisa. Conforme os manuais de análise de dados, as técnicas podem ser divididas em interdependentes e dependentes e estão associadas ao tipo de relação estudada (Hair et al., 2009). As análises de interdependência possuem como objetivo reduzir, classificar e agrupar observações e/ou variáveis. Dentro desse grupo estão as técnicas de análise de agrupamento, análise de componentes principais e análise fatorial, por exemplo. Já as análises de dependência correspondem ao grupo de técnicas que busca estimar modelos que expressem a relação entre variáveis. Nesse sentido, encontram-se as variadas técnicas de regressão (ex.: linear, logística, multinomial, binomial negativa, quantílica) e de modelagem de equações estruturais, por exemplo (Favero & Belfiore, 2017). A seguir serão apresentados alguns estudos que utilizaram técnicas das duas perspectivas, dependência e interdependência.

Em uma pesquisa sobre valores culturais e prevalência de empreendedorismo social, Canestrino et al. (2020) aplicaram uma análise de agrupamento em uma das primeiras etapas da pesquisa para reunir aqueles países que possuíam características culturais mais próximas. Para isso, os pesquisadores utilizaram dados do *Global Leadership and Organizational Behavior Effectiveness* - GLOBE, projeto que coleta percepção de gestores em vários países acerca das dimensões culturais propostas por Hofstede. A partir disso, foi possível identificar três grupos com atributos relativamente semelhantes. O primeiro cluster foi dominado por países do norte europeu e foi cunhado como simpático, o segundo foi composto por países asiáticos e africanos e recebeu o rótulo de pragmático e o último grupo foi caracterizado como progressista e foi formado principalmente por países do sul europeu e américa latina.

Em um exemplo das técnicas de dependência, Benatti et al. (2021) buscou avaliar a relação entre o registro de MEI em municípios paulistas e diferentes indicadores de natureza econômica (Produto Interno Bruto Municipal – PIB-M – e Índice Firjan de Desenvolvimento Municipal – IFDM). Para isso, os autores levantaram dados de diferentes bases secundárias e aplicaram uma regressão quantílica em dois modelos: ambos com o registro de MEI como variável independente, mas diferindo nas variáveis

dependentes (PIB-M e IFDM). O estudo permitiu identificar que o MEI exerce maior efeito sobre municípios de menor tamanho e de faixas de baixo e médio crescimento do IFDM.

Outro exemplo de pesquisa que empregou técnicas de dependência foi o estudo realizado por Pagotto, Borges, et al. (2022) que utilizaram dados do PSED 2 para determinar a associação de diferentes formas de capital – humano, financeiro e social – no desenvolvimento de capacidades inovadoras em empresas nascentes. Dentre as variáveis investigadas, recursos financeiros pessoais, escolaridade, capital social utilizado para acessar infraestrutura física foram determinantes para o desenvolvimento das capacidades inovadoras ao longo do tempo em empresas nascentes (Pagotto, Borges, et al. 2022).

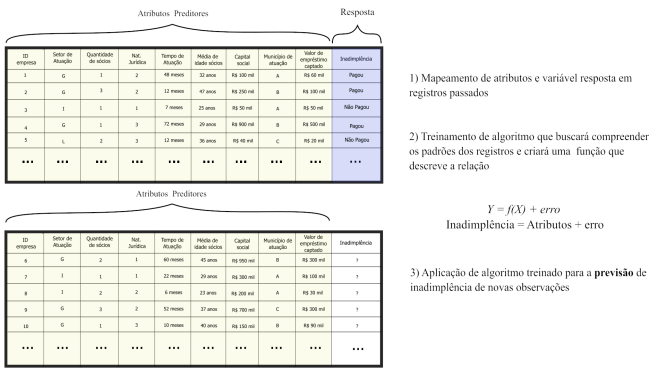
As técnicas mencionadas anteriormente têm se consolidado e se desenvolvido no contexto da estatística há certo tempo. Geralmente fazem parte dos currículos de cursos de Análise Multivariada em cursos de graduação e pós-graduação. No entanto, diante da difusão crescente de ferramentas de machine learning, pesquisadores da área de empreendedorismo têm estimulado a adoção dessa abordagem também nos estudos da área (Chalmers et al., 2021; Maula & Stam, 2020; Prüfer & Prüfer, 2020).

Apesar da estatística e *machine learning* terem alicerces em dados e inclusive técnicas semelhantes, os objetivos, meios e instrumentos das duas abordagens possuem algumas diferenças. Por um lado, a estatística foca primordialmente em inferência enquanto o machine learning possui maior ênfase sobre previsão (Bzdok et al., 2018). Outras características de distinção entre ambas as abordagens se desdobram a partir dessa classificação inicial e serão discutidas após o exemplo a seguir.

Entende-se por previsão a capacidade de se antecipar um resultado futuro com base em atributos presentes (James et al., 2013). Para fins de ilustração, imagine o caso, retratado na Figura 6, de um gestor de uma política pública que queira criar um modelo preditivo para avaliar se as empresas que estão recorrendo a uma linha de crédito retornarão o empréstimo concedido depois de três anos.

Figura 6

Criação de modelo preditivo para inadimplência de empresas



Nota: Elaborado pelos autores.

Para atender esse objetivo, o gestor poderá treinar e validar um modelo de machine learning a fim de identificar padrões em dados passados de empresas que passaram pela mesma situação. O algoritmo vai mapear o conjunto vasto de variáveis (ex.: número de empreendedores, gênero dos empreendedores, setor de atuação, capital social, natureza jurídica, localização, caráter familiar, dentre outras) e, partir disso, buscará padrões para formar uma função que descreve a relação. Uma prática usual após o treinamento do algoritmo é validar a capacidade de previsão dele em um particionamento do seu banco de dados original, chamado de banco de teste, para verificar se o modelo responde bem a um subconjunto dos dados que não participou da etapa de treinamento.

A partir da função desenvolvida com base na identificação dos padrões passados e tomados os devidos cuidados na etapa de validação, o gestor poderá ler um novo conjunto de dados que possui as características dos empreendimentos do seu território atualmente e assim prever a chance de pagarem o empréstimo depois do período desejado. Nota-se que o objetivo aqui é realizar previsão. Nessas circunstâncias, a função criada, a depender do algoritmo adotado, poderá ter baixa interpretabilidade. Assim, apesar de se compreender que ela consegue prever bem novas observações, o que está por trás nem sempre é de fácil compreensão.

Agora considerando a seguinte situação: o pesquisador quer obter mais interpretabilidade e compreender de que modo algumas variáveis afetam o pagamento do empréstimo. Nesse caso, o investigador estará trabalhando sob uma perspectiva de inferência, que está, tradicionalmente, mais associada à estatística (Bzdok et al., 2018).

A partir do exemplo, cabe destacar alguns desdobramentos da relação previsão/inferência. Métodos baseados em machine learning lidam melhor com a identificação de padrões em bases extensas, compostas por muitas variáveis, enquanto a estatística tem maior foco um conjunto menor de variáveis, mas com maior extensão de observações. Ademais, devido à flexibilidade para calcular padrões, alguns algoritmos de machine learning podem ter um bom poder preditivo ao criar funções sofisticadas que descrevem as relações investigadas, porém, podem fornecer baixa interpretabilidade, exigida para realizar inferências (Bzdok et al., 2018).

COMO AVANÇAR NA LITERATURA EM EMPREENDEDORISMO COM O SUPORTE DE FERRAMENTAS COMO O R

A presente seção reúne práticas que podem ser adotadas para o avanço da pesquisa quantitativa em empreendedorismo com o suporte de ferramentas como o R. Os pontos destacados nesta subseção são uma compilação de discussões de editorias sobre métodos quantitativos na pesquisa em empreendedorismo, como o uso de análises exploratórias de dados, medidas para aprimorar estudos quantitativos e a publicidade de análises.

Primeiramente, pesquisadores devem se apoiar mais em técnicas de análise exploratória dos dados. Normalmente, tais técnicas são recomendadas previamente à aplicação de modelagens multivariadas avançadas, uma vez que permitem a revelação de padrões na distribuição das variáveis, bem como a identificação de dados omissos ou outliers. No entanto, elas também são especialmente úteis para a elucidação de fenômenos pouco compreendidos. Dentre as técnicas de análise exploratória, inserem-se as análises descritivas (contemplando medidas além da média e desvio-padrão, como mínimo e máximo), análises de cluster, análise de componentes principais e identificação de padrões em ferramentas de visualização de dados. A aplicação de técnicas exploratórias de dados – como a modelagem de tópicos, clusterização, análise de redes - podem apresentar insights de grande valor para as pesquisas (Wennberg & Anderson, 2020).

Anderson et al. (2019) indicam três fatores que são importantes para o aprimoramento de artigos teórico-empíricos na área de empreendedorismo: 1) a pergunta de pesquisa que motiva o estudo; 2) as condições que contribuem para o aprimoramento de inferências causais e 3) os procedimentos adotados para minimizar a parcialidade dos pesquisadores. Em relação ao primeiro ponto, é importante que a pergunta de pesquisa esteja sustentada de modo qualificado e que o método seja adequado para respondê-la (Maula & Stam, 2020). Sobre o aprimoramento de inferências causais, há um estímulo para desenhos de pesquisas experimentais, a despeito da viabilidade de aplicação, inclusive devido ao rigor destas

para lidar com problemas de endogeneidade e a capacidade de demonstrar relações de causalidade (Anderson et al., 2019; Maula & Stam, 2020).

Um ponto a ser evitado é a "caçada dos asteriscos". Trata-se de um comportamento de pesquisadores onde existe um viés pela necessidade da obtenção de resultados significativos em análises, ocasionando práticas conhecidas como *p-hacking* e HARKing[5]. Mais que os asteriscos, deve-se buscar bons insights e procedimentos de pesquisa rigorosos, como enfatizam Anderson et al. (2019, p. 4) enfatizam, "*Pesquisadores podem publicar bons estudos de empreendedorismo, realizando perguntas interessantes e aplicando designs de pesquisa rigorosos independente de identificar resultados significativos*". Por outro lado, pesquisadores devem estar atentos ao tamanho do efeito identificado nos resultados dos modelos. Afinal, um p-valor significativo não é sinônimo de que a variável preditora possuirá um efeito prático relevante para a variação de uma variável dependente (Maula & Stam, 2020).

Outra boa prática que tem sido incentivada é a publicidade dos dados e códigos (Anderson et al., 2019; Maula & Stam, 2020). Plataformas como o Researchgate e Data Mendeley concedem *Document Object Identifier* (DOI) para as bases disponibilizadas por pesquisadores. Já o código das análises pode ser documentado por meio de ferramentas como Rmarkdown (formato de arquivo para R que permite gerar relatórios), Google Colab, Jupyter Notebook e Github.

AGENDA

Diante do que foi apresentado, compreende-se que pesquisadores da área de empreendedorismo podem aproveitar da crescente disponibilidade de dados bem como do ferramental de softwares cada vez mais versáteis e poderosos. Diante disso, a presente seção tem como objetivo levantar possíveis caminhos que pesquisadores podem trilhar a partir das discussões levantadas nesse editorial.

Conforme apresentado na seção 3, existe um grande volume de dados disponíveis de modo não estruturado. Nesse sentido, pesquisadores já têm explorado a potencialidade desse tipo de dado, realizando investigações a partir de dados de redes sociais (Obschonka et al., 2017; Pagotto, Barbosa, et al. 2022), grandes veículos de imprensa (Suarez et al., 2020), e plataformas de financiamento coletivo. Considerando esse tipo de dado e as pesquisas passadas, pesquisas futuras podem tentar responder algumas das seguintes perguntas: Quais são as representações que a grande mídia constrói sobre o empreendedorismo? Quais os discursos emitidos por veículos de grandes mídias sobre empreendedores e empresários no Brasil? (Suarez et al., 2020).

Muitos dos levantamentos nacionais como as pesquisas do IBGE não utilizam o termo empreendedor dentre as categorias laborais. Os dois grupos que apresentam maior relação com empreendedorismo são os trabalhadores por conta própria e os empregadores. Portanto, um segundo caminho de pesquisas seria aprofundar investigações acerca dos trabalhadores por conta própria. Na literatura internacional já se observa esforços para lançar um olhar mais aprofundado sobre essa categoria ocupacional; exemplo disso é a edição especial da Small Business Economics, em 2020, sobre o assunto (Burke & Cowling, 2020). No Brasil, apesar de incipiente, já existem experiências documentadas sobre o uso das bases do IBGE para a realização de pesquisas sobre empreendedorismo (e.g., Almeida et al., 2017).

De acordo com discussões nacionais e internacionais, sabe-se que os trabalhadores por conta própria são um perfil crescente (IBGE, 2021), diversificado – nas palavras de Santiago e Vasconcelos (2017), vai do catador ao doutor – (Burke & Cowling, 2020; Moortel & Vanroelen, 2017) e, em média, mais vulnerável que trabalhadores empregados (Gindling & Newhouse, 2013; Santiago & Vasconcelos, 2017). Portanto, mais pesquisas que busquem compreender esse perfil, seu contexto e seu processo empreendedor são necessárias. A partir de bases de dados brasileiras que contemplam o perfil ocupacional, estudos podem ser realizados para segmentar melhor o trabalhador por conta própria brasileiro.

Outra avenida de pesquisas futuras envolve a exploração da potencialidade dos dados para investigar o fenômeno empreendedor em diferentes níveis e a exploração em uma perspectiva multinível. Bases da RFB, o ICE e a MUNIC podem ser associadas a outras bases também a nível municipal para explorar o efeito de variáveis contextuais e institucionais – como segurança, cultura e leis – sobre o empreendedorismo (Muñoz-Fernández et al., 2019). O estudo de Morais et al. (2022) é um exemplo do emprego desta estratégia ao conjugar bases de diversas fontes (ex.: FIRJAN, RAIS, CAGED, DATASUS, INEP, IBGE) para avaliar a associação de variáveis socioeconômicas (ex.: renda, escolaridade, saúde) e a proporção de MEI a nível municipal. Considerando o nível país, bases como o GEM podem ser aplicadas para compreender melhor a associação do empreendedorismo também a outras condições contextuais, como democracia (Audretsch & Moog, 2020).

Declaração de Conflito de Interesse

Os autores declaram não existir conflito de interesses.

Notas de fim:

1 Acessível em: https://datasebrae.com.br.
2 Rotinas prontas criadas pelos desenvolvedores dos pacotes.
3 https://cran.r-project.org/web/packages/available_packages_by_name.html
4 https://rstudio.cloud.
5 Prática associada à escolha de variáveis a partir de análises prévias com resultados significativos.

Declaração de contribuições individuais dos autores

Papéis	Contribuição por autor	
	Pagotto DP	Borges C
Conceitualização	■	■
Metodologia	■	■
Software	■	
Validação	■	■
Análise formal	■	■
Pesquisa / Levantamento	■	
Recursos	N.A.	
Curadoria dos dados	N.A.	
Escrita - Rascunho original	■	■
Escrita - Revisão e edição	■	■
Visualização dos dados	N.A.	
Supervisão / Orientação	■	■
Administração do Projeto	N.A.	
Financiamento	N.A.	

REFERÊNCIAS

Almeida, F. M., Valadares, J. L., & Sediyama, G. A. S. (2017). A Contribuição do Empreendedorismo para o Crescimento Econômico dos Estados Brasileiros. *REGEPE - Revista de Empreendedorismo e Gestão de Pequenas Empresas*, 6(3), 466–494. https://doi.org/10.14211/regepe.v6i3.552

Anderson, B. S., Wennberg, K., & McMullen, J. S. (2019). Editorial: Enhancing quantitative theory-testing entrepreneurship research. *Journal of Business Venturing*, 34(5), 105928. https://doi.org/10.1016/j.jbusvent.2019.02.001

Arenius, P., Engel, Y., & Klyver, K. (2017). No particular action needed? A necessary condition analysis of gestation activities and firm emergence. *Journal of Business Venturing Insights*, 8(June), 87–92. https://doi.org/10.1016/j.jbvi.2017.07.004

Audretsch, D. B., Belitski, M., Chowdhury, F., & Desai, S. (2021). Necessity or opportunity? Government size, tax policy, corruption, and implications for entrepreneurship. *Small Business Economics*. https://doi.org/10.1007/s11187-021-00497-2

Audretsch, D. B., & Moog, P. (2020). Democracy and Entrepreneurship. *Entrepreneurship: Theory and Practice*, 1–25. https://doi.org/10.1177/1042258720943307

Barbosa, R., & Borges, C. (2021). A Saúde do Empreendedor no Brasil: Uma Análise dos Dados do Sistema de Informação de Agravos de Notificação (SINAN). *Future Studies Research Journal: Trends and Strategies*, 13(1), 28–41. https://doi.org/10.24023/futurejournal/2175-5825/2021.v13i1.532

Barros, A. A., & Pereira, C. M. M. A. (2008). Empreendedorismo e Crescimento Econômico: uma Análise Empírica. *Revista de Administração Contemporânea*, 12(4), 975–993. https://doi.org/10.1590/S1415-65552008000400005

Benatti, L. N., da Silva, E. E., & Prearo, L. C. (2021). Microempreendedores individuais e o desenvolvimento econômico nos municípios paulistas de 2010 a 2014. *REGEPE - Revista de Empreendedorismo e Gestão de Pequenas Empresas*, 10(2), e1676-e1676. https://doi.org/10.14211/regepe.e1676

Burke, A., & Cowling, M. (2020). On the critical role of freelancers in agile economies. *Small Business Economics*, 55(2), 393–398. https://doi.org/10.1007/s11187-019-00240-y

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, 15(4), 233–234. https://doi.org/10.1038/nmeth.4642

Canestrino, R., Ćwiklicki, M., Magliocca, P., & Pawełek, B. (2020). Understanding social entrepreneurship: A cultural perspective in business research. *Journal of Business Research*, 110(July 2019), 132–143. https://doi.org/10.1016/j.jbusres.2020.01.006

Chalmers, D., MacKenzie, N. G., & Carter, S. (2021). Artificial Intelligence and Entrepreneurship: Implications for Venture Creation in the Fourth Industrial Revolution. *Entrepreneurship: Theory and Practice*, 45(5), 1028–1053. https://doi.org/10.1177/1042258720934581

Cooper, D. R., & Schindler, P. S. (2014). *Business Research Methods*. In Business Research Methods (12th ed.). McGraw-Hill Irwin.

Cristi, M. A. A. (2018). Los métodos positivista y fenomenológico, una explicación desde las ciencias naturales y sociales. *Revista Pesquisa Qualitativa*, 6(12), 541. https://doi.org/10.33361/rpq.2018.v.6.n.12.219

Dyer, W. G., Dyer, W. J., & Gardner, R. G. (2013). Should My Spouse Be My Partner? Preliminary Evidence From the Panel Study of Income Dynamics. *Family Business Review*, 26(1), 68–80. https://doi.org/10.1177/0894486512449354

Favero, L. P., & Belfiore, P. (2017). *Manual de análise de dados*: estatística e modelagem multivariada com Excel®, SPSS® e Stata® (1st ed.). GEN.

Ferreira, A. D. M. F., Loiola, E., & Guedes Gondim, S. M. (2020). Produção científica em empreendedorismo no Brasil: uma revisão da literatura de 2004 a 2020. *Gestão & Planejamento-G&P*, 21. https:// doi.org/10.21714/2178-8030gep.v.21.5618

Fischer, B. B., Schaeffer, P. R., Vonortas, N. S., & Queiroz, S. (2018). Quality comes first: university-industry collaboration as a source of academic entrepreneurship in a developing country. *The Journal of Technology Transfer*, 43(2), 263-284. https://doi.org/10.1007/s10961-017-9568-x

Gil, A. C., & Silva, S. P. M. (2015). O Método Fenomenológico na Pesquisa sobre Empreendedorismo no Brasil. *Revista de Ciências da Administração*, 17(41), 99–113. https://doi.org/10.5007/2175-8077.2015v17n41p99

Gindling, T. H., & Newhouse, D. (2013). Self-Employment in the Developing World. In *Background Paper to the 2013 World Development Report* (Issue September 2012). https://doi.org/10.1016/j.worlddev.2013.03.003

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise Multivariada de Dados*. Bookman.

Hechavarría, D. M., & Ingram, A. E. (2019). Entrepreneurial ecosystem conditions and gendered national-level entrepreneurial activity: a 14-year panel study of GEM. *Small Business Economics*, 53(2), 431–458. https://doi.org/10.1007/s11187-018-9994-7

Hornik, K. (2020). *R FAQ*. Frequently Asked Questions on R. https://cran.r-project.org/doc/FAQ/R-FAQ.html#What-is-R_003f

Hox, J. J., & Boeije, H. R. (2005). Data Collection, Primary vs. Secondary. In *Encyclopedia of Social Measurement* (pp. 593–599). https://doi.org/10.1016/B0-12-369398-5/00041-4

IBGE. (2021). *Indicadores IBGE - Pesquisa Nacional por Amostra de Domicílios Contínua* - Quarto Semestre de 2020.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* - with Applications in R. Springer.

Kaminski, J. C., & Hopp, C. (2020). Predicting outcomes in crowdfunding campaigns with textual, visual, and linguistic signals. *Small Business Economics*, 55(3), 627–649. https://doi.org/10.1007/s11187-019-00218-w

Koumbarakis, P., & Volery, T. (2022). Predicting new venture gestation outcomes with machine learning methods. *Journal of Small Business Management*, 1-34. https://doi.org/10.1080/00472778.2022.2082453

Kwapsiz, A., & Hechavarria, D. (2018). Women don't ask: an investigation of startup financing and gender. *Venture Capital*, 20(2), 159–190. https://doi.org/10.1080/13691066.2017.1345119

Maula, M., & Stam, W. (2020). Enhancing Rigor in Quantitative Entrepreneurship Research. *Entrepreneurship: Theory and Practice*, 44(6), 1059–1090. https://doi.org/10.1177/1042258719891388

McDonald, S., Gan, B. C., Fraser, S. S., Oke, A., & Anderson, A. R. (2015). A review of research methods in entrepreneurship 1985-2013. *International Journal of Entrepreneurial Behavior & Research*. https://doi.org/10.1108/IJEBR-02-2014-0021

Moortel, D. D., & Vanroelen, C. (2017). *Classifying self-employment and creating an empirical typology*. http://hdl.voced.edu.au/10707/449386.

Morais, M. C. A., Emmendoerfer, M. L., Vitória, J. R., Mendes, W. A. Socioeconomic determinants of the individual micro-entrepreneur (IME). *REGEPE - Revista de Empreendedorismo e Gestão de Pequenas Empresas*, 11(3), e2070. https://doi.org/10.14211/ibjesb.e2070

Muñoz-Fernández, Á., Assudani, R., & Khayat, I. (2019). Role of context on propensity of women to own business. *Journal of Global Entrepreneurship Research*, 9(1). https://doi.org/10.1186/s40497-019-0160-8

Nassif, V. M. J., Silva, N. B., Ono, A. T., Bontempo, P. C., & Tinoco, T. (2010). Empreendedorismo: área em evolução? Uma revisão dos estudos e artigos publicados entre 2000 e 2008. *RAI-Revista de Administração e Inovação*, 7(1), 175-193.

Naudé, W. (2010). Promoting Entrepreneurship in Developing Countries: *Policy Challenges* (Issue 4).

Neergaard, H., & Ulhoi, J. P. (2007). Introduction: Methodological variety in enterpreneurship research. In *Handbook of Qualitative Research Methods in Entrepreneurship* (pp. 1–14). https://doi.org/10.1108/GM-04-2013-0043

Obschonka, M., & Audretsch, D. B. (2020). Artificial intelligence and big data in entrepreneurship: a new era has begun. *Small Business Economics*, 55(3), 529–539. https://doi.org/10.1007/s11187-019-00202-4

Obschonka, M., Fisch, C., & Boyd, R. (2017). Using digital footprints in entrepreneurship research: A Twitter-based personality analysis of superstar entrepreneurs and managers. *Journal of Business Venturing Insights*, 8, 13–23. https://doi.org/10.1016/j.jbvi.2017.05.005

Oliveira Junior, A. B. D., Gattaz, C. C., Bernardes, R. C., & Iizuka, E. S. (2018). Pesquisa em empreendedorismo (2000-2014) nas seis principais revistas brasileiras de administração: lacunas e direcionamentos. *Cadernos EBAPE.BR*, 16, 610-630. https://doi.org/10.1590/1679-395167644

Pagotto, D., Barbosa, R., Borges, C., & Nassif, V. (2022). Sentimentos Negativos de Empreendedores e a Covid-19: Uma Análise de Tweets. *Revista Inteligência Competitiva*, 12(1), e0414-e0414. https://doi.org/10.24883/IberoamericanIC.v12i.2022.e0414

Pagotto, D. P., Borges, C. V., Almeida, M. I. S., Hoffmann, V. E. (2022). Forms of Capital, innovation capability and innovation in new ventures. *REGEPE - Revista de Empreendedorismo e Gestão de Pequenas Empresas*, 11(2), 1-11, https://doi.org/10.14211/regepe.e1952

Pagotto, D., Teixeira, D. M., Miranda Filho, S. S., Borges, C., & Arantes, F. P. (2020). A evolução do empreendedorismo por mulheres em Goiás. In *Perfil da Empreendedora Goiana* - o empreendedorismo por mulheres e seus desafios (pp. 1–102). Sebrae - Goiás.

Prüfer, J., & Prüfer, P. (2020). Data science for entrepreneurship research: studying demand dynamics for entrepreneurial skills in the Netherlands. *Small Business Economics*, 55(3), 651–672. https://doi.org/10.1007/s11187-019-00208-y

Reynolds, P. D., Curtin, R. T. Business Creation in the United States: Panel Study of Entrepreneurial Dynamics II Initial Assessment. *Foundations and Trends® in Entrepreneurship*, v. 4, n. 3, p. 155–307, 2008.

Reynolds, P. D., Hechavarria, D., Tian, L. R., Samuelsson, M., & Davidsson, P. (2016). Panel study of entrepreneurial dynamics: A five cohort outcomes harmonized data set. *Res. Gate,* Revision 1, 1–48. https://doi.org/10.13140/RG.2.1.2561.7682

Rossi, M. de F. P. (2018). O trabalhador por conta própria: empreendedorismo e autoemprego na Região Metropolitana de Belo Horizonte/MG. *Revista Ciências Do Trabalho*, 12, 37–53.

Santiago, C. E. P., & Vasconcelos, A. M. N. (2017). Do catador ao doutor: Um retrato da informalidade do trabalhador por conta própria no Brasil. *Nova Economia*, 27(2), 213–246. https://doi.org/10.1590/0103-6351/2588

Semrau, T., & Hopp, C. (2016). Complementary or compensatory? A contingency perspective on how entrepreneurs' human and social capital interact in shaping start-up progress. *Small Business Economics*, 46(3), 407–423. https://doi.org/10.1007/s11187-015-9691-8

Suarez, J. L., White, R. W., Parker, S. C., & Jiménez-Mavillard, A. (2020). Entrepreneurship bias and the mass media: evidence from big data. *Academy of Management Discoveries*, 7(2), 1–49. https://doi.org/10.5465/amd.2018.0177

von Bloh, J., Broekel, T., Özgun, B., & Sternberg, R. (2020). New(s) data for entrepreneurship research? An innovative approach to use Big Data on media coverage. *Small Business Economics*, 55(3), 673–694. https://doi.org/10.1007/s11187-019-00209-x

Warhuus, J. P., Frid, C. J., & Gartner, W. B. (2021). Ready or not? Nascent entrepreneurs' actions and the acquisition of external financing. *International Journal of Entrepreneurial Behaviour and Research*, 27(6), 1605–1628. https://doi.org/10.1108/IJEBR-09-2020-0586

Wennberg, K., & Anderson, B. S. (2020). Editorial: Enhancing the exploration and communication of quantitative entrepreneurship research. *Journal of Business Venturing*, 35(3), 105938. https://doi.org/10.1016/j.jbusvent.2019.05.002

BIOGRAFIAS DOS AUTORES

Daniel do Prado Pagotto *é mestre em Administração pela Universidade Federal de Goiás e atualmente é doutorando em Administração pelo Programa de Pós-Graduação em Administração da Universidade de Brasília (PPGA-UnB). Atua como Coordenador Adjunto do Laboratório de Pesquisa em Empreendedorismo e Inovação da UFG (LAPEI-UFG). Suas pesquisas envolvem as áreas de empreendedorismo e inovação em serviços, com ênfase no uso de métodos quantitativos e dados secundários. Suas pesquisas vem sendo publicadas em revistas como Revista de Inteligência Competitiva, Humanidades & Inovação, REGEPE e eventos.*

E-mail: danielppagotto@gmail.com.

Cândido Borges *é professor de empreendedorismo da Universidade Federal de Goiás, Goiânia, Brasil, onde também é diretor do Laboratório de Pesquisa em Empreendedorismo e Inovação (LAPEI-UFG) e professor do Programa de Pós-Graduação em Administração da UFG. Ele obteve seu Ph.D. pela HEC Montréal e Pós-Doc pela EAESP-FGV. Sua pesquisa atual se concentra em novos empreendimentos, autoemprego e política de empreendedorismo. Seus artigos foram publicados em periódicos como Future Studies Research Journal, Humanidades & Inovação, REGEPE, RAUSP, entre outros.*

E-mail: candidoborges@gmail.com.