REGEPE ENTREPRENEURSHIP AND SMALL BUSINESS JOURNAL

PUBLISHER ANEGEPE Associação Nacional de Estudos em Empreendedorismo e Gestão de Pequenas Empresas

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

Editorial

Open access

Check for updates

Quantitative research in entrepreneurship using the R software for data analysis

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

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

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

Editorial Details

Article History
Available online : *May 02, 2023*

JEL Code: *C00, L26*

Article ID: *2384*

Editor-in-Chief[1] or Adjunct[2]:
[1] *Dr. Edmundo Inácio Júnior*
Universidade Estadual de Campinas, UNICAMP

Associate Editor:
Dra. Rose Mary Almeida Lopes
ANEGEPE

Executive[1] or Assistant[2] Editor:
[2] *M. BA. João Paulo Moreira Silva*

Translation / Proofreading:
Nina Mesquita Petek

Item relacionado (isTranslationOf):
https://doi.org/10.14211/regepe.esbj.e2257

How to cite:
Pagotto, D. do P., & Borges, C. (2023). Quantitative research in entrepreneurship using R software for data analysis. *REGEPE Entrepreneurship and Small Business Journal*, 12(2), e2384. https://doi.org/10.14211/regepe.esbj.e2384

***Corresponding author:**

Daniel do Prado Pagotto
danielppagotto@gmail.com

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.

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

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.

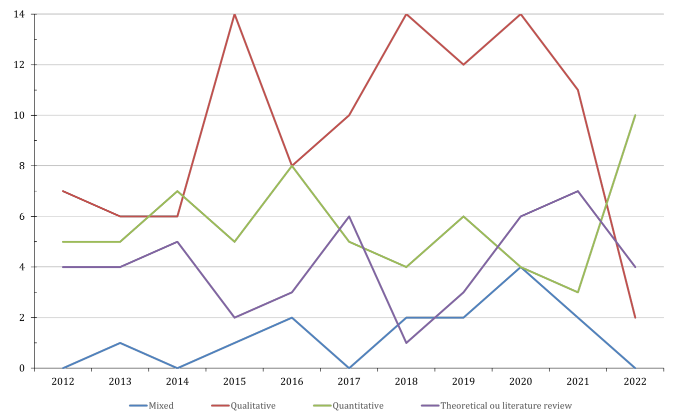
Palavras-chave: Métodos quantitativos. Software R. Dados secundários.

INTRODUCTION

The qualitative approach is the most used method in entrepreneurship research in Brazil. An analysis of papers published on Revista de Empreendedorismo e Gestão de Pequenas Empresas (REGEPE) between 2012 and 2022 found a higher number of articles which followed a qualitative approach - 104 of the former and 62 of the latter. Furthermore, as illustrated in Figure 1, only in the last year there was an inversion on the prevalence of qualitative over quantitative methods.

Figure 1

Evolution of publications by approach



Note: Elaborated by the authors.

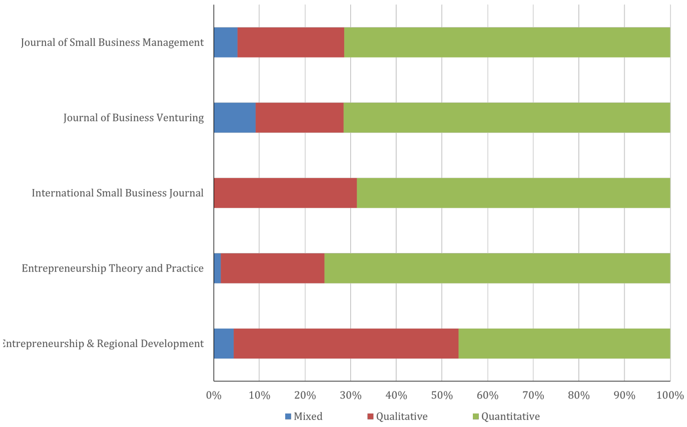
Various reviews and bibliometric studies published in Brazil recently came to the same conclusion. In a review of the publications of the Encontro de Estudos Sobre Empreendedorismo e Gestão de Pequenas Empresas - EGEPE and the Encontro Nacional da Associação de Pós-Graduação e Pesquisa em Administração (Enanpad) between 2000 and 2008, Nassif et al. (2010) discovered a predominance of studies that employed qualitative methods: 60.7% of the 219 theoretical-empirical articles were qualitative. Oliveira et al. (2018) investigated entrepreneurship articles published in six management journals between 2000 and 2014 and noticed a prevalence of publications using qualitative methodologies. A total of 54 empirical studies were included in his search, with 51.9% being qualitative, 11.1% being mixed, and 37% being quantitative. Ferreira et al. (2020) revealed that 44% of the 179 articles published between 2004 and 2020 were qualitative, 27% quantitative, and 25% theoretical.

This characteristic distinguishes national research from the international field, where quantitative studies predominate. McDonald et al. (2015) conducted a survey in six of the major international journals on entrepreneurship from 1985 to 2013 and found that, in a sample of 3749 papers, the majority (55%) employed a quantitative approach. An updated analysis on the same journals used by McDonald et al. (2015) presented the same pattern. Except for Entrepreneurship & Regional Development Journal, the others publish quantitative studies more frequently. Besides, in a universe of 362 empirical articles, 69.06% were of a quantitative nature (see Figure 2).

The greater number of qualitative studies is not necessarily a problem. Qualitative research is essential for the development of scientific knowledge in applied social sciences (Cristi, 2018), and it is no different in the field of entrepreneurship (Gil & Silva, 2015; Neergaard & Ulhoi, 2007). The issue is that the quantitative approach in entrepreneurship research has a history of low use in the national territory. The lack of quantitative research may prevent researchers from leveraging the benefits that this approach can provide, such as the ability to cover representative samples to validate theories developed and explored initially through qualitative methods, and generalization through sample designs and appropriate analysis techniques (Cooper & Schindler, 2014). Failure to advance quantitative research may represent a barrier to the development of the entrepreneurship field in Brazil.

Figure 2

Proportion of studies by type and journal



Note: Elaborated by the authors.

The reason for the low number of quantitative publications is multifaceted. However, one possible explanation could be the limited use of secondary data. In fact, considering the survey carried out at REGEPE, cited above, in quantitative or mixed methods research, almost twice as many studies used primary data - 49 used primary and 27 used secondary data. This finding was consistent with that found by Oliveira Junior et al. (2018), who discovered that 79.6% of studies employed primary surveys in a universe of 54 national empirical articles of entrepreneurship. Secondary data can increase the quantity and quality of quantitative research by avoiding the costs and time associated with collecting primary data and allowing studies with a greater number of observations (Hox & Boeije, 2005).

Knowing the available databases, as well as mastering the tools and methodologies for accessing and using these bases could improve the use of secondary data sources in research. In this regard, this article presents two contributions to the advancement of entrepreneurship research using a quantitative approach and secondary data: First, public national and international databases for entrepreneurship research are presented; second, a tutorial on the use of Software R for data in entrepreneurship is introduced.

SECONDARY DATA SOURCES FOR ENTREPRENEURSHIP RESEARCH

Obtaining data for entrepreneurship research is difficult, as it is in many other areas of Social Sciences. On the one hand, there is a scarcity of secondary sources on the early stages of business creation. On the other hand, from the standpoint of primary surveys, accessing entrepreneurs is tricky because they are busy individuals whose businesses are constantly changing, making it difficult to capture all the phenomena (Maula & Stam, 2020).

However, in recent years, there has been an increase in the amount of data collected from different sources, such as database records, web scraping techniques, and videos (Maula & Stam, 2020; Obschonka & Audretsch, 2020). Because they use techniques for extracting, interpreting, and analyzing unstructured data, the last two sources - web scraping and videos - have great potential for qualitative and quantitative research (e.g., texts, images, videos). The first source is characterized by the data's structured nature. Each will be discussed in greater detail below.

Structured data are the most used secondary sources in entrepreneurship research. There are governmental and institutional databases, company records, and surveys in this universe devoted primarily to the study of the entrepreneurial phenomenon (for example, the Panel Study of Entrepreneurial Dynamics (PSED) and the Global Entrepreneurship Monitor (GEM)). Benatti et al. (2021), for example, used data from the Microempreendedor Individual (MEI), extracted from the Data Sebrae repository[1], to assess the effect of this category of business on economic development of São Paulo state's municipalities. Audretsch et al. (2021) used a variety of data sources, including the GEM, to investigate the impact of institutional variables on opportunity and necessity entrepreneurship at the national level. Section 3 will cover a few of these databases in greater detail.

Exploration of unstructured sources - texts and videos - is unusual in entrepreneurship research, but it is becoming more common as data analysis tools, personal computer processing power, and cloud processing technologies advance. Web scraping is a method of collecting and extracting information from web pages (Prüfer & Prüfer, 2020). Obschonka et al. (2017), for example, investigated personality traits of "superstar" entrepreneurs and managers using data from Twitter user publications. Pagotto, Barbosa, et al. (2022) used Twitter to analyze the sentiment associated with tweets from entrepreneurs during the early stages of the Covid-19 pandemic. Previous experiences include extracting and analyzing texts from mainstream media such as The New York Times and Financial Times to assess differences in the content of publications about entrepreneurs (Suarez et al., 2020), as well as audio and video analysis of crowdfunding platforms to predict the success of fundraising campaigns (Kaminski & Hopp, 2020).

Considering the two sources mentioned, structured data investigations have been a reality in entrepreneurship research for some time. Unstructured data analysis, on the other hand, is already possible because there are programs with user-friendly interfaces that perform unstructured data processing, as well as libraries in programming languages such as R and Python. Thus, unstructured data analysis is viewed as a novel approach to measuring and comprehending phenomena in the field of entrepreneurship (Maula & Stam, 2020; von Bloh et al., 2020).

This presentation demonstrates that the analysis of unstructured data proves a promising path due to a number of factors, including increased data availability, amplification of the processing power of domestic and remote machines, traditional qualitative research software with more advanced textual analysis features (e.g., Nvivo, Atlas.ti), and packages in programming languages dedicated to this purpose. Data analysis tools and software that were previously used primarily in quantitative research can now be added to qualitative approaches, particularly when working with unstructured data, occasionally contributing to a narrowing between both perspectives and thus strengthening the investigation in the field. Despite the innovative nature of unstructured data, as demonstrated in section 2, we still have many opportunities to study entrepreneurship phenomena using structured secondary data.

STRUCTURED DATABASES IN ENTREPRENEURSHIP

The objective of this section is to present a few secondary databases on entrepreneurship, demonstrating their relevance and potential, and providing examples of their use in scientific studies. Table 1 lists some of the entrepreneurship bases. The foundations of PSED, GEM, and different national bases of governmental organizations are briefly described in the following paragraphs.

The Panel Study of Entrepreneurial Dynamics (PSED) was a survey led by Paul Reynolds with the purpose of collecting panel data from a representative sample of American entrepreneurs. Two editions of the PSED were conducted in the United States, and the most recent version, PSED 2, included the monitoring of emerging business through interview rounds conducted between 2006 and 2011. The PSED's main distinguishing feature is its longitudinal design, which makes possible to map various activities along the entrepreneurial process, such as identifying the opportunity, legalizing the company, making the first sale, and reaching the financial break-even point (Reynolds & Curtin, 2008). PSED data and supporting materials are freely available on the page: http://www.psed.isr.umich.edu/psed.

Table 1

Databases for entrepreneurship research

Databases	Sample Profile
Panel Study of Entrepreneurial Dynamics	US nascent stage entrepreneurs. Variations are found in other countries such as Australia, Sweden
Global Entrepreneurship Monitor	Data by country on conditions to undertake and attitudes towards entrepreneurial activity
Global Accelerator Learning Initiative	Companies that have gone through business accelerator programs
Brazilian Federal Revenue Service	Contains the national registration of legal entities (CNPJ) of companies and their partners
IBGE Bases of the Integrated Household Survey System	Surveys on characteristics of the Brazilian population, including breakdowns by groups such as employers and self-employed. Examples of databases: Continuous National Household Sample Survey (PNADc), National Health Survey (PNS), Basic Municipal Information Survey (MUNIC), Agro Census

Note: Elaborated by the authors.

Other countries, such as Australia, China, and Sweden, have undertaken initiatives similar to the PSED, allowing the creation of a single harmonized database containing observations from all of these surveys (Arenius et al., 2017; Reynolds et al., 2016; Warhuus et al., 2021). For several reasons, the PSED base or surveys derived from it have great potential for further research, such as: 1) researchers have encouraged longitudinal studies of entrepreneurship because company formation is a dynamic process (Maula & Stam, 2020); 2) the databases contain a wide range of data on topics such as entrepreneur characteristics, the entrepreneurial process, the nascent company, financing, business strategies, social capital, community support, and motivations. Because of this large amount of data, the foundation has spread in entrepreneurship studies on topics such as family entrepreneurship (Dyer et al., 2013), social capital (Semrau & Hopp, 2016), and female entrepreneurship (Kwapsiz & Hechavarria, 2018), predicting the emergence and abandonment of new ventures (Koumbarakis & Volery, 2022), among others.

The Global Entrepreneurship Monitor (GEM) is a survey released in 1999 in multiple countries with the objective of tracking aspects of the population's entrepreneurial attitudes and behavior, as well as the perception of contextual conditions for entrepreneurship. These two dimensions of GEM analysis are translated into two annual surveys: 1) the Adult Population Survey, which investigates questions related to the adult population's perception of identifying business opportunities in their locality, perception of the capabilities to start a business, and initial entrepreneurship rate, among others; and 2) the National Expert Survey, which is a survey aimed at specialists to capture the perception of variables in the entrepreneurial context, such as funding for entrepreneurship, government support, taxation, and bureaucracy, among others. It should be noted that the GEM also publishes reports and studies on a variety of topics, including social entrepreneurship, family entrepreneurship, and female entrepreneurship.

The GEM bases, like the PSED, are widely used in entrepreneurship studies, and are frequently combined with other surveys, which broadens the investigation of entrepreneurship in relation to other phenomena. Here are two examples of database composition: Hechavarra and Ingram (2019) connected the GEM to World Bank databases to investigate the impact of the ecosystem on male and female entrepreneurship prevalence; Audretsch et al. (2021) used a combination of multiple data sources - Worldwide Governance Indicators (WGI), International Monetary Fund government spending, and GEM - to assess the impact of national institutions on the rate of entrepreneurship by necessity and opportunity. Section 5 will include an exercise for connecting the GEM to another base.

Although they are not solely dedicated to entrepreneurship research, certain government databases in Brazil hold great promise for the country's researchers. Some of them are listed below; they are publicly available and can be used to build quantitative studies using secondary data.

The Pesquisa Nacional por Amostra de Domicílio (PNADc), the Pesquisa Nacional de Saúde (PNS), the Censo Agropecuário, and the Pesquisa de Informações Básicas Municipais (MUNIC) are all conducted by the Brazilian Institute of Geography and Statistics (IBGE). Moreover, the Federal Revenue Service of Brazil (RFB) publishes data on the National Register of Legal Entities (CNPJ). Through the Department of Informatics of the Unified Health System (DATASUS), the Ministry of Health compulsorily consolidates a list of diseases that affect the population through the Sistema de Informações de Agravos e Notificações (SINAN), as well as variables on morbidity and mortality, which are listed, respectively, on the Sistema de Informações Hospitalares (SIH) and the Sistema de Informações sobre Mortalidade (SIM). The Ministry of Education compiles datasets on education at the municipal level. In collaboration with Endeavor, the National School of Public Administration (ENAP) recently published the latest Entrepreneurial Cities Index (ICE - 2020), which includes data on the 100 largest Brazilian municipalities. Some of these databases disaggregate data at the municipality level (for example, MUNIC and ICE), while others disaggregate data at the individual level (e.g., SINAN).

When dealing with individual bases, an observation must be made before presenting examples of their application: in some of these surveys, entrepreneurs can be identified as employers or self-employed workers. The former is frequently associated with opportunity entrepreneurs, whereas the latter with necessity entrepreneurs (Naudé, 2010). However, such an association must be made with caution, because we find businesses started by opportunity or necessity in both groups. The use of self-employed workers as equivalents to entrepreneurs is a point of contention in the literature, and it requires theoretical and methodological advances that allow for the improvement of analyses.

Previous studies have been carried out using these bases, such as the application of SINAN to identify the profile of diseases that affect entrepreneurs (Barbosa & Borges, 2021), the use of RFB data to map female entrepreneurship in the state of Goiás (Pagotto et al., 2020), the use of multiple national databases to assess the association of socioeconomic factors and the proportion of MEIs in Minas Gerais municipalities (Morais et al., 2022), the use of PNADc to investigate the characteristics of self-employed workers (Rossi, 2018) informality (Santiago & Vasconcelos, 2017) and the relationship between entrepreneurship and economic growth (Barros & Pereira, 2008). In addition to the examples provided in the preceding paragraphs, we find Brazilian authors who have used other secondary data sources in conducting research published in high-impact journals, such as Fischer et al. (2018), who employed data from FAPESP and CNPq in a study on academic entrepreneurship.

Given the above-mentioned possibilities, the following two sections will be devoted to an introduction to the R software, including a brief presentation of the program and, in sequence, the application of an exploratory analysis resulting from the combination of two databases, the GEM and the WGI.

THE R SOFTWARE

Statistical packages have always been associated with quantitative entrepreneurship research. The R software is one of the tools that has gained traction in recent years. R is a programming environment and language that focuses on statistical analysis (Hornik, 2020). R, unlike traditional software used in applied social sciences such as SPSS and Stata, is free of charge. Furthermore, because it is a tool with a programming language interface and due to the functions included in the hundreds of packages that can be installed, it has greater functionality versatility. Furthermore, if the researcher has a built script, the analyses can be reproducible, which contributes to greater research transparency, a condition that is increasingly valued by the scientific community of entrepreneurship (Anderson et al., 2019; Maula & Stam, 2020).

When packages are added to the software, they perform various functions[2] such as data reading and processing, visualization, and quantitative analysis. Such packages are frequently created and improved by R users, allowing the community to contribute to the tool's continuous advancement. Table 2 lists a few R packages and the functionalities they provide. It should be noted that this is by no means an exhaustive list. The RProject[3] website contains a complete list of packages as well as their documentation.

Table 2

R Language Packages

Packages	Functionalities
Reading: readxl, vroom, foreign	Readxl allows reading MS Excel files. Vroom loads files with larger volume of data quickly. The foreign package includes functions for reading files in other programs' formats, such as SPSS and Stata
Processing: dplyr, tidyr, lubridate	The dplyr package brings together an essential set of functions that allow filtering, selecting, grouping, summarizing and joining two or more databases. Tidyr contains functions for resizing your database, which is required for some visual and statistical analysis. Lubridate is devoted to treatments involving date/time data
Visualization: ggplot2, plotly, leaflet, DT, Shiny	Ggplot2 is the basis for creating graphs. Plotly allows the creation of interactive graphs. The leaflet has functions dedicated to creating maps. DT can generate interactive tables. Shiny allows the creation of web applications, such as interactive dashboards
Quantitative analyzes: survey, stats, tidymodels, psych, laavan	Many statistical techniques are included in the stats package, such as tests for comparing one or more groups (T-test, Wilcoxon, ANOVA) and regressions. The survey package is typically used on bases resulting from complex sampling surveys (eg PNADc). Tidymodels is a metapackage that combines several other packages dedicated to executing machine learning algorithms' workflows. Psych, for example, has implementations dedicated to factor analysis. Finally, laavan includes functions for structural equation modeling
Textual analysis: Tidytext, wordcloud, syuzhet, rtweet, bibliometrix	The tidytext package includes a large number of text-handling functions. You can create word clouds with the wordcloud package. The syuzhet generates sentiment analysis, which categorizes sentences into positive and negative valences and some sentiments defined by a lexical dictionary. Rtweet is a Twitter post extraction support package. Finally, bibliometrix is a package that aids in the execution of bibliometric analyses

Note: Elaborated by the authors.

Researchers can use R alone. However, the language is typically manipulated using the RStudio® software, which is an integrated development environment with a more intuitive interface that provides a better experience of use.

A BRIEF TUTORIAL ON USING R FOR DATA IN ENTREPRENEURSHIP

We performed a few reading operations, data processing, and exploratory data analysis for this tutorial. However, it is expected that some of the lessons presented here, such as the use of joins, will be useful in expanding the horizons of possibilities for entrepreneurship research by allowing the combination of multiple databases. As demonstrated in Section 3, studies involving GEM frequently connect it to other bases.

R and RStudio® must be installed on the computer or accessed via the RStudio Cloud tool[4] to perform the analyses. Furthermore, the spreadsheets containing the bases used in this case study, as well as the data dictionary from the annex of this document, must be accessed. The following procedures will be followed.

1. *Loading necessary packages for data processing and analysis*
2. *Reading data from spreadsheets and in comma separated values (.csv) format*
3. *Data analysis*
4. *Merging two databases*
5. *Performing exploratory data analysis and data visualization*

Let's begin by loading the packages and reading the bases that will be used in the example (see Box 1). Two databases will be used for this, which were originally combined in a previous study (Audretsch et al., 2021): the Adult Population Survey (APS) from the Global Entrepreneurship Monitor in aggregate format and the Worldwide Governance Indicator (WGI). The aggregated GEM APS base, as previously stated, includes the results of a survey on perceptions of entrepreneurial behavior and attitudes by country.

Box 1

```
# Installing packages that will be used. Once you have the packages installed,
# there is no need to run theses codes again
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

# Loading packages that will be used
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

# reading the databases through the read_csv function and saving them in the
# wgi and gem_aps objects
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 = ";")
```

The WGI database, on the other hand, provides information on the quality of governance in countries, including perceptions of corruption, the applicability of laws, political stability, among other things. As shown in Table 3, the following variables from each base were chosen for this analysis.

Table 3

Group of variables

Variable	Base	Description
Economy	GEM, WGI	Country
Continent	GEM	Continent
Entrepreneurship as a good career choice	GEM	Percentage of people aged 18 to 64 who agree with the statement: "In your country, most people think starting a business is a good career path."
Total early-stage Entrepreneurial Activity (TEA) Rate	GEM	Percentage of the population aged 18 to 64 who is a nascent entrepreneur or runs a new business.
Rule of Law	WGI	The extent to which agents are trusted and follow society's rules, as well as the quality of contract enforcement, property rights, police, and the judiciary.
Regulatory quality	WGI	Perception of the government's ability to develop and implement sound policies and regulations that promote private-sector development.
Political Stability	WGI	Perception of the likelihood of unconstitutional measures causing instability or seizing power, violence, including politically motivated conditions, and terrorism.
Voice Accountability	WGI	Perception of the extent to which citizens in the country can participate in the governing body, exercise free expression/assembly, and access to free media.

Note: Elaborated by the authors.

It is not the scope of the tutorial to delve into aspects of inferential statistics or machine learning, which would require greater theoretical depth to propose a model, as well as the leveling of knowledge in quantitative methods and assumption tests of statistical models.

If you're using RStudio®, the bases will be loaded in the Environment tab, which is usually located in the upper right corner of the program. Let's take a look at the variables using the dplyr package's glimpse() function. The glimpse function output shows that the WGI base has nine columns (variables) and 202 rows (observations). The first observations for each variable are displayed in front of it (see Box 2).

The goal now is to join both datasets. It is critical that both have a corresponding variable. The data dictionary and the initial inspection with the glimpse() function show that the *code* and *abrev* variables in the wgi and gem_aps databases are equivalent. The left_join function will then be used. We are telling R in the code at Box 3 to join the gem_aps and wgi datasets according to the abrev and code columns. Afterwards, the result will be saved in an object called gem_wgid.

Box 3

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

Next, only the variables relevant to the study will be selected using the select function from the dplyr package (see Box 4).

Box 4

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

The skim function of the skimr package will now be used to perform a descriptive analysis of the base (see Box 5).

Box 2

```
# glimpse function serves to inspect the base, including the number of
# observations, variables and variable types

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~
```

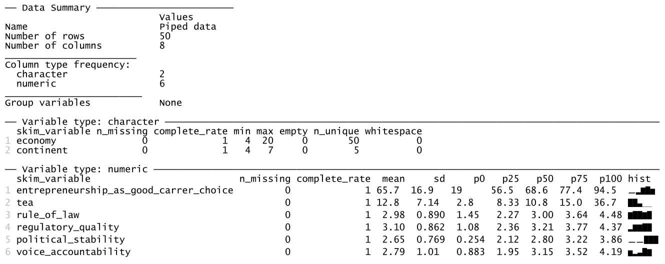
Box 5

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

The result of this function is a set of measures, such as mean, standard deviation, percentiles, and a simple histogram (see Figure 3).

Figure 3

Results of the skim() function



```
── Data Summary ────────────────────────
                           Values
Name                       Piped data
Number of rows             50
Number of columns          8
_______________________
Column type frequency:
  character                2
  numeric                  6
________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 economy               0             1   4  20     0       50          0
2 continent             0             1   4   7     0        5          0

── Variable type: numeric ────────────────────────────────────────────────────────────────
  skim_variable                          n_missing complete_rate  mean     sd     p0   p25  p50  p75  p100 hist
1 entrepreneurship_as_good_carrer_choice         0             1 65.7  16.9    19     56.5 68.6 77.4 94.5 ▁▁▃▇▅
2 tea                                            0             1 12.8   7.14   2.8    8.33 10.8 15.0 36.7 ▇▅▂▁▁
3 rule_of_law                                    0             1  2.98  0.890  1.45   2.27 3.00 3.64 4.48 ▇▇▅▇▇
4 regulatory_quality                             0             1  3.10  0.862  1.08   2.36 3.21 3.77 4.37 ▂▇▇▇▇
5 political_stability                            0             1  2.65  0.769  0.254  2.12 2.80 3.22 3.86 ▁▁▂▇▇
6 voice_accountability                           0             1  2.79  1.01   0.883  1.95 3.15 3.52 4.19 ▅▂▃▇▅
```

Note: Elaborated by the authors.

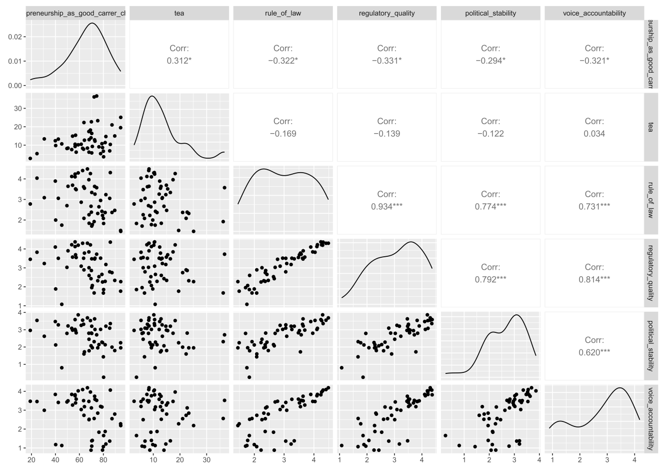
Finally, let's use the GGally package's ggpairs() function to generate a correlation matrix of the variables (see Box 6). Because the identification variables for country (economy) and continent (continent) are categorical, it was decided to remove them from the analysis using the function select (-variable name). The results are shown in Figure 4.

Box 6

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

Figure 4

Result of the ggpairs() function



Note: Elaborated by the authors.

It is possible to determine that the variable entrepreneurship as a good career choice had a negative and significant correlation with the institutional variables. These, in turn, were highly correlated with one another, which is understandable given the phenomena they measure. Again, this case study is limited to analyzing data using R language functions and does not intend to delve into theoretical aspects.

Finally, let's dig a little deeper into the relationship between two variables: political stability and entrepreneurship as good career choice (see Box 7). The ggplot data visualization function was used for this. Within the aes argument, the x and y coordinates are linked to the variables political stability and entrepreneurship as a good career choice, respectively, in the first parenthesis.

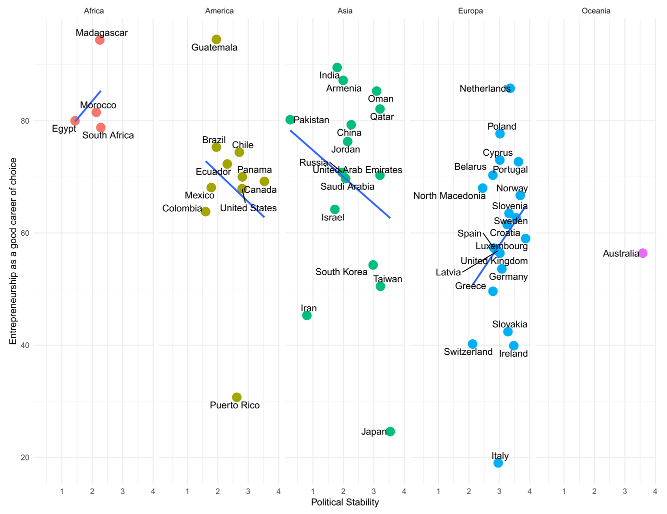
Box 7

```
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")
```

Next, we must specify the data layout format: points (geom_point()) or a smooth line describing the relationship (geom_smooth()). We can add parameters to both functions (e.g.: color the points according to the continents and increase the size of the points for better visualization). The geom_text_repel() function adds texts to each point based on the economy variable, while facet_grid divides the data into multiple panels based on the continents variable. Finally, the function theme_minimal() adds a minimalist design. The xlab() and ylab() functions modify the axis titles based on the text we entered. The results is shown in Figure 5.

Figure 5

Relationship between variables Entrepreneurship as good career of choice and political stability



Note: Elaborated by the authors.

This tutorial is also available in video format on the Youtube channel of the Laboratório de Pesquisa em Empreendedorismo e Inovação of Universidade Federal de Goiás (LAPEI – UFG). In 2021, LAPEI-UFG promoted a R course applied to entrepreneurship research in collaboration with 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). The course consisted of three modules that were delivered synchronously. The course had 161 participants from various education and research institutions throughout Brazil. The recordings had over 1500 views on YouTube® until October 2022. An assessment provided at the end of the training showed that the modules were rated between "satisfactory" (35%) and "very satisfactory" (65%). Participants emphasized the didactics and the quality of the materials available as strong points. Improvement opportunities included the division of shorter modules, more meetings, and meetings held outside of business hours at times.

ANALYTICAL APPROACHES TO ENTREPRENEURSHIP RESEARCH

The data analysis process must be tailored to the research question. The techniques, according to data analysis manuals, can be divided into interdependent and dependent and are associated with the type of relationship studied (Hair et al., 2009). The goal of interdependence analyses is to reduce, categorize, and group observations and/or variables. Techniques such as cluster analysis, principal component analysis, and factor analysis fall into this category. Dependency analysis refers to a class of techniques that attempt to estimate models that express the relationship between variables. In this regard, various regression techniques (e.g., linear, logistic, multinomial, negative binomial, quantile) and structural equation modeling are available (Favero & Belfiore, 2017). Some studies will be presented below that used techniques from both perspectives, dependence and interdependence.

Canestrino et al., (2020) used a cluster analysis in one of the early stages of their research on cultural values and the prevalence of social entrepreneurship to identify countries with similar cultural characteristics. The researchers used data from the Global Leadership and Organizational Behavior Effectiveness (GLOBE) project, which collects managers' perceptions of Hofstede's cultural dimensions across multiple countries. This allowed three groups with relatively similar characteristics to be identified. The first cluster was dominated by northern European countries and was labeled as friendly, the second by Asian and African countries and was labeled pragmatic, and the third by countries from southern Europe and Latin America and was labeled progressive.

Benatti et al. (2021) used dependency techniques to assess the relationship between MEI registration in municipalities in São Paulo and various economic indicators (the Municipal Gross Domestic Product – GDP-M – and Firjan Municipal Development Index – IFDM). The authors collected data from various secondary sources and used quantile regression in two models, both with the MEI record as an independent variable but two different dependent variables (GDP-M and IFDM). According to the study, the MEI has a greater impact on smaller municipalities as well as the IFDM's low and medium growth ranges.

Pagotto, Borges, et al. (2022) employed PSED 2 data to determine the association of different forms of capital - human, financial, and social - in the development of innovative capabilities in start-ups, which is another example of research that used dependency techniques. Among the variables studied, personal financial resources, education, and social capital employed to access physical infrastructure were determinants of the development of innovation capabilities in emerging companies over time (Pagotto, Borges, et al. 2022).

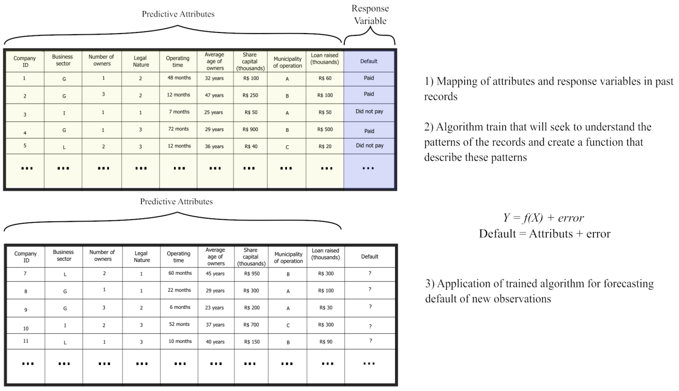
For some time, these techniques have been consolidated and developed in the context of statistics. They are typically covered in Multivariate Analysis courses at the undergraduate and graduate levels. However, given the increasing availability of machine learning tools, entrepreneurship researchers have encouraged the use of this approach in their research (Chalmers et al., 2021; Maula & Stam, 2020; Prüfer & Prüfer, 2020).

Although statistics and machine learning are both based on data and use similar techniques, the two approaches have distinct goals, methods, and tools. On the one hand, statistics is primarily concerned with inference, whereas machine learning is more concerned with prediction (Bzdok et al., 2018). Other distinguishing features of both approaches emerge from this preliminary classification and will be discussed after the following example.

Predicting is defined as the ability to predict a future outcome based on current characteristics (James et al., 2013). As an example, consider the case depicted in Figure 6 of a public policy manager who wants to develop a predictive model to determine whether companies that access a credit line will repay the loan after three years.

Figure 6

Creation of a predictive model



Company ID	Business sector	Number of owners	Legal Nature	Operating time	Average age of owners	Share capital (thousands)	Municipality of operation	Loan raised (thousands)	Default
1	G	1	2	48 months	32 years	R$ 100	A	R$ 60	Paid
2	G	3	2	12 months	47 years	R$ 250	B	R$ 100	Paid
3	I	1	1	7 months	25 years	R$ 50	A	R$ 50	Did not pay
4	G	1	3	72 monts	29 years	R$ 900	B	R$ 500	Paid
5	L	2	3	12 months	36 years	R$ 40	C	R$ 20	Did not pay
...	...	...	...	...	...	...	...	...	...

1) Mapping of attributes and response variables in past records

2) Algorithm train that will seek to understand the patterns of the records and create a function that describe these patterns

$Y = f(X) + error$

Default = Attributs + error

Company ID	Business sector	Number of owners	Legal Nature	Operating time	Average age of owners	Share capital (thousands)	Municipality of operation	Loan raised (thousands)	Default
7	L	2	1	60 months	45 years	R$ 950	B	R$ 300	?
8	G	1	1	22 months	29 years	R$ 300	A	R$ 100	?
9	G	3	2	6 months	23 years	R$ 200	A	R$ 30	?
10	I	2	3	52 monts	37 years	R$ 700	C	R$ 300	?
11	L	1	3	10 months	40 years	R$ 150	B	R$ 90	?
...	...	...	...	...	...	...	...	...	...

3) Application of trained algorithm for forecasting default of new observations

Note: Elaborated by the authors.

To achieve this goal, the manager will be able to train and validate a machine learning model to identify patterns in past data from companies that have faced similar challenges. The algorithm will map a large set of variables (e.g., number of entrepreneurs, gender of entrepreneurs, sector of activity, social capital, legal nature, location, family character, and so on) and look for patterns to form a function that describes the relationship. Following training, it is common practice to validate the algorithm's predictive capacity in a partitioning of its original database, known as the test dataset, to determine whether the model responds well to a subset of the data that did not participate in the training stage.

The manager will be able to read a new set of data that has the characteristics of the projects in his territory today and thus predict the chance of paying the loan after the desired period using the function developed based on the identification of past patterns and due care taken in the validation stage. It should be noted that the goal here is to perform prediction. Under these conditions, the function created may be difficult to interpret, depending on the algorithm used. As a result, while it is understood that it can accurately predict new observations, what lies behind it is not always clear.

Consider the following scenario: a researcher wishes to improve interpretability and comprehend how certain variables affect loan repayment. In this case, the investigator will approach the problem from an inference standpoint, which is traditionally associated with statistics (Bzdok et al., 2018).

Some consequences of the prediction/inference relationship are highlighted in the example. Machine learning methods are better at identifying patterns in large databases with many variables, whereas statistics focuses on a smaller set of variables with a wider range of observations. Furthermore, because of the flexibility with which patterns can be calculated, some machine learning algorithms can have good predictive power by creating sophisticated functions that describe the investigated relationships; however, they can provide low interpretability, which is required to perform inferences (Bzdok et al., 2018).

HOW TO ADVANCE IN THE LITERATURE ON ENTREPRENEURSHIP WITH THE SUPPORT OF TOOLS SUCH AS R

This section brings practices that can be used to advance quantitative research in entrepreneurship with the help of tools such as R. The points highlighted in this subsection are a collection of editorial discussions on quantitative methods in entrepreneurship research, such as the use of exploratory data analyses, actions to improve quantitative studies, and analysis publicity.

First, researchers should rely on exploratory data analysis techniques more frequently. Such techniques are typically recommended prior to the use of advanced multivariate modeling because they allow the discovery of patterns in variable distributions as well as the identification of missing data or outliers. They are, however, particularly useful for elucidating poorly understood phenomena. Descriptive analyses (including measures beyond the mean and standard deviation, such as minimum and maximum), cluster analysis, principal component analysis, and pattern identification in data visualization tools are examples of exploratory analysis techniques. The use of exploratory data techniques such as topic modeling, clustering, and network analysis can provide valuable research insights (Wennberg & Anderson, 2020).

Anderson et al. (2019) identify three important factors for the advancement of theoretical-empirical articles in the field of entrepreneurship: 1) the research question that drives the study; 2) the conditions that help improve causal inferences; and 3) the procedures used to reduce researcher bias. Concerning the first point, it is critical that the research question is qualified, and that the method used to answer it is adequate (Maula & Stam, 2020). Regarding the improvement of causal inferences, there is a stimulus for experimental research designs, regardless of their applicability, including their rigor in dealing with endogeneity problems and the ability to demonstrate causality relationships (Anderson et al., 2019; Maula & Stam, 2020).

The "hunting for asterisks" is one thing to avoid. This is a researcher's behavior in which there is a bias due to the need to obtain significant results in analyses, which leads to practices such as p-hacking and HARKing[5] . As Anderson et al. (2019, p. 4) emphasizes, "Researchers can publish good entrepreneurship studies, asking interesting questions and applying rigorous research designs regardless of identifying significant results." On the other hand, researchers must be aware of the magnitude of the effect identified in the model results. After all, a significant p-value does not imply that the predictor variable will have a practical effect on the variation of a dependent variable.

Another useful practice that has been promoted is the public dissemination of data and codes. Databases provided by researchers are assigned a Document Object Identifier (DOI) by platforms such as Researchgate and Data Mendeley. Rmarkdown (a file format for R that allows for the generation of reports), Google Colab, Jupyter Notebook, and Github can all be used to document the analysis performed.

AGENDA

Given what has been presented, entrepreneurship researchers can benefit from the increasing data availability as well as increasingly versatile and powerful software tools. As a result, the purpose of this section is to suggest potential research directions based on the discussions raised in this study.

As demonstrated in Section 3, there is a large volume of unstructured data available. Researchers have already investigated the potential of this type of data, conducting studies using data from social networks (Obschonka et al., 2017; Pagotto, Barbosa, et al. 2022), large media outlets (Suarez et al., 2020), and crowdfunding platforms. Future research may attempt to answer some of the following questions based on this type of data and previous research: What are the representations of entrepreneurship that the mainstream media creates? What are the major media outlets' discourses on entrepreneurs in Brazil? (Suarez et al., 2020).

Many national surveys, such as IBGE, do not include the term "entrepreneur" among job categories. Self-employed workers and employers are the two most closely related groups to entrepreneurship. Hence, a second research avenue would be to delve deeper into studies of self-employed workers. Efforts to explore deeper into this occupational category can already be seen in the international literature; one example is a special edition of Small Business Economics on the subject due out in 2020 (Burke & Cowling, 2020). Although still in its early stages, documented experiences with the use of IBGE databases to conduct entrepreneurship research exist in Brazil (e.g., Almeida et al., 2017).

According to national and international discussions, self-employed workers are a growing profile (IBGE, 2021), diverse - ranging from garbage collectors to doctors, in the words of Santiago and Vasconcelos (2017) - (Burke & Cowling, 2020; Moortel & Vanroelen, 2017), and, on average, more vulnerable than employed workers. Therefore, more research into this profile, its context, and entrepreneurial process is required. Furthermore, studies can be conducted using Brazilian databases that consider the occupational profile to better segment the Brazilian self-employed worker.

Another path for future research is to assess the potential of data to investigate the entrepreneurial phenomenon at various levels and from a multilevel perspective. Datasets of the RFB, ICE, and MUNIC can be linked to other bases at the municipal level to investigate the impact of contextual and institutional variables on entrepreneurship (Muñoz-Fernández et al., 2019). Morais et al. (Audretsch & Moog, 2020) used this strategy by combining data from various sources (e.g., FIRJAN, RAIS, CAGED, DATASUS, INEP, IBGE) to assess the relationship between socioeconomic variables (e.g., income, education, health) and the proportion of MEIs at the municipal level. Regarding the country level, datasets such as the GEM can be used to better understand the relationship between entrepreneurship and other contextual conditions such as democracy (Audretsch & Moog, 2020).

Endnotes

[1] Avaiable on: https://datasebrae.com.br.

[2] Ready-made routines created by package developers.

[3] https://cran.r-project.org/web/packages/available_packages_by_name.html.

[4] https://rstudio.cloud.

[5] A practice that involves the selection of variables from previous analyses yielding significant results.Cerum constemerest iae din dicem. Grae etrudem utus es

Conflit of interest statement

The authors declare that there is no conflict of interest.

Authors' statement of individual contributions

Roles	Contributions	
	Pagotto DP	Borges C
Conceptualization	■	■
Methodology	■	■
Software	■	
Validation	■	■
Formal analysis	■	■
Investigation	■	
Resources	N.A.	
Data Curation	N.A.	
Writing - Original Draf	■	■
Writing - Review & Editing	■	■
Visualization	N.A.	
Supervision	■	■
Project administration	N.A.	
Funding acquisition	N.A.	

REFERENCES

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. *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 enterprenuership 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. 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. http://dx.doi.org/10.1561/0300000022

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. https://rct.dieese.org.br/index.php/rct/article/view/177

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

AUTHOR BIOGRAPHIES

Daniel do Prado Pagotto *obtained his master deegree on Business Administration at Universidade Federal de Goiás and currently is working towards a Ph.D. at University of Brasilia's Business Administration Program. Daniel is adjunct coordinator of Entrepreneurship and Innovation Research Lab from UFG (LAPEI-UFG). His current research focuses on entrepreneurship and service innovation, applying quantitative methods and secondary data. His papers have been published in journals sucha as Revista de Inteligência Competitiva, Humanidades & Inovação, REGEPE among others.*

E-mail: danielppagotto@gmail.com.

Cândido Borges *is Professor of entrepreneurship at the Universidade Federal de Goiás, Goiânia, Brazil where he is also Director of the Entrepreneurship and Innovation Research Laboratory (LAPEI-UFG) and Professor at UFG's Management Pos-Graduate Program. He obtained his Ph.D. at HEC Montréal and a Post-Doc from EAESP-FGV. His current research focuses on new ventures, self-employment and entrepreneurship policy. His papers have been published in journalssuch as Future Studies Research Journal, Humanidades & Inovação,REGEPE, RAUSP, among others.*

E-mail: candidoborges@gmail.com.