Global Entrepreneurship Analytics: Using GEM Data

Book · July 2020

DOI: 10.4324/9780429316715

CITATIONS
3

READS
980

2 authors:

Milenka Argote
Bidem Co
68 PUBLICATIONS **151** CITATIONS
SEE PROFILE

León Darío Parra Bernal
Universidad EAN
81 PUBLICATIONS **199** CITATIONS
SEE PROFILE

Global Entrepreneurship Analytics

This innovative book proposes new methodologies for the measurement of entrepreneurship by applying techniques of demography, engineering, mathematics and statistics.

Using the data from the Global Entrepreneurship Monitor (GEM), statistical demographic techniques are used for the evaluation of data quality (EDQ), and a new methodology for the estimation of Specific Entrepreneurship Rates (SER) and the Global Entrepreneurship Rate (GER) is proposed. At the same time the authors present artificial intelligence techniques such as Fuzzy Time Series (FTS) to forecast data series of the entrepreneurial population. Finally, they present a case study of the implementation of Big Data in Entrepreneurship using GEM data that shows the latest technological trends for the management of data, in support of making more accurate decisions. Being a methodological book, the techniques presented can be applied to any dataset in different areas. Readers will learn new methodologies of analysis and measurement of entrepreneurship using data from the Global Entrepreneurship Monitor. They will be able to access the experience of the authors through each of the applied cases in which the reader is taken by the hand, both through the scientific method and through the methodology of construction of more accurate metrics in entrepreneurship, with less error.

This book will be of value to students at an advanced level, academics and researchers in the fields of Entrepreneurship, Business Analytics and Research Methodology.

Milenka Linneth Argote Cusi is Founder of Business Intelligence and Demography SAS (www.bidem.com.co).

León Darío Parra Bernal is Associate Professor at EAN University, Bogotá, Colombia (www.universidadean.edu.co).

Routledge Focus on Business and Management

The fields of business and management have grown exponentially as areas of research and education. This growth presents challenges for readers trying to keep up with the latest important insights. Routledge Focus on Business and Management presents small books on big topics and how they intersect with the world of business research.

Individually, each title in the series provides coverage of a key academic topic, while collectively the series forms a comprehensive collection across the business disciplines.

Distributed Leadership and Digital Innovation
The Argument for Couple Leadership
Caterina Maniscalco

Public Relations Crisis Communication
A New Model
Lisa Anderson-Meli and Swapna Koshy

Implicative Marketing
For a Sustainable Economy
Florence Touzé

Global Entrepreneurship Analytics
Using GEM Data
Milenka Linneth Argote Cusi and León Darío Parra Bernal

For more information about this series, please visit: www.routledge.com/Routledge-Focus-on-Business-and-Management/book-series/FBM

Global Entrepreneurship Analytics

Using GEM Data

Milenka Linneth Argote Cusi and León Darío Parra Bernal

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2021
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
Names: Argote Cusi, Milenka Linneth, author. | Parra Bernal, León Darío, author.
Title: Global entrepreneurship analytics : using GEM data / Milenka Linneth Argote Cusi, León Darío Parra Bernal.
Description: New York, NY : Routledge, 2021. |
Series: Routledge focus on business and management | Includes bibliographical references and index.
Identifiers: LCCN 2020016116 | ISBN 9780367321178 (hardback) | ISBN 9780429316715 (ebook)
Subjects: LCSH: Entrepreneurship--Statistical methods.
Classification: LCC HB615 .A746 2021 | DDC 338/.040727--dc23
LC record available at https://lccn.loc.gov/2020016116

ISBN: 978-0-367-32117-8 (hbk)
ISBN: 978-0-429-31671-5 (ebk)

Typeset in Times New Roman
by MPS Limited, Dehradun

Dedicated to all those innovative researchers in the study of measurement methods for reducing uncertainty, especially for the Global Entrepreneurship Monitor whose worldwide initiative is a benchmark in the measurement of entrepreneurship.

Contents

Illustrations

Figures

Tables

Preface

The main motivation of this book is the application of new methodologies for the measurement of entrepreneurship. Consistent with the philosophy of entrepreneurship whose differential is innovation, it is necessary to innovate in new metrics on entrepreneurship in support of making more accurate decisions.

The measurement of entrepreneurship is not simple, since it is a complex concept that defines a state or action of the individual that is related to many factors, especially those of economics and survival. Thus, this book delves into the concepts of measurement and its practical application to discover a number or data that ultimately tries to measure as objectively as possible the phenomenon under study. There is no doubt that in order to make a "good" measurement, the rigor of the methods used should provide the scientific support required to have a level of confidence in the metrics generated.

Data is required for measurement. The issue of data is crucial, since it is the raw material used to create a measure. That is why, given the availability of official Global Entrepreneurship Monitor Data that collects information on entrepreneurship worldwide in a standardized way, this book uses this rich source of information in all chapters for the case of Colombia because it is available to the authors. It should be noted that the authors have tried to explain the methodologies presented in this book so that they can be applied using other databases or other cases, and that is the most important value of this intellectual production.

Being consistent with the thread of the book, it begins with a transcendental theme, before transforming the data into useful information for decision making, the Evaluation of Data Quality (EDQ) in Chapter 1. Thus, the book begins with a methodology for the evaluation of the quality of the data that uses statistical and demographic techniques and that was presented in its first version at the GEM data use seminar for scientific publications in 2018 held in Bogotá, Colombia.

Chapter 2 presents a new entrepreneurship metric, the Global Entrepreneurship Rate (GER), created similarly to the Global Fertility Rate, an internationally known indicator calculated in a standardized way worldwide for monitoring the number of children per woman. In this sense, the theoretical mathematical concepts of this rate are taken up and by analogy applied to the case of entrepreneurship to find a measure of the average annual number of ventures per adult between the ages of 18 and 64. Behind the GER summary indicator are the Specific Entrepreneurship Rates (SER) that are estimated and explained in this chapter. The metrics created are consistent and capture the heterogeneity of entrepreneurship behavior at different ages.

Chapter 2 is the input for Chapter 3. With a new metric on entrepreneurship that eliminates sample bias, Fuzzy Time Series (FTS) is applied to forecast GER data series using GEM data. It is an innovative methodology that comes from Artificial Intelligence (AI) which models with greater certainty phenomena that are not subject to assumptions such as linearity and balance. The importance of Chapter 3 is that it ventures into sophisticated techniques of forecasting methods that nowadays apply to organizations with a high level of maturity.

Chapter 4 takes another step in the use of new technologies for data management and the generation of new metrics, this time integrating different data sources. This chapter presents the methodology developed for the implementation of Big Data technologies for the exploitation of GEM data. The project, which was carried out between 2017–2018, has generated interesting lessons that can be taken into account to initiate and execute projects of this magnitude. The importance of this chapter is the transfer of knowledge in the implementation of projects in Big Data in general and its usefulness in the case of the GEM that would allow it to integrate information from different official sources of data, which constitute a valuable input for the construction of new, more complex metrics for decision making.

We hope that this book is an open source of future applications as well as a source of innovations that will allow us to leave our comfort zone, expand knowledge boundaries and integrate new methodologies from other areas which provide tools to perform measurement processes that are more accurate or otherwise contain less error.

Sincerely,

The Authors

1 Methodology for the Evaluation of Data Quality

The GEM Case

1.1 Introduction

A key process of data science is the stage prior to data analysis, namely the evaluation of data quality (EDQ). It is important to note that data science begins with the collection of data, without which we could not make science. There are several data collection methods: censuses, surveys, public records, and nowadays we can add data from social networks. Each of these methods entails its own scope and limits while using different techniques from statistics and demography.

Data quality control starts from its collection, which is why there are techniques that allow us to minimize error in the process of collection. Once the data is collected it is stored in databases called "microdata" that refer to the original data. The process of typing and organizing data in structures such as databases requires a particular knowledge of data design and management; in addition, knowledge of the technologies is also necessary to carry out this process efficiently (servers, database managers, software, hardware, etc.).

Once we have a set of organized and systematized data understandable to the "general public", the process that gives meaning to this chapter begins: the evaluation of data quality. In the frame of data science, the difference between Data Cleaning, Data Tidying, Data Mining and Data Analytics is that EDQ is applied after data is cleaned, organized and standardized and prior to DA. It is a stage in which specific techniques are applied that derive from the experience of the data scientist and the knowledge of the data source, and probably due to this it is a generally omitted phase in the different data science projects or applied studies, or a phase relegated to a technical aspect, when it is of vital importance for the validity and reliability of the data. If an EDQ is not performed, surely the information obtained from the data will be erroneous.

It is feasible to find in the literature researches on data quality in areas related to engineering and health, and there are methodological proposals, each from a different perspective. It is a process that depends on creativity and both the integral knowledge and techniques of the area to design instruments or processes (i.e., play with the data) with the objective of evidencing possible errors or inconsistencies in the information that can call into question the information obtained from the data. However, the experience in EDQ allows us to design a methodology that can be applied in any area.

1.2 The Data Contains Errors

It is important to accept that data contains errors; that is to say, it is perfectible. The analysis of uncertainty associated with data with the idea of collecting information for government, institutions or for scientific research is a challenge. In other words, from the moment the need for data arises, an abstraction level is born to design the instrument and select the techniques required to collect data according to the objective of the research or intervention.

Errors can be summarized into sample and non-sample errors during the process of collecting the information from the beginning until it is available in a database (Argote, 2003). In the conceptual phase of the idea or need to obtain data, errors are tied to the level of abstraction of the person who conceived it and they are expressed in concepts, definitions and measures that will be specified in a design. The design phase of the instrument or the data collection source is also subject to human error, since, if there is no good design of the instruments, the experiment will be unsuccessful. Another source of error is the selection of the population from whom the information referring to the sample design is collected, which may result in sample errors that limit the inferences that can be made from the data.

At the moment of the implementation of the fieldwork or of what the data entry itself is, there are several non-sampling errors tied to all the actors involved in the process such as surveyors, interviewers and supervisors, all of whom can make mistakes in their declarations or in the collection and verification processes. The data is collected in a repository depending on the available technology (e.g., paper, google forms, excel, etc.). There is a process of typing data that can lead to further human or interface errors. It is assumed that there is a database design behind the typing process, with certain quality requirements to ensure that data is organized so it can be optimally

exploited later. However, this phase can still be subject to error if it is not done by a specialist in design and database management.

Finally, once the data is organized in structures that allow its manipulation by computational means, which is constituted in a product that includes or is the sum of the errors that could have accumulated during the process, consequently it is a product that should be evaluated.

1.3 What Is the Evaluation of Data Quality?

According to the literature, quality is a relative concept. The concept of quality arises in the production sector given the need to measure its goals not only by the level of production but also by the quality of each of the products offered to customers. In this sense, what is understood by quality in a given sector is defined by the interest group through different dimensions which are considered important for quality measurement. According to Heredia and Vilalta (2009), what can be quality for some may differ from quality for others; however, it should be oriented towards improvement.

Quality is a desired characteristic of things and products, since it is synonymous with well done, well dealt with, well built, etc. Heredia and Vilalta (2009), in their article on the importance of evaluating data quality for companies, define data quality on the basis of the implications of their poor quality. "Poor data quality affects business management in various ways. Obviously a primary affectation of the poor data quality is its effect on decision making."

In this regard data quality refers to the fact that data, as well as products and services, must have quality, as the implications of not possessing it affect accurate decision making. The evaluation of data quality is a process that involves different methodologies, techniques and technologies prior to its use in knowledge management (Heredia & Vilalta, 2009; Argote, 2003). The data quality measurements are conducted effectively through dimensions such as those mentioned by Hadhiatma (2018): namely completeness, consistency and precision. Other authors, depending on the area, use other dimensions within which the most important ones (which are taken into account in this chapter) are coherence, validity, truthfulness and reliability (Zúñiga & Sánchez, 2012; Azeroual, Saake & Abuosba, 2019; Hadhiatma, 2018).

1.4 Methodologies to Evaluate Data Quality (MEDC)

Given the different perspectives and data source, the methodologies are diverse; however, it is feasible to find points of convergence.

So, the literature has several EDQ cases mostly from the quantitative, engineering, information technology areas and others in the health and demography areas, demonstrating the extremes in which this field moves, from the very technical to the most social, making the intermediate areas such as entrepreneurship, that comes from the economy, lag behind in the application of these techniques.

Zúñiga and Sánchez (2012) apply an interesting methodology for EDQ to the enrollment of a university in Costa Rica. It deals with the methodology of Arkady Maydanchik (2007) who applied for the analysis of an institutional database carried out in 2011, information of which was evaluated from 1980–2011. An abstract of the methodology is shown in Table 1.1.

On the other hand, Azeroual, Saake and Abuosba (2017) present a methodology related to measurements for data quality and data cleansing for a Research Information System (RIS). According to the authors, the creation of measurements depends on the type of data analyzed: *Laissez faire,* few or rare changes in the nature of the data (errors of incidental nature so they can be ignored); *Reactive approach*, important data and with rare changes (errors are corrected but its causes are not accounted for and it does not need monitoring) and *Proactive approach*, important data that is frequently changed (oriented towards prevention and elimination of the sources of error and tied to constant monitoring). Once a perspective is adopted based

Table 1.1 The Methodology of Arkady Maydanchik

Steps	*Description*
Collect Data	Both steps are correlated with an exploratory process and with knowledge of the data
Data Analysis and Data Tidying	
Generate Data for Test	Refers to a process to generate new data and different kinds of tests about data
Design of Rules of Data Quality	Refers to the set of restrictions to evaluate data quality
Implementation of Data Quality Rules	Refers to implementation of the rules in code
Adjust the Rules of Data Quality	An iterative process
Tabulation of Aggregate Results	According the preview process it generates tables and aggregate indicators
Dashboard of Data Quality	The result of EDQ is presented in a dashboard

Source: Own elaboration based on Zúñiga & Sánchez (2012: 41).

on the identification of the type of data according to the above, the authors present seven steps for EDQ:

1 Identification of the data to be checked for data quality (typically research-related operational data or data in decision-making reports).
2 Decide which dimensions are used and their weighting.
3 Work out an example of good and bad data for each criterion.
4 Apply the test criterion to the data.
5 Evaluate the test result and whether the quality is acceptable or not.
6 Make corrections where necessary.
7 Repeat the process periodically and observe trends in data quality.

Concomitantly, it is important to remember that nowadays the types of information available on the web include structured and unstructured data, which creates the need to expand the concepts of EDQ. In this sense Hadhiatma (2018) made a revision of the EDQ methods and techniques for Linked Open Data (LOD), whose definition refers to open data at a lower level. This research is important, as it helps us to understand EDQ needs in more complex scenarios such as Big Data, WEB data, WEB semantic, etc. There is evidence that more sophisticated methods are required to identify patterns at a lower level, known as the *"Ontology"* level, where problems of completeness, consistency and precision also tend to occur. That is, the progress in this area is still incipient, since the main problem with the implementation of applications such as Big Data are data quality, data interoperability and data management (Hadhiatma, 2018: 5).

The methodology proposed in this chapter, although it has its encounters with the previous ones, has its differences. They coincide in three key processes at a general level: data exploration, generation of rules or parameters, and evaluation of the results in relation to the parameters and correction. In terms of the added value presented by our methodology, it defines analysis dimensions based on the knowledge of demographic and statistical techniques and does so in a more operational way adapted to the dimensions "data mining" and the "timeline". In addition, the proposed methodology includes a comparison process, which goes beyond the identification of patterns to the construction of standard key indicators where the results can be compared with other data sources. Finally, the methodology is focused on the case of population data, since it makes use of the masculinity index to evaluate the balance of the data sample (although for other

areas you can build a simile of a ratio between variables of interest), and uses a theoretical definition of a rate that in mathematical terms represents a frequency of cases in relation to a total. The proposed data source and methodology are detailed below.

1.5 Data and Methodology

This chapter takes as a data source the cases corresponding to the Adult Population Survey (APS) of GEM Colombia (2017), which includes 2098 cases whose methodology and characteristics are explained in detail in Reynolds et al. (2005) (see Table 1.2). The APS considers a set of closed and some open questions for the description of economic activity. The questionnaire starts with filter questions such as age and whether the person is currently trying to start a self-employment business or for an employer. Subsequently, a set of questions is formulated to characterize different aspects of entrepreneurship start-up, business ownership, type of business, reasons for entrepreneurship, source of investment, evaluation of expectations over time, networks, internationalization, potential growth, and support programs for entrepreneurship. There is a control question to identify whether the interviewee is the current owner of an independent business of the declared entrepreneurship. If so, this allows for the registration of other entrepreneurships. Finally, a demographic block is applied in which sex, age, employment, household size, household income, educational level, marital status and socioeconomic status, among others, are considered.

Table 1.2 Distribution of the Adult Population Sample aged 18–64, GEM Colombia Data, 2017

Age Group	*2017*	*Percentage*
18–19	115	5.5
20–24	273	13
25–29	270	12.9
30–34	225	10.7
35–39	258	12.3
40–44	206	9.8
45–49	211	10.1
50–54	196	9.3
55–59	179	8.5
60–64	165	7.9
Total	**2,098**	**100**

Source: Own elaboration.

The proposed methodology is based on a value chain that goes from the simplest to the most complex in relation to the number of variables to be considered in order to evaluate the data. The first phase of EDQ begins at the moment the data is available to the user, and involves a set of processes related to the exploration, identification of key variables and in-depth analysis of the behavior of these variables; thus this stage coincides with the definition of "mining data" in the sense that it digs deeply into the data in order to find the "vein", that is to say, the coherence thereof.

Following the value chain set out in Figure 1.1, the simplest indicator to evaluate the behavior of population data in demography is the masculinity index, which is part of the methodology proposed, since it allows us to visualize in a simple way the data distribution under the approximate parameter of 50/50 for the sex ratio. Any behavior that moves away from this standard deserves a thorough evaluation. In addition, since the analysis is performed at individual ages, an analysis of the age declaration can be performed which is susceptible to an incorrect declaration.

Another proposed indicator is the response rates dealing with non-sampling errors (Argote, 2007). The objective is to identify key variables from which to find out whether the population has responded or not responded. Many times, researchers assume that the key variables of their research include data and that they are consistent; however,

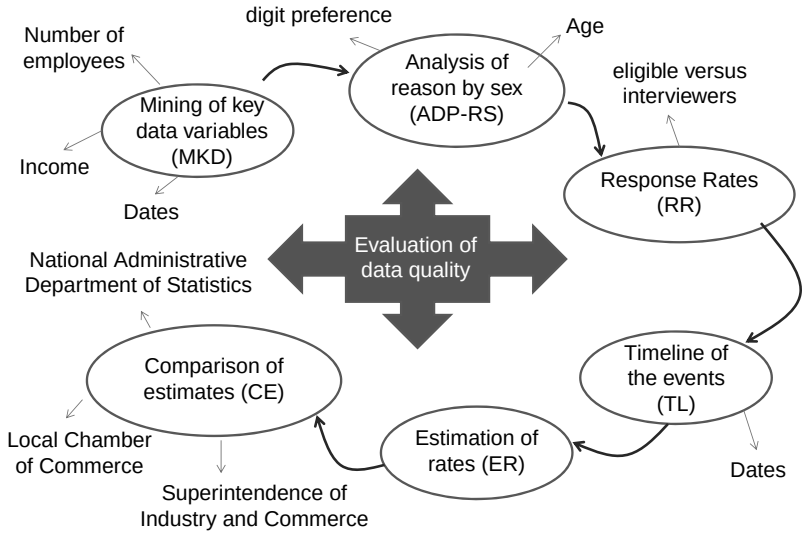

Figure 1.1 The Evaluation of Data Quality (EDQ) Methodology.

Source: Own elaboration.

when doing their analysis, the opposite is true. The response rates, as the name implies, are quotients selected according to the group of interest, and require a knowledge of mathematics and demography for the development of the response tables (see Figure 1.1).

Once you have acceptable response rates or a maximum accepted non-response, the evidence is available to continue with the next EDQ stage. The center of the next stage is "time". This consists of selecting those key variables related to time such as dates, consecutive events, correlated events, etc. in order to perform an analysis of the timeline of the events. This type of variable is often found in demographic and health surveys, which are designed to collect information on women's birth dates. Other longitudinal design studies (e.g., the Mexican National Urban Employment Survey (NUES)) perform the registration of the dates of entry into a new job, with the aim of conducting longitudinal analysis of the individual's history in the labor market. Likewise, the registration of dates allows for survival analysis. This analysis allows for the verification of the consistency with which the events were declared and for the identification of anomalies.

The following process responds to more complex mathematical concepts from the point of view of demography and the probability theory, since the aim is to create measurements of social phenomena (indicators) in support of decision making. Rates are the most used measurements and have a complex theoretical definition in a mathematical sense that is usually difficult to estimate using dates (Argote, 2007, 2009). The aim of a rate is the measurement of the frequency of occurrence of an event in relation to the population at risk of experiencing it. It is a measure that approaches the probability of occurrence. At this stage, a knowledge of statistics, mathematics, demography and information technology is required for programming algorithms to estimate a rate.

Finally, the indicators and charts, the result of the previous processes, constitute a set of measurements obtained from the data source, the input for the next stage. It is a process of comparing the consistency of estimates with other sources. In this case, when it comes to rate indicators defined as standard such as birth or death rates, if two sources have the same measure the comparison is feasible. For non-traditional measurements it is necessary to apply the same methodology for the sources to be compared or to otherwise find approximations, so that the consistency of estimates can be evaluated (as Argote did in 2018) to compare different estimates of the Mexican population projection (Argote, 2015).

The previous summarized explanation of the methodology is detailed below in terms of the procedure and mathematical formulation in the case of the APS GEM data.

1.5.1 Mining of Key Data Variables (MKD)

This first phase consists in exploring the variables, knowing and identifying them both in the instruments for their collection and in the database to which they must correspond (see Table 1.3). Mining implies the following procedures:

- Revision of the names of the variables, labels and their categories in the database for the user's understanding.
- Analyze and execute frequency commands of demographic variables (sex, age, marital status, income level, educational level, etc.), variables that allow us to know the population or sample under study. This data must be consistent with similar samples or with other data sources.

This stage implies carrying out the following procedures in a database processing program (SPSS, STATA, TABLEAU):

- Recode variables for easy location. In big databases the location of the variables is difficult, and in order to maintain the premise of not modifying the original database it is recommended to copy the variables of interest in others that are located at the end of the database for quick location.
- Recategorize variables to standardize. For those cases in which variables have multiple categories such as economic activity or simple ages, it is recommended to recode to simplify the number of categories according to such standards as international codes, five-year age groups, etc.
- Variable crossing allows us to evaluate the coherence between variables. It is important to carry out crossing of variables such as age versus sex, as it allows us to observe the coherence of the data distribution. Regarding entrepreneurship with APS GEM, it is interesting to cross the variables of the current number of employees versus the expected number of employees in the next five years, as it allows us to identify anomalies.

1.5.2 Analysis of Digit Preference and Reason by Sex (ADP-RS)

The analysis of the data by individual ages allows us to evaluate the digit preference in the declaration of age, and its distribution by sex allows us to construct the masculinity index to evaluate the distribution by sex of the data. These two demographic variables, namely sex and

Table 1.3 Key Variables on Entrepreneurship, APS GEM

Key Questions (Variables)	*Description*
bstart	Q1A1. Are you, alone or with others, currently trying to start a new business, including any self-employment or selling any goods or services to others?
bjobst	Q1A2. Are you, alone or with others, currently trying to start a new business or a new venture for your employer as part of your normal work?
suacts	Q1B. Over the past 12 months have you done anything to help start this new business?
subustype	Q1F. What kind of business is this?
sunowjob	Q1H1. Not counting the owners, how many people are currently working for this business?
suyr5job	Q1H2. Not counting owners, how many people will be working for this business five years from now?
ownmge	Q2A. Are you, alone or with others, currently the owner of a business you help manage, self-employed, or selling any goods or services to others?
ombustype	Q2F. What kind of business is this?
futsup	Q3A. Are you, alone or with others, expecting to start a new business, including any type of self-employment, within the next three years?
discent	Q3B. Have you, in the past 12 months, sold, shut down, discontinued or quit a business you owned and managed, any form of self-employment, or selling goods or services to anyone?
Demographic Variables	
Gender	A. What is your gender?
Age	B. What is your current age (in years)?
hhsize	E. How many members make up your permanent household, including you?
cohhinc	F. Which of these ranges best describes the total annual income of all the members of your household, including your income, as one combined figure?
coreduc	G. What is the highest level of education you have completed?
costrata	M. Survey vendor to indicate stratum which corresponds to the respondent, if applicable to sample.
dtsurv	N. Date of survey (dd.mm.yy).

Source: Own elaboration.

age, are determinant variables of many social behaviors and tend to carry considerable weight in most mathematical models designed to explain population behavior.

The masculinity index is the ratio between the number of men at age x and the number of women at the same age. This is the reason that measures the number of men per 100 women.

$$IM = \frac{H_x}{M_x} \tag{1}$$

The reason it is an indicator for EDQ is because there are parameters for the behavior of the reason by sex and this allows for the evaluation of how large a gap there is between the behavior of the data samples in relation with the parameters. In this case, the ratio is expected to be close to one for all ages.

In Figure 1.1 we observed the behavior of IM for the 2017 Colombia GEM data. The behavior of the curve is erratic according to the individual ages; a pattern cannot be distinguished and it is striking that the approximated 50/50 rule applies in only a few cases; that is to say, close to 1. This shows us that the characteristics of the sample are unique features. On the contrary, GEM methodology does not take into account age and sex intervals for select APS samples, which can present imbalances, as observed in Figure 1.2.

The revision of the age declaration is relevant to EDQ because it is an indicator of the consistency of the data. According to several demographic studies (e.g., Pressat, 2000), there is a preference for declaring even numbers or those ending in zero. These are cases where people declare a younger age, among others. Therefore, by looking at individual ages, we can observe whether there are patterns that are not consistent.

1.5.3 Response Rates (RR)

The idea of response rates is to calculate the percentage of people eligible to answer a question, and who really answered. The response rates are expected to be very close to 100 percent because it tells us how the population responded. Response is understood to be that which is in the range of possible answers that are different to non-responding (NR) or lost data (MISSING).

For the estimation of response rates, matrices are constructed to allow for the calculation of reasons. The matrices represent group and subgroup arrangements according to five-year age groups, which by convention is the pattern to organizing population data. The simplest

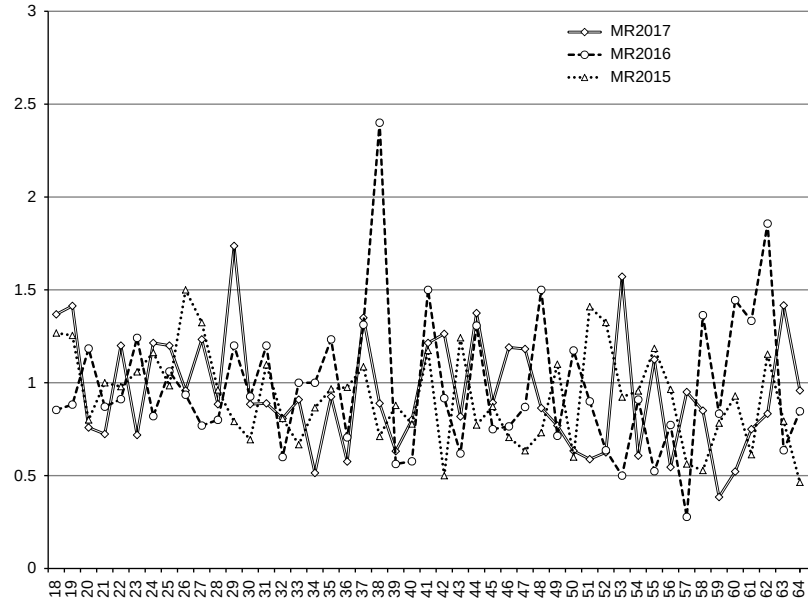

Figure 1.2 Masculinity Rate, GEM Colombia Data, 2017.
Source: Own elaboration using GEM data.

matrix is the distribution of the total population by five-year age groups (denominators of rates).

For the calculation of subgroups that make up the possible numerators, filters are used to select the cases that have an answer. In this sense, the question to be evaluated is selected in advance; for example, in the GEM case, we have control questions: Have you started an entrepreneurship in the past 12 months? (YES/NO). The response rate for this variable has as a numerator the number of cases that answered "Yes" between the total population of that age group (See Table 1.4).

$$TR = \frac{Answers_{x+5}}{Eligibles_{x+5}} \tag{2}$$

1.5.4 Timeline of the Events (TL)

Time is one of the key variables in statistics and demography. From a statistical point of view, the entities under study – in this case the

Table 1.4 Entrepreneurship Response Rates, GEM Colombia Data, 2017

Age	*Women*				*Men*			
	Total Women in the Sample	*Eligible*	*Respondent*	*Answer Rate*	*Total Men in the Sample*	*Eligible*	*Respondent*	*Answer Rate*
18–19	48	9	7	77.8	70	20	15	75
20–24	142	42	30	71.4	125	40	26	65
25–29	139	39	28	71.8	154	38	26	68.4
30–34	127	38	24	63.2	108	35	21	60
35–39	139	50	29	58	116	32	20	62.5
40–44	100	25	16	64	100	31	22	71
45–49	107	30	24	80	97	30	14	46.7
50–54	114	40	25	62.5	93	24	13	54.2
55–59	90	16	7	43.8	79	31	16	51.6
60–64	86	12	7	58.3	64	17	10	58.8
Total	**1,092**	**301**	**197**	**65.4**	**1,006**	**298**	**183**	**61.4**

Source: Own elaboration using data from APS GEM.

population or the companies – are not static entities but exhibit a behavior over time. Therefore, when modeling real phenomena, it is crucial to have a record of the time.

Unfortunately, not all cases have a record of dates that correspond to the events we wish to study. In these cases it is feasible to perform an analysis of the sequence of events, since some events precede others. For example, in the case of fertility, individuals are first born before entering school, but if this sequence is detected upside down it is an anomaly. In the case of entrepreneurship, the idea of entrepreneurship is expected to be followed by formal registration and so on. If questions about this process exist, it is feasible to evaluate a timeline of events.

The starting point for the sequence of events with or without a date is the bivariate analysis between consecutive events (variables). In this way it is possible to construct variable crossing tables to evaluate inconsistencies that may arise in this regard. For the sequence analysis of more variables there are more sophisticated methods such as life trajectory analysis or survival analysis that are used in longitudinal studies.

In the case of entrepreneurship with 2017 GEM Colombia data, a table is constructed by five-year age groups in which the number of people who claim to be undertaking an entrepreneurship (bstart), the number of people who take any action on their entrepreneurship are recorded (suacts), then if you declare that at the time of the interview you own a business (ownmge), and finally those who respond that they closed a business (discent) (see Table 1.5).

1.5.5 Estimation of Rates (ER)

Estimating a rate is not a simple exercise. The definition of different demographic measures tells us about it. According to Moreno, López, and Corcho (2000), the rate, comprising a number and denominator, is a type of quotient among others (detailed in Chapter 2, this volume) which represents the methodology for estimating the Global Entrepreneurship Rate. According to the above, a theoretical rate is defined as follows:

$$Rate = \frac{Number\ of\ events\ which\ occur\ in\ a\ determined\ period\ of\ time}{Exposition\ time\ of\ the\ individuals\ until\ they\ experiment\ the\ event}$$

As an example, the Global Fertility Rate is defined as standard worldwide and whose calculation is carried out as follows:

$$GFT = \frac{Number\ of\ births\ per\ woman}{Time\ of\ exposure\ x_i}$$

Table 1.5 Timeline of Entrepreneurship Events, GEM Colombia Data, 2017

Edad	*Women*					*Men*				
	Total Women in the Sample	*Eligible*		*Respondent*		*Total Men in the Sample*	*Eligible*		*Respondent*	
		Bstart	*Suacts*	*Ownmge*	*Discent*		*Bstart*	*Suacts*	*Ownmge*	*Discent*
		Start a Business	*Do an Action of Their Undertaking*	*Owner*	*Close Business*		*Start a Business*	*Do an Action of Their Undertaking*	*Owner*	*Close Business*
18–19	48	9	7	5	0	70	20	15	9	4
20–24	142	42	30	28	7	125	40	26	23	5
25–29	139	39	28	25	10	154	38	26	21	9
30–34	127	38	24	30	8	108	35	21	18	8
35–39	139	50	29	36	5	116	32	20	26	7
40–44	100	25	16	20	7	100	31	22	26	6
45–49	107	30	24	27	3	97	30	14	28	11
50–54	114	40	25	28	8	93	24	13	31	10
55–59	90	16	7	14	7	79	31	16	29	7
60–64	86	12	7	10	4	64	17	10	14	7
Total	**1,092**	**301**	**197**	**223**	**59**	**1,006**	**298**	**183**	**225**	**74**

Source: Own elaboration.

Being the denominator the time of exposure is from the age of 18 up until the birth event experience (in the reproductive age period for a woman, i.e., aged 18–45).

In the case of entrepreneurship (see Chapter 2), the global entrepreneurship rate is defined as follows:

$$GET = \frac{Number\ of\ undertakings}{Exposure\ time\ to\ experience\ an\ event}$$

This is where the exposure time is the time that elapses from the age of 18 up to the moment of undertaking an enterprise in the productive age (aged 18–64). The exposure time is measured in person years (Pressat, 2000). However, if there is no data on dates, by convention the population of the mid-term age group is taken as an average (see details in Chapter 2).

The Global Entrepreneurship Rate (GER) calculations for the case of Colombia using GEM 2017 data are presented in Table 1.6. A rate of 2,053 ventures is estimated in the productive life of an adult under the assumption of constant risk undertaken. Table 1.6 and Figure 1.3 also show the specific rates of entrepreneurship that integrate the GER, which allow for measuring entrepreneurship by five-year age groups at rates per 1,000 adults.

1.5.6 Comparison of Estimates (CE)

The construction of indicators is a key part of EDQ; however, it is important to compare the results with other sources in order to validate them. In this sense, this phase requires expert knowledge of the subject of the data source in order to be aware of other sources of information with useful data to construct similar indicators that allow for comparison.

Concerning entrepreneurship in Colombia, there are several sources of information. The search for data on the subject begins at national level with official institutions such as DANE[1] and CCB[2] and at the international level with institutions such as The World Bank, Euromonitor, GEM, Doing Business, etc.

The comparison requires official sources that have the required data. In the case of Colombia, the official data source at individual level is owned by the DANE institution which carried out many surveys and aggregated data population related to employment and unemployment at national level in 2017 through the Great Integrated Household Survey (GIHS) and data from the micro-business module of 2015. The data of the EAP[3] in the first semester of 2017 is taken as denominator and as numerator the amount of micro-businesses of the GIHS of the

Table 1.6 Specific Entrepreneurship Rates and GER by Five-year Age Groups

Five-Year Age Groups	*Total*			
	Total Population	*Do an Action of Their Undertaking*	*Specific Annual Rate*	*Rates by 1,000*
18–19	118	22	0.093	93.22
20–24	267	56	0.042	41.948
25–29	293	54	0.037	36.86
30–34	235	45	0.038	38.298
35–39	255	49	0.038	38.431
40–44	200	38	0.038	38
45–49	204	38	0.037	37.255
50–54	207	38	0.037	36.715
55–59	169	23	0.027	27.219
60–64	150	17	0.023	22.667
Total	**2,098**	**380**	**0.411**	
Global Entrepreneurship Rate:			2.053	

Source: Own elaboration.

Note: The first group is not a five-year age group because the APS interviews adults aged 18–64, but the method adjusts the two years of this group.

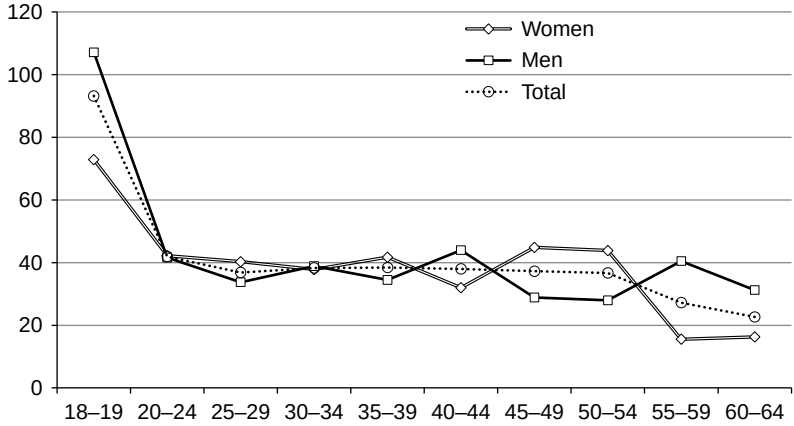

Figure 1.3 Specific Entrepreneurship Rates per Five-year Age Groups, GEM Colombia Data, 2017.

Source: Own elaboration.

Table 1.7 Estimation of the Proportion of Entrepreneurships in Colombia, 2017

Numerator	*2015*	*EAP 2017*	*PEP*
Total Number of Small Enterprises	4,662	24,581	18.97
Total Number of Small Enterprises of Employer	603	24,581	2.45
Total Number of Small Enterprises of Self-Employer	4,059	24,581	16.51

Source: Own elaboration.

Notes:
The source of the data for 2015 is the Big Survey Integrated of Homes of Colombia. The EAP was taken from the National Department Administrative of Statistics of Colombia, 2017.
EAP: Economically Active Population
PEP: Percentage of Entrepreneurship Population.

2015 national total, information of which is available. It is possible to adopt this approach, since there is no national data source similar to that collected by GEM Colombia.

As can be seen in Table 1.7, an approach to the Entrepreneurial Activity Rate (EAR) defined by the GEM is estimated to mean *Total Early Entrepreneurship Activity (TEA)* (see Chapter 2, this volume) which in this case takes as denominator the economically active population in Colombia in the first half of 2017 and as numerator the amount of micro-businesses of the 2016 total national GIHS for an EAR of 18.97 percent of enterprises in Colombia. The EAR of Colombia, according to the GEM in 2017, decreased significantly compared to 2016, from 27.6 percent to18.9 percent (GEM, 2017).

1.6 Conclusions

EDQ is a phase of data science which is not simple, but it is fundamental. An EDQ methodology allows us to formulate the scientific argument to support the quality of the source data and based on this data to perform more sophisticated processes such as DA predictions, and AI simulations in support of adequate decision making and close to reality (Argote, 2015).

EDQ is a thorough process of identifying patterns and anomalies. This chapter has presented techniques from engineering, mathematics, demography and statistics. According to the literature review, there are general processes that are part of EDQ; however, the proposed methodology follows a value chain from the simple to the complex that allows for concrete results to be obtained in a practical way.

The proposed methodology, unlike the revised ones, emphasizes the estimation of rates, analysis of the timelines and comparison of the

results with other data sources, which guarantees the valuation of coherence, validity and quality of the data considering both sampling and non-sampling errors. The methodology is available to all audiences, since it makes use of instruments such as tables, matrices and Excel and the SPSS for data registration.

Notes

1 National Administrative Department of Statistics.
2 Bogotá Chamber of Commerce.
3 Economically Active Population (EAP).
1 Theoretical definition of rate: the number of people exposed to experiencing the event between the exposure time measured in person years. Because the theoretical definition requires a longitudinal data design, it is neither easy nor feasible to calculate it frequently (Argote, 2007). However, under certain demographic assumptions, the law of large numbers and the central limit theorem, it can be assumed that the denominator approaches the average population over time during the period in which the rate is measured. Review the detail of the definition of a rate, specific dates and Global Entrepreneurship Rate in this chapter. This chapter uses the results of the calculation of this measure for 2008–2017 proposed and generated as a new entrepreneurship metric, presented in the *Training Seminar: Using GEM Data for Scientific Publications*. Bogotá, Colombia, June 2018, sponsored by the Global Entrepreneurship Monitor Consortium.
2 In the work of Argote (2007) there is a rigorous analysis on the subject of assumptions in demography and her thesis that methods that capture the nonlinear dynamics of populations are required.
3 The details of the estimation for the GER and the SER are found in Chapter 2. In this chapter only the results of the methodology developed in Chapter 2 are used.
1 Unstructured data (or unstructured information) is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy but may also contain data such as dates, numbers and facts. This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents.

(https://en.wikipedia.org/wiki/Unstructured_data)

References

Argote Cusi, M. (2003). Evaluación de la calidad del dato. In Estimación de la distribución estadística de la Tasa Global de Fecundidad mediante remuestreo, retrieved from: http://conocimientoabierto.flacso.edu.mx/tesis/103.

Argote Cusi, M. (2007). Estimación de la distribución estadística de la Tasa Global de Fecundidad. *Papeles de Población*, 54(13), 87–113.

Argote Cusi, M. (2009). Comparación y evaluación de la distribución

estadística del estimador de la tasa global de fecundidad de Bolivia en 1998 y 2003. *Papeles de Población*, 62(15), 201–222.
Argote Cusi, M. (2015). Sensitivity analysis of projections population. *Papeles de Población*, 84(21), 45–67.
Argote-Cusi, M. L. (2018). El uso de lógica difusa en proyecciones de población: el caso de México. *Papeles de población*, 24(95), 273–301.
Azeroual, O., Saake, G. & Abuosba, M. (2019). Data quality measures and data cleansing for research information systems. *arXiv preprint* arXiv:1901.06208.
Hadhiatma, A. (2018). Improving data quality in the linked open data: A survey. *Journal of Physics: Conference Series*, 978(1), 012026.
Heredia, R., Jobany, J. & Vilalta Alonso, J. A. (2009). La calidad de los datos: su importancia para la gestión empresarial. *Libre Empresa*, 6(1), 43–50.
Maydanchik, A. (2007). *Data Quality Assessment*. Technics Publications, Basking Ridge, NJ.
Moreno-Altamirano, A., López-Moreno, S. & Corcho-Berdugo, A. (2000). Principales medidas en epidemiología. *Salud pública de México*, 42, 337–348.
Pressat, R. (2000). *El análisis demográfico: Métodos, resultados, aplicaciones*. Fondo de cultura Económica, Mexico City.
Reynolds, P., Bosma, N., Autio, E., Hunt, S., De Bono, N., Servais, I. & Chin, N. (2005). *Global Entrepreneurship Monitor: Design and Implementation 1998–2003* (No. 1101). Global Entrepreneurship Research Association.
Zúñiga Segura, L. & Sánchez Godínez, E. (2012). Calidad de datos y su evaluación: un caso de estudio. *Calidad en la Educación Superior*, 3(2), 33–49.
Acs, Z. J. & Szerb, L. (2007). Entrepreneurship, economic growth and public policy. *Small Business Economics*, 28(2–3), 109–122.
Acs, Z. & Szerb, L. (2009). The Global Entrepreneurship Index (GEINDEX). *Foundations and Trends® in Entrepreneurship*, 5(5), 341–435. http://dx.doi.org/10.1561/0300000027.
Acs, Z. J., Autio, E. & Szerb, L. (2014). National systems of entrepreneurship: Measurement issues and policy implications. *Research Policy*, 43(3), 476–494.
Acs, Z., Desai, S. & Klapper, F. (2007). A comparison of GEM and the World Bank Group entrepreneurship data. In *Third GEM Research Conference: Entrepreneurship, Economic Development and Public Policy*.
Acs, Z., Szerb, L. & Lloyd, A. (2018). The global entrepreneurship index. In A. Lloyd (Ed.), *Global Entrepreneurship Index 2018*. The Global Entrepreneurship and Development Institute, Washington, DC, pp. 1–44.
Akerlof, G. & Shiller, J. R. (2016). La economía de la manipulación, como caemos como incautos en las trampas del mercado. *Deusto S.A. Ediciones*, Grupo Planeta, Madrid.
Argote Cusi, M. (2003). Evaluación de la calidad del dato. In Estimación de la distribución estadística de la Tasa Global de Fecundidad mediante re-muestreo, descargado: http://conocimientoabierto.flacso.edu.mx/tesis/103.
Argote Cusi, M. (2007). Estimación de la distribución estadística de la Tasa Global de Fecundidad. *Papeles de Población*, 54(13), 87–113.

Argote Cusi, M. (2009). Comparación y evaluación de la distribución estadística del estimador de la tasa global de fecundidad de Bolivia en 1998 y 2003. *Papeles de Población*, 62(15), 201–222.

Argote Cusi, M. (2015). Sensitivity analysis of projections population. *Papeles de Población*, 84(21), 45–67.

Argote Cusi, M. & Parra, L. (2018). Working Paper: Evaluación de la Calidad del Dato GEM. Training Seminar using GEM data for Scientific Publications, developed 4–8 June 2018, Bogotá, Colombia. https://universidadean.edu.co/es/noticias/fuimos-sede-de-importante-seminario-de-datos-del-gem.

Astebro, Z. A. T. & Robinson, D. A. (2016). Public policy to promote entrepreneurship: A call to arms. *Small Business Economics*, 47, 35–51.

Audretsch, D. B., Kuratko, D. F. & Link, A. N. (2016). Dynamic entrepreneurship and technology-based innovation. *Journal of Evolutionary Economics*, 26(3), 603–620.

Cantillon, R. (2015 [1755]). *An Essay on the Nature of Trade in General*. The Liberty Fund, Indianapolis, IN.

Chowdhury, F., Terjesen, S. & Audretsch, D. (2015). Varieties of entrepreneurship: Institutional drivers across entrepreneurial activity and country. *European Journal of Law and Economics*, 40(1), 121–148.

Entrepreneurship Research Conference (2017). Babson College Entrepreneurship Research Conference (BCERC) and Doctoral Consortium co-sponsored by The University of Oklahoma, USA, June 7–10.

GEM (2017). Global Report 2017. www.gemconsortium.org/report/49812.

GEM (2018). Training seminar for scientific publications using GEM data. https://universidadean.edu.co/es/noticias/fuimos-sede-de-importante-seminario-de-datos-del-gem.

Gómez, L., López, S., Hernández, N., Galvis, M., Varela, R., Moreno, J., Pereira, F., Parra, L., Matíz, F., Cediel, G. & Martínez, P. (2018). *GEM Colombia: Estudio de la Actividad Empresarial en 2017*. UNINORTE, Barranquilla.

Gut Allan (2013). *Probability: A Graduate Course* (Vol. 75). Springer Science & Business Media, New York.

Henrekson, M. & Sanandaji, T. (2014). Small business activity does not measure entrepreneurship. *Proceedings of the National Academy of Sciences*, 111(5), 1760–1765.

Kantis, H., Federico, J. & Menéndez, C. (2012). Políticas de fomento al emprendimiento dinámico en América Latina: tendencias y desafíos. CAF Working Paper No. 2012/09, August 2012.

Maritz, A., Zolin, R., De Waal, A., Fisher, R., Perenyi, A. & Eager, B. (2015). Senior entrepreneurship in Australia: Active ageing and extending working lives. *International Journal of Organizational Innovation*, 7(1), 1–39.

Moreno-Altamirano, A., López-Moreno, S. & Corcho-Berdugo, A. (2000). Principales medidas en epidemiología. *Salud pública de México*, 42, 337–348.

Parra, L. & Argote, M. (2013). La gestión en el proceso de creación empresarial: el caso de IN3 de la Universidad EAN de Colombia. Emprendimiento: diferentes aproximaciones. Universidad EAN: Bogotá.

Parra, L. & Argote, M. (2017). Data analytics to characterize university-based companies for decision making in business development programs. In E. Rodriguez, *Data Analytics Applications in Latin America and Emerging Economies*, pp. 187–205. CRC Press, Abingdon, pp. 187–205.

Parra, L. & Argote, M. (2018). *Academia, emprendimiento e investigación empresarial: homenaje a la Universidad EAN en sus 50 años.* Ediciones Universidad EAN, Bogotá.

Parra, L., Argote, M. & Farro, T. (2018). Emprendimiento Universitario: Análisis de contraste entre la Universidad EAN – Colombia y la Universidad Continental – Perú. In L. Parra & M. Argote, *Academia, emprendimiento e investigación empresarial: homenaje a la Universidad EAN en sus 50 años*. Ediciones Universidad EAN, Bogotá.

Pressat, Roland (2000). *El análisis demográfico: Métodos, resultados, aplicaciones*. Fondo de cultura Económica, Mexico City.

Preston, S. H., Heuveline, P. & Guillot, M. (2001). *Demography: Measuring and Modelling Population Processes*. Blackwell, Oxford.

Reynolds, P., Bosma, N., Autio, E., Hunt, S., De Bono, N., Servais, I. & Chin, N. (2005). Global entrepreneurship monitor: Data collection design and implementation 1998–2003. *Small Business Economics*, 24(3), 205–231.

Rowland, D. T. (2003). *Demographic Methods and Concepts*. Oxford University Press, New York.

Schumpeter, J. A. (1980 [1934]). Change and the Entrepreneur. In *The Theory of Economic Development*. Oxford University Press, Oxford.

Siegel, J. S. & Swanson, D. A. (2004). *The Methods and Materials of Demography* (2nd edn). Elsevier, San Diego, CA.

Stewart, I. (2007). *La historia de las matemáticas, en los últimos 10.000 años*. Editorial Crítica, Barcelona.

Urbano, D., Aparicio, S. & Audretsch, D. (2018). Twenty-five years of research on institutions, entrepreneurship, and economic growth: What has been learned? *Small Business Economics*, 53(1), 21–49.

Abbasov, A. & Mamedova, M. (2003). Application of fuzzy time series to population forecasting. *Vienna University of Technology*, 1, 545–552.

Aladag, S., Aladag, C. H., Mentes, T. & Egrioglu, E. (2012). A new seasonal fuzzy time series method based on the multiplicative neuron model and SARIMA. *Hacettepe Journal of Mathematics and Statistics*, 41(3), 145–163.

Alho, J. M. (2014). Forecasting demographic forecasts. *International Journal of Forecasting*, 30(4), 1128–1135.

Alho, J. M. & Spencer, B. D. (1985). Uncertain population forecasting. *Journal of the American Statistical Association*, 80(390), 306–314.

Alho, J. & Spencer, B. (2006). *Statistical Demography and Forecasting*. Springer Science & Business Media, New York.

Alho, J., Alders, M., Cruijsen, H., Keilman, N., Nikander, T. & Pham, D. Q. (2006). New forecast: Population decline postponed in Europe.

Statistical Journal of the United Nations Economic Commission for Europe, 23(1), 1–10.
Argote Cusi, M. (2007). Estimation of the statistical distribution of the Global Fertility Rate. *Papeles de Población*, 54(13), 87–113.
Argote Cusi, M. (2009). Comparison and evaluation of the statistical distribution of the estimator of the total fertility rate of Bolivia in 1998 and 2003. *Papeles de Población*, 62(15), 201–222.
Argote Cusi, M. (2012). Analysis of sensitivity of births to small changes in the Global Fertility Rate. *Papeles de Población*, 72(18), 85–112.
Argote Cusi, M. (2015). Análisis de sensibilidad de proyecciones de población. *Papeles de Población*, 84, 45–67, April/June.
Argote Cusi, M. L. (2016). *Uso de la lógica difusa en proyecciones de población. Paper presented in the XIII National Meeting of Demographic Research, 22–24 June, UNAM*, Mexico.
Argote Cusi, M. L. (2018). El uso de lógica difusa en proyecciones de población: el caso de México. *Papeles de población*, 24(95), 273–301.
Argote Cusi, M. & Parra Bernal, L. D. (2017). Data analytics to characterize university-based companies for decision making in business development programs. *In Data Analytics Applications in Latin America and Emerging Economies*. CRC Press, Abingdon, pp. 187–205.
Bas, E., Uslu, V. R., Aladag, C., Yolcu, U. & Egrioglu, E. (2014). A modified genetic algorithm for forecasting fuzzy time series. *Applied Intelligence*, 41, 453–463.
Bosma, N., Coduras, A., Litovsky, Y. & Seaman, J. (2012). GEM Manual: A report on the design, data and quality control of the Global Entrepreneurship Monitor. *Global Entrepreneurship Monitor*, 9.
Burney, S. A., Ali, S. M. & Khan, M. S. (2018). A novel high order Fuzzy Time Series forecasting method with higher accuracy rate. *International Journal of Computer Science and Network Security*, 18(5), 13–40.
CEPAL (2009). Proyección de población. Demographic Observatory Latin American and Caribbean, Year IV, No. 7, April. Publication developed by Guiomar Bay.
Chen, S. M. (2002). Forecasting enrollments based on high-order fuzzy time series. *Cybernetics and Systems: An International Journal*, 33, 1–16.
Chen, S. M. & Chen, D. C. (2011). TAIEX forecasting based on fuzzy time series and fuzzy variation groups. *IEEE Transactions on Fuzzy Systems*, 19, 1–12.
Chen, S. M. & Hsu, C. C. (2004). A new method to forecast enrollments using fuzzy time series. *International Journal of Applied Science and Engineering*, 2(3), 234–244
Chen, S. M. & Hwang, J. R. (2000). Temperature prediction using fuzzy time series. *IEEE Transactions on Systems,Man, and Cybernetics-Part B: Cybernetics*, 30, 263–275.
Chen, S. M. & Kao, P. Y. (2013). TAIEX forecasting based on fuzzy time

series, particle swarm optimization techniques and support vector machines. *Information Sciences*, 247, 62–71.
Chen, S. M. & Tanuwijaya, K. (2011). Fuzzy forecasting based on high-order fuzzy logical relationships and automatic clustering techniques. *Expert Systems with Applications*, 38, 15425–15437.
Chen, T. L. & Shiu, S. Y. (2007). A new clustering algorithm based on self-updating process. *In JSM Proceedings, Statistical Computing Section*, Salt Lake City, UT, pp. 2034–2038.
Dubois, D., Ostasiewicz, W. & Prade, H. (2000). Fuzzy sets: History and basic notions. *In Fundamentals of Fuzzy Sets*. Springer, Boston, MA, pp. 21–124.
Egrioglu, S., Bas, E., Aladag, C. H. & Yolcu, U. (2016). Probabilistic fuzzy time series method based on artificial neural network. *American Journal of Intelligent Systems*, 62(2), 42–47.
Ghosh, H., Chowdhury, S. & Prajneshu, S. (2015). An improved fuzzy time series method of forecasting based on L-R fuzzy. *Journal of Applied Statistics*, 43(6), 1128–1139.
Jang J. S. R., Sun, C-T. & Mizutani, E. (1997). *Neuro-Fuzzy and Soft Computing. A Computational Approach to Learning and Machine Intelligence*. Prentice-Hall, London.
Keyfitz, N. (1981). The limits of population forecasting. *Population and Development Review*, 27(4), 579–593. Ed. Population Council, Washington, DC.
Kopcso, D. & Pachamanova, D. (2017). Case article: Business value in integrating predictive and prescriptive analytics models. *INFORMS Transactions on Education*.
Kosko, B. (1992). *Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence*, vol. 1. Prentice Hall, London.
Lee, R. D. & Tuljapurkar, S. (2000). Population forecasting for fiscal planning: Issues and innovations. Ed. UC Berkeley, CEDA. Available at https://escholarship.org/content/qt7n02r268/qt7n02r268.pdf.
Lee, S. M. (1998). Asian Americans: Diverse and growing. *Population Bulletin*, 53(2), 1.
Martín del Brío, B. & Sanz Molina, A. (2002). *Redes neuronales y sistemas difusos*. Alfaomega, Mexico City.
Martín, B., Medrano, N., Pollán, T. & Sanz, A. (1970). *Redes neuronales y sistemas borrosos: un libro de texto en español.* Ed. Universidad de Zaragoza.
Mohammad, A. & Hamisu, A. A. (2017). A novel two-factor high order fuzzy time series with applications to temperature and futures exchange forecasting. *Nigerian Journal of Technology*, 36(4), 1124–1134.
Nauck, D. & Kruse, R. (1997). A neuro-fuzzy method to learn fuzzy classification rules from data. *Fuzzy Sets and Systems*, 89, 277–288.
Rayer, S. (2008). Population forecast errors: A primer for planners. *Journal of Planning Education and Research*, 27(4), 417–430.
Rutkowska, D. (2002). Type 2 fuzzy neural networks: An interpretation based on fuzzy inference neural networks with fuzzy parameters. In *Fuzzy*

Systems. Proceedings of the 2002 IEEE International Conferenc, vol. 2, pp. 1180–1185.
Sasu, A. (2010). An application of fuzzy time series to the Romanian population. *Bulletin of the Transilvania University of Brasov*, 3, 52.
Silverman, E., Bijak, J., Hilton, J., Cao, V. D. & Noble, J. (2013). When demography met social simulation: A tale of two modelling approaches. *Journal of Artificial Societies and Social Simulation*, 16(4), 9.
Song, Q. & Chissom, B. S. (1993). Fuzzy time series and its models. *Fuzzy Sets and Systems*, 54(3), 269–277.
Song, Q. & Chissom, B. S. (1994). Forecasting enrollments with fuzzy time series: Part II. *Fuzzy Sets and Systems*, 62(1), 1–8.
Sullivan, J. & Woodall, W. H. (1994). A comparison of fuzzy forecasting and Markov modeling. *Fuzzy Sets and Systems*, 64(3), 279–293.
Taleb, N. N. (2007). *El Cisne Negro, el impacto de la altamente improbable*. Editorial Paidós, Madrid.
Vovan, T. (2019). An improved fuzzy time series forecasting model using variations of data. *Fuzzy Optimization and Decision Making*, 18(2), 151–173.
Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338–353.
Zadeh, L. A. (1973). Outline of a new approach to the analysis of complex systems and decision processes. *Systems, Man and Cybernetics, IEEE Transactions*, 1, 28–44.
Zhi-xin, J., Hong-bin, Z. & An-min, X. (2009). Research in method of complex system reliability evaluation based-on fuzzy sets. In *Intelligent Systems and Applications, 2009. ISA 2009. International Workshop*, pp. 1–4.
Acs, Z. J. & Szerb, L. (2007). Entrepreneurship, economic growth and public policy. *Small Business Economics*, 28(2–3), 109–122.
Acs, Z., Astebro, T., Audretsch, D. & Robinson, T. (2016). Public policy to promote entrepreneurship: A call to arms. *Small Business Economics*, 47, 35–51. DOI. 10.1007/s11187-016-9712-2.
Agarwal, A., Xie, B., Vovsha, I., Rambow, O. & Passonneau, R. (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media* (pp. 30–38). Association for Computational Linguistics, UK.
Aguilar, L. J. (2016). *Big Data, Análisis de grandes volúmenes de datos en organizaciones*. Alfaomega Grupo, Madrid.
Álvarez, P., García, S. I., Menéndez, C., Federico, J. & Kantis, H. (2016). El ecosistema emprendedor de la Ciudad Autónoma de Buenos Aires. Una mirada exploratoria. *Pymes, Innovación y Desarrollo*, 4(1). UBA, Buenos Aires, Argentina.
Anghelache, C., Anghel, M. G. & Solomon, A. G. (2017). National accounts system: Source of information in macroeconomic forecast. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, 7(2), 76–82.
Argote, M. & Parra, L. (2016) Marco conceptual para el análisis de brechas tecnológicas en el sector metalmecánico. In L. Parra, *Análisis de brechas*

tecnológicas en el sector metalmecánico desde el estudio de casos de contraste. EAN University, Bogotá, p. 95.
Audretsch, D. (2012). Entrepreneurship research. *Management Decision*, 50(5), 755–764.
Autio, E., Rannikko, H., Handelberg, J. & Kiuru, P. (2014). *Analyses on the Finnish High-Growth Entrepreneurship Ecosystem*. Aalto University Publication Series BUSINESS + ECONOMY, 1.
Belitski, M., Chowdhury, F. & Desai, S. (2016). Taxes, corruption, and entry. *Small Business Economics*, 47(1), 201–216.
Benavente, J. M. & Crespi, G. (2016). Towards a theoretical approach to national systems of innovation. *Estudios de Economía*, 22(2), 243.
Caffo, B., Peng, R. D. & Leek, R. H. (2016). *Executive Data Science: A Guide to Training and Managing the Best Data Scientists*. Lean Publishers, Victoria, BC.
Canibano, L. & Sánchez, M. P. (2009). Intangibles in universities: Current challenges for measuring and reporting. *Journal of Human Resource Costing & Accounting*, 13(2), 93–104.
Chen, C. L. P. & Zhang, C-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, 314–347.
Chen, H., Chiang, R. H. & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188.
Crespi, G., Katz, J. & Olivari, J. (2017). Innovation, natural resource-based activities and growth in emerging economies: The formation and role of knowledge-intensive service firms. *Innovation and Development*, 1–23.
Curry, E. (2016). The Big Data value chain: Definitions, concepts, and theoretical approaches. In J. Cavanillas, E. Curry & W. Wahlster (eds), *New Horizons for a Data-driven Economy*. Springer, Champaign, IL.
Eatwell, J. (2016). International capital liberalization: The impact on world development. *Estudios de economía*, 24(2), 219.
Elgendy, N. & Elragal, A. (2016). Big Data analytics in support of the decision-making process. *Procedia Computer Science*, 100, 1071–1084.
Estrin, S. & Mickiewicz, T. (2013). Entrepreneurship in transition economies: The role of institutions and generational change. In M. Miniti, *The Dynamics of Entrepreneurship: Evidence from the Global Entrepreneurship Monitor Data*. Oxford University Press, Oxford, pp. 181–208.
Fuerlinger, G., Fandl, U. & Funke, T. (2015). The role of the state in the entrepreneurship ecosystem: Insights from Germany. *Triple Helix*, 2(1), 3.
Gobble, M. A. (2013) Big Data: The next big thing in innovation. *Research and Technology Management*, 56(1), 64–66.
Hausman, D., McPherson, M. & Satz, D. (2016). *Economic Analysis, Moral Philosophy, and Public Policy*. Cambridge University Press, Cambridge.
Horita, F. E., de Albuquerque, J. P., Marchezini, V. & Mendiondo, E. M. (2017). Bridging the gap between decision-making and emerging big data

sources: An application of a model-based framework to disaster management in Brazil. *Decision Support Systems*, 97, 12–22.
Iamsiraroj, S. (2016). The foreign direct investment–economic growth nexus. *International Review of Economics & Finance*, 42, 116–133.
Kantis, H., Federico, J. & Ibarra, S. (2014). Índice de condiciones sistémicas para el emprendimiento dinámico. In *Una herramienta para la acción en America Latina*. BID, Washington, DC.
Kantis, H., Postigo, S., Federico, J. & Tamborini, F. (2002). El surgimiento de emprendedores de base universitaria: en qué se diferencian? Evidencias empíricas para el caso de Argentina. In *Presentado en: RENT XVI Conference, Barcelona*.
Katz, J. (2017). The Latin American transition from an inward-oriented industrialization strategy to a natural resource-based model of economic growth. *Institutions and Economies*, 7(1), 9–22.
Kelley, D., Singer, S. & Herrington, M. (2016). *2015/2016 Global Report*. GEM Global Entrepreneurship Monitor, Babson College, Universidad del Desarrollo, Universiti Tun Abdul Razak, Tecnológico de Monterrey, International Council for Small Business (ICSB).
Kim, G. H., Trimi, S. & Chung, J. H. (2014). Big-data applications in the government sector. *Communications of the ACM*, 57(3), 78–85.
Kościelniaka, H. & Puto, A. (2015). BIG DATA in decision making processes of enterprises. *Procedia Computer Science*, 65, 1052–1058.
Leyden, D. P. & Link, A. N. (2013). Knowledge spillovers, collective entrepreneurship, and economic growth: The role of universities. *Small Business Economics*, 41(4), 797–817.
Leyden, D. P. & Link, A. N. (2014). A theoretical analysis of the role of social networks in entrepreneurship. *Resources Policy*, 43(7), 1157–1163.
Link, A. N. & Link, J. R. (2007). *Government as Entrepreneur*. Oxford University Press, New York.
Mallinger, M. & Stefl, M. (2015). Big Data decision making. *Graziadio Business Review*, 18(2).
Martinez-Lopez, L. & Martinez-Lopez, F. J. (2010). Intelligent e-services and multi-agent systems for B2C e-commerce. *Internet Research*, 20(3).
McAfee, A. & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, 90(10), 60–68.
Minniti, M. (2013). *The Dynamics of Entrepreneurship: Evidence from Global Entrepreneurship Monitor Data*. Oxford University Press, Oxford.
Navicke, J., Rastrigina, O. & Sutherland, H. (2014). Nowcasting indicators of poverty risk in the European Union: A microsimulation approach. *Social Indicators Research*, 119(1), 101–119.
Nunns, J. & Rosenthal, S. (2016). Financial transaction taxes in theory and practice. *National Tax Journal*, 69(1), 171–216.
Obschonka, M. (2017). The quest for the entrepreneurial culture: Psychological Big Data in entrepreneurship research. *Current Opinion in Behavioral Sciences*, 18, 69–74.

Ostrom, E. (2014). Collective action and the evolution of social norms. *Journal of Natural Resources Policy Research*, 6(4), 235–252.

Pak, A. & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, 10, 1320–1326.

Parra, L. & Piñeros, A. (2016) ¿Cuál es el rol del Estado en la promoción del Sector Metalmecánico en Colombia? In L. Parra, *Análisis de brechas tecnológicas en el sector metalmecánico desde el estudio de casos de contraste* (p. 95). Ed. EAN University, Bogotá.

Reynolds, P. D., Camp, S. M., Bygrave, W. D., Autio, E. & Hay, M. (2001). *GEM Global Entrepreneurship Report, 2001 Summary Report*. Kauffman Center for Entrepreneurial Leadership at the Ewing Marion Kauffman Foundation, Kansas City, KA.

Reynolds, P. D., Bosma, N., Autio, E., Hunt, S., De Bono, N., Servais, I., Lopez-Garcia, P. & Chin, N. (2005). Global Entrepreneurship Monitor: Data collection design and implementation 1998–2003. *Small Business Economics*, 24(3), 205–231.

Rothaermel, F. T., Agung, S. D. & Jiang, L. (2007). University entrepreneurship: A taxonomy of the literature. *Industrial and Corporate Change*, 16(4), 691–791.

Schiller, S., Goul, M., Iyer, L., Sharda, R. & Schrader, D. (2014). Panel: Build Your Dream (not just Big) Analytics Program. In *Proceedings of the Twentieth Americas Conference on Information Systems (AMCIS)*, Savannah, GA.

Schiller, S., Goul, M., Iyer, L. S., Sharda, R., Schrader, D. & Asamoah, D. (2015). Build Your Dream (not just Big) Analytics Program. *Communications of the Association for Information Systems*, 37(40). Available at http://aisel.aisnet.org/cais/vol37/iss1/40.

Schroeck, R., Shockley, J., Smart, D., Romero-Morales, P. & Tufano (2012). Analytics: The real-world use of big data. How innovative enterprises extract value from uncertain data. IBM Institute for Business Value. Retrieved from www-03.ibm.com/systems/hu/resources/the_real_word_use_of_big_data.pdf.

Stiglitz, J. E. & Greenwald, B. C. (2016). *La creación de una sociedad del aprendizaje: Una nueva aproximación al crecimiento, el desarrollo y el progreso social*. La Esfera de los Libros, Madrid.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G. & Gnanzou, D. (2015). How "big data" can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165, 234–246.