Spanish Journal of Statistics

VOLUME 7, NUMBER 1, 2025

INE

Spanish Journal of Statistics

NIPO: 222-24-003-9
ISSN: 2695 - 9070
DOI: :https://doi.org/10.37830/SJS.2025.1

Published by the Spanish National Statistical Institute

Spanish Journal of Statsitics

VOLUME 7, NUMBER 1, 2025

Contents

SPANISH JOURNAL OF STATISTICS
Vol. 7 No. 1, 1025, Pages 5 – 8
doi:https://doi.org/10.37830/SJS.2025.1.01

Editorial, Spanish Journal of Statistics

Presentation of Volume 7, 1, 2025

José María Sarabia

Editor-in-Chief Spanish Journal of Statistics

University of Cantabria, Spain

Dear readers and dear members of the statistical community,

It is my pleasure to present Volume 7, Issue 1, corresponding to the year 2025. This volume consists of six interesting articles, five articles in the general statistics section and one invited article, related to the 2024 Spanish National Award in Statistics.

The first article is entitled "Copula based inference for certain types of actuarial datasets: A brief survey", whose authors are Indranil Ghosh, Carman Greenway, Hannah Powers, and Brandon Kelly Kortan, from the University of North Carolina, Wilmington, USA. Copulas provide a powerful framework for modeling bivariate and multivariate dependence structures. In this study, the authors investigate several forms of dependence characterized by well-established dependence measures, including Spearman's ρ, Kendall's τ, and Blomqvist's β, for selected actuarial datasets obtained from the CAS data repository in the `R` software package. The research primarily focuses on insurance claim datasets, the proposed copula-based methodology can be readily applied to other actuarial data types and broader domains. Using the `CDAvine` package in `R`, the authors identify the best-fitting bivariate copula for each dataset and subsequently examine various structural properties of these optimal copula models. This approach also provides a basis for extending the analysis to multivariate dependence using vine copula constructions, which will be addressed in a separate article.

The next article is entitled, "Gender gap and spatial disparities in the evolution of literacy in Spain, 1860-1910" and its authors are José Manuel Gutiérrez, University of Salamanca, Spain and Gloria Quiroga, Complutense University of Madrid, Spain. This article examines the dynamics of Spanish literacy between 1860 and 1910, a period in which local councils were responsible for public elementary education. To this end, the authors construct a harmonized series of literacy rates for the population aged ten or older, disaggregated by sex and province. Marked spatial differences and a very large gender gap are observed. Five clusters are identified according to provincial male literacy rates in 1860; these clusters retain explanatory power throughout the period and for both sexes. A parsimonious statistical model of the evolution of male literacy, incorporating linguistic variables, shows considerable temporal stability in its spatial distribution. The model for female literacy displays similarities to that of male literacy, although in this case the initial condition (in 1860) is explained not by female literacy but by male literacy. Overall, the evolution of literacy in Spain between 1860 and 1910 did not follow the spatial pattern of economic modernization. Moreover, there was no correlation between birth rates and children's literacy rates for either sex, nor between urbanization and literacy. In the broader Western European context, Spain's literacy trajectory during this period was largely a failure, except for the provinces in the top cluster.

The third article is entitled "Feasibility of Implementing Accelerometers in the Spanish Health Survey", whose authors are Borja del Pozo Cruz, Department of Sport Sciences, European University of Madrid, Spain; Rosa M. Alfonso Rosa, Department of Human Motricity and Sport Performance, University of Seville, Spain and Jesús del Pozo-Cruz, Department of Physical Education and Sports, University of Seville, Spain. Accurately measuring physical activity, sedentary behavior, and sleep is vital for public health monitoring, but self-reported data are often biased. Accelerometers provide objective information, yet their feasibility within the Spanish Health Survey, ESdE, has not been assessed. In this study the authors evaluated the integration of thigh-worn accelerometers in the ESdE by analyzing participant compliance, device usability, data return, and comparisons with self-reported measures. A total of 100 adults aged 30–90 were recruited through five provincial delegations of the National Statistics Institute, INE, with each delegation enrolling 20 participants equally divided between home-based collection and prepaid-return groups. All participants wore a thigh-mounted

SENS accelerometer continuously for 7 to 10 days using a water-resistant patch, with two patches provided in case replacement was needed. INE staff administered the ESdE questionnaire and coordinated device logistics. Valid accelerometry data were obtained from 98 participants, showing excellent compliance. Device return rates were 100% in the collection group and 85% in the prepaid-return group. Comparison with self-reported data was only possible for sedentary behavior, where participants consistently underestimated sitting time. Agreement between self-reports and accelerometry was low, and Bland–Altman plots revealed a clear negative bias. These findings demonstrate the feasibility of incorporating accelerometry into national surveys like the ESdE, with high participant adherence and minimal operational issues. The objective data provided by accelerometers can complement self-reported measures and capture domains such as sleep and incidental activity, which are often overlooked. The authors conclude that their inclusion in future surveys could improve the accuracy and usefulness of lifestyle surveillance in Spain.

The fourth article is entitled, "Semiparametric von Mises kernel circular density estimator", whose authors are Yasmina Ziane, Nabil Zougab and Kahina Bedouhene, from the Research unit LaMOS, Faculty of Exact Sciences, Bejaia University, Algeria and Kahina Bedouhene, from the Department of Mathematics, University of Tizi-Ouzou, Algeria. In this article, the authors propose estimating the circular density function using a bias-corrected semiparametric circular kernel method based on the von Mises kernel. This approach applies a multiplicative bias correction to the initial parametric model to improve both the estimator's quality and its bias properties. Two semiparametric estimators-Hjort and Glad (1995) (HG) and Jones, Signorini, and Hjort (1999) (JSH) are applied to estimate the probability density of circular data with support $[0, 2\pi]$. Their properties, including bias, variance, and integrated mean square error (MSE), are presented. A comparative study is performed to evaluate the performance of the semiparametric estimators (HG and JSH). The commonly used cross-validation technique is adapted for bandwidth selection. Finally, a simulation study and an application with real data for circular data illustrate, in terms of integrated squared bias (ISB) and integrated squared error (ISE), that the JSH and JLN semiparametric estimators with the von Mises kernel outperform the classical and HG estimators.

The fifth article in this volume is entitled, "Objective Bayesian goodness-of-fit tests for the alpha-skew-normal distribution", whose authors are José Rodolfo Olmos-Zepeda and Sergio Pérez-Elizalde, both in the Department of Statistics and Data Science, Colegio de Postgraduados, Mexico. The family of alpha-skew-normal (ASN) distributions constitutes a flexible class of three-parameter probability models defined by their location, scale, and shape. The shape parameter controls both asymmetry and uni-/bimodality, enabling the distribution to capture unimodal or bimodal patterns with varying degrees of skewness. The authors introduce an objective Bayesian goodness-of-fit test for assessing whether a random sample arises from an ASN distribution when the parameters are unknown. The proposed test statistics are constructed from the empirical distribution function, whose sampling distributions depend exclusively on the shape parameter. Their prior predictive distributions, serving as null distributions, are obtained by integrating out the shape parameter with respect to a proper approximation to Jeffreys prior—specifically, a Cauchy prior—selected for its analytical tractability. Critical values are obtained through Monte Carlo simulation. A thorough simulation study shows that the proposed tests maintain the nominal significance level under a wide range of conditions and display strong power against various alternative distributions. The methodology is further illustrated through real-data applications, demonstrating its practical relevance.

The sixth article is an invited contribution based on the speech delivered by Professor Concepción Bielza upon receiving the 2024 Spanish National Statistics Prize, a ceremony made especially notable by the presence of His Majesty the King of Spain. The article offers a personal perspective on her research career, characterized by the long-standing interplay between statistics, artificial intelligence, and their applications in neuroscience and industry. After beginning her professional career in statistical decision theory and probabilistic graphical models, Bayesian networks soon became the central framework of her work, enabling rigorous reasoning under uncertainty across a wide range of fields.

Finally, I would like to express my gratitude to all the authors contributing to this volume for choosing our journal as a platform for disseminating their research. I also wish to acknowledge the work of the editors and reviewers, whose efforts help uphold a high standard of scientific quality.

Spanish Journal of Statistics
Vol. 7 No. 1 2025, Pages 9–38
doi:https://doi.org/10.37830/
SJS.2025.1.02

Regular Article

Copula based inference for certain types of actuarial datasets: A brief survey

Indranil Ghosh[1], Carman Greenway[2], Hannah Powers[3], and Brandon Kelly Kortan[4]

[1]University of North Carolina, Wilmington, ghoshi@uncw.edu
[2]University of North Carolina, Wilmington, cdg8092@uncw.edu
[3]University of North Carolina, Wilmington, hep6882@uncw.edu
[4]University of North Carolina, Wilmington,bkk6389@uncw.edu

Received: May 26, 2025. Returned: -. Revised: -. Accepted: October 6, 2025.

Abstract: Copula is a useful tool for modeling bivariate/multivariate dependency structures among others. In this paper, we aim to study various types of dependence indicated by well-known measures of dependence such as Spearman's ρ, Kendall's τ, and Blomqvist's β etc., for certain types of actuarial datasets, which is obtained from the CAS datasets in R software package. Although our primary focus is on the insurance claim datasets, the adopted copula-based procedure can be mimicked in other types of actuarial datasets and other domains as well. On using the CDA vine package in R, we find the best fitted bivariate copula for a given dataset, and subsequently study various structural properties of the derived best fitted bivariate copula. The adopted strategy can be envisioned in identifying and exploring multivariate dependence via Vine copula strategy which will be discussed in a separate article.

Keywords: Bivariate copula; Measures of Association; Kendall's tau; Copula fitting

MSC: 62H05, 62H20, 91G70

1 Introduction

Copula, Latin for "link, tie, or bond," in the mathematical world, is a function that allows us to combine univariate distributions to obtain a joint distribution with a particular dependency structure, according to the work of Durante and Sempi (2016). While copulas were first formally introduced to the mathematical and statistical world in 1959, the idea was not a foreign concept, as primitive versions of copulas were seen in Wassily Hoeffding's works as early as 1940. Hoeffding established possible bounds for these functions and studied measures of dependence invariant under strictly increasing transformations. Maurice Fréchet's work in 1951 also had ideas similar to the copula present, as his work on bounds for joint distributions with given marginals laid the foundation for copulas.

Copulas, introduced by Abe Sklar in 1959, are a fundamental resource in defining dependency structure between random variables. Sklar's theorem allows for a transformation of dependency

structure to a simpler form involving a joint uniform cumulative distribution function in a transformed random variable and marginal probability density functions. The idea is to use the probability integral transform

$$Y = \int_{-\infty}^{x} f_U(u)du = F_X(x),$$

for some random variable X to define a uniformly distributed random variable Y. While the original theorem makes no mention of such a condition, it is often preferable to use continuous marginal cumulative distribution functions $F_{X_n}(X_n) = U_n$, so that each $U_n \sim U(0,1)$. Using Y as dummy variable, one may write

$$F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[F_X(X) \leq y] = \mathbb{P}[X \leq F_X^{-1}(y)] = F_X(F_X^{-1}(y)) = y.$$

The dependence between two random variables, say X and Y, is completely described by the joint distribution function $F_{X,Y}(x,y)$. The major motivation of separating $F_{X,Y}(x,y))$ in two parts: the one which describes the dependence structure, and the other one which describes the marginal behavior, leads to the concept of copula. To every bivariate distribution function $F_{X,Y}(x,y))$, with continuous marginals $F_X(x)$ and $F_Y(y)$, corresponds to a unique function

$C : [0,1] \times [0,1] \to [0,1]$, such that

$$F_{X,Y}(x,y)) = C\left(F_X(x), F_Y(y)\right),$$

for $(x,y) \in (-\infty,\infty) \times (-\infty,\infty)$. Extension of this definition to a higher dimension (say, in d - dimension, where $d \geq 3$) can certainly be envisioned. This could be summarized as follows.

A d-dimensional copula is a function $C : [0,1]^d \to [0,1]$ that satisfies:

(i) $C(u_1, u_2, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_d) = 0$ for all $1 \leq i \leq d$, and $0 \leq u_k \leq 1$ for $k = 1, \ldots, d$ with $k \neq i$.

(ii) $C(1, \ldots, 1, u, 1, \ldots, 1) = u$ for all $0 \leq u \leq 1$, in each of the d arguments. That is, the copula equals u when one argument is u and all others are 1.

(iii) For any vectors $\mathbf{s} = (s_1, \ldots, s_d)$ and $\mathbf{w} = (w_1, \ldots, w_d)$ such that $s_i \leq w_i$ for all $i = 1, \ldots, d$, the following inequality holds:

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots+i_d} C(u_1^{(i_1)}, \ldots, u_d^{(i_d)}) \geq 0,$$

where $u_j^{(1)} = s_j$ and $u_j^{(2)} = w_j$ for each $j = 1, \ldots, d$. That is, the copula is non-decreasing in d dimensions.

The copula is then the joint cumulative distribution function $C(\vec{u})$, where $\vec{u} = (u_1, u_2, \ldots, u_d)$ of the transformed random variables u and contains all dependency information of the initial random variables x_d. According to Sklar's (1959) theorem,

$$C(\vec{u}) = C(F_{X_1}(x_1), ..., F_{X_d}(x_d)) = \mathbb{P}[U_1 \leq u_1, ..., U_d \leq u_d],$$

such as elliptical, Archimedean, and Pickands.

Sklar's formal introduction of copulas provided a powerful tool to model and analyze dependency structures between random variables, regardless of marginal distributions. His theorem is not

limited to the mathematical world, as fields such as finance, hydrology, and engineering all need an understanding of the joint behavior of random variables through certain copulas, see, Nadarajah (2017) for pertinent details in this context.

The Gaussian copula family of models was first introduced to the financial field by Oldrich Vasicek, see, for pertinent details, see, Trivedi (2007) and the references cited therein. He first introduced the copula model at a corporate loan firm to help reduce concentration of risk in specific geographic regions and industries. From there, the use of copulas in finance increased heavily in 2000, when David X. Li applied them to model default correlations in credit risk. This work skyrocketed the popularity of the Gaussian copula model to price collateralized debt obligations (CDOs) and assess the risk of credit portfolios. However, the Gaussian copula, like all other types of copulas, has limitations that users must be mindful of, particularly in capturing tail dependencies. The overuse of the Gaussian copula has been linked to the 2007-2008 financial crisis because of failure to predict extreme joint movements, which led to significant financial loss. This shows the need to determine the correctness of assumptions underlying the use of specific copulas. Next, we discuss another type of copula which is useful in financial risk modeling, namely, the Gumbel copula.

The Gumbel Copula, named after Emil Julius Gumbel, a German mathematician known for his contributions to the Extreme Value Theorem, is a part of the Archimedean copula tree that specializes in deciphering the dependence structures between random variables, mostly focusing on upper tail dependence, scenarios where extreme high values in one variable are associated with extreme high values in another, see, Tinungki (2023) and the references cited therein for an extensive discussion on this matter. The Gumbel copula has been applied to various fields, such as risk management in the financial field, actuarial science, modeling insurance risks, and hydrology, modeling joint distributions in events such as extreme rainfall and river discharge, helping to assess flood risks among others. The Gumbel Copula also has its limitations. While it is great for modeling the upper tail dependence, it lacks a similar strength in magnitude for modeling the lower tail dependence, which is the tendency of variables to jointly exhibit low values. In addition, there is another family (alias collection of copulas) which are quite useful in actuarial data application, popularly known as Archimedean copula.

The term "Archimedean Copula" was first introduced in statistical literature by Christian Genest and Jock MacKay in their work, see, Genest (1986). The most common Archimedean are the Gumbel Copula, as previously discussed, the Clayton Copula, the Frank Copula, the Ali-Mikhail-Haq (AMH) Copula, the Joe Copula, and the Nelsen Copula etc. All of these are designed to model dependencies between random variables, with each copula model specializing in different fields with different tail dependency modeling.

Although not that much of use (as compared to a Gaussian copula), the t-Copula, also known as the Student's t-Copula, is another copula used to model dependence structures between multiple random variables. The term t-Copula first gained notoriety in 2005, through the work of Stefano Demarta and Alexander J. McNeil. In their work, they analyzed the t-Copula's properties and applications, which contributed significantly to its adoption in statistical modeling. The t-Copula is heavily applied in the risk management and financial fields, especially in portfolio management. The t-copula can also be divided into two special cases, namely, the Skewed t-Copula and the Grouped t-Copula. The Skewed t-Copula extension introduces asymmetry into the dependence structure, which allows for different behaviors in the upper and lower tail structures, which is extremely useful in the financial field. Meanwhile, the Grouped t-Copula model allows for varying degrees of freedom parameters for different groups of variables, which allows us to get a more nuanced modeling of dependencies in heterogeneous data sets, see, Venter(2002). The t-copula, and their extensions have drawbacks, especially in terms of computation. In the next, we discuss the motivation to carry out this project.

This paper leverages the copula models available in the VineCopula package in R to explore about the dependency structure among concomitant variables present in an various types of insurance data. As the misuse of copula models was partly to blame for the financial collapse of 2008, we concern ourselves with tail dependence, or dependence of extreme events. An example of the importance of tail dependence in insurance data is as follows: Suppose that there is a strong upper tail dependence of average charges to a health insurer and the number of stays at the hospital. This would imply that the overall cost is much higher. If the insurer finds that many of its policyholders have frequent hospital visits, this could pose a significant risk to the insurer. The copula fit to the data, as in this example, gives important insights into dependency and tail dependence so that the appropriate decision(s) can be made.

Our work intends to fit bivariate copulas to understand these tail risks, as well as on the overall dependence. The remainder of this paper is organized as follows. In Section 2, we describe various dependence measures and the associated copula version of it. Section 3 briefly outlines various datasets from insurance domains and we provide adequate rationale on the selection of these data sets as well as variable selection. In Section 4, we provide the details of the copula fitting to each of these datasets. Finally, some concluding remarks are made in Section 5.

2 Dependence Measures

In the study of bivariate dependence, there are multiple measures specific to each types of different scenarios that an experimenter/researcher can envision. In this paper, we focus on three distinct types of popular dependence measures, each of which can be obtained via a copula which are:(a) Pearson's correlation coefficient; (b) Kendall's Tau, and (c) Blomqvist's β.

- Pearson's correlation coefficient is given by

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

  where $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$, and where $|\rho| \leq 1$.
  In terms of copula, Spearman's ρ_S can be written as

$$\rho_S = 12 \iint_{[0,1]^2} C(u,v)\, dC(u,v) - 3 = 12 \iint_{[0,1]^2} [C(u,v) - uv]\, du\, dv.$$

  However, Pearson's correlation is highly sensitive to outliers as it involves central moments, as a single extreme value can distort the coefficient and produce a misleading output. Pearson's correlation assumes the relationship between two variables is linear, and does not capture nonlinear associations. Additionally, it assumes that the data is approximately Gaussian which may produce unreliable results if the data is skewed or otherwise non-Gaussian. In cases where assumptions are violated or the relationship is non-linear, different methods should be used, such as Kendall's τ. In the next, we discuss some useful details on Kendall's τ.
- Kendall's τ, developed by Maurice Kendall in 1938, is a non-parametric measure of the strength and direction of the relationship between two variables. Unlike Pearson's correlation, which is designed for linear relationships, Kendall's Tau is useful for ordinal (ranked) data or when the relationship between variables is nonlinear. Kendall's Tau compares the relative ordering of pairs and focuses primarily on counting the number of concordant and discordant pairs. A pair is said to be concordant if:

$$(X_1 - X_2)(Y_1 - Y_2) > 0,$$

where both variables move in the same direction (both increasing or both decreasing). A pair is said to be discordant if:

$$(X_1 - X_2)\,(Y_1 - Y_2) < 0,$$

where one variable increases while the other decreases, and (X_1, Y_1) and (X_2, Y_2) are two pairs of random variables from a joint distribution function. If a pair holds the same rank, it is considered tied.

There are two variations of Kendall's Tau: Tau-a and Tau-b. Tau-a is used when there are no tied ranks in the data, while Tau-b is the most commonly used version when ties are present. The formula for Kendall's Tau is given by:

$$\tau = \frac{C - D}{\sqrt{(C + D + T_1)(C + D + T_2)}},$$

where C is the number of concordant pairs, D is the number of discordant pairs, and T_1, T_2 are the number of tied pairs. Alternatively, τ can be written as:

$$\tau = \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) > 0] - \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

Let X and Y be continuous random variables with copula C. Then Kendall's tau is:

$$\tau = 4 \iint_{[0,1]^2} C(u, v)\, dC(u, v) - 1$$

The result ranges from -1 (perfect negative rank correlation) to $+1$ (perfect positive rank correlation), with 0 indicating no correlation. Kendall's Tau is widely used in the social sciences, medical research, and economics. However, it has limitations, such as a computational complexity of $O(n^2)$ and sensitivity to ties in data. Next, we consider the role of Blomqvist's β as a measure of dependence.

- Blomqvist β, introduced by Nils Blomqvist in 1950, is a non-parametric measure of statistical dependence. It measures the strength of association between two variables based on medians rather than considering the full distribution ranks. Unlike Pearson's correlation coefficient, which measures linear relationships, or Kendall's Tau, which assesses monotonic relationships, Blomqvist's β focuses on how observations cluster around their median values. Blomqvist's β is particularly useful when dealing with ordinal data, non-Gaussian distributions, and datasets with outliers. It is defined as:

$$\beta = 4\mathbb{P}\left[X > M_X, Y > M_Y\right] - 1,$$

where M_X and M_Y are the medians of X and Y, respectively, and $P\left(X > M_X, Y > M_Y\right)$ is the probability that both variables fall above their medians. The copula based expression of Blomqvist β is defined as

$$\beta = 4C\left(\tfrac{1}{2}, \tfrac{1}{2}\right) - 1.$$

Blomqvist's β ranges from -1 (strong negative association) to $+1$ (strong positive association), with 0 indicating no association. While less commonly used, Blomqvist's β is applied in finance for skewed financial returns, Denuit (2005); medical research for disease rates, Nooraee (2014), and social sciences for survey analysis, Agresti (2010). However, it has weaknesses, including loss of information due to median-based classification, sensitivity to ties in data, and poor performance for weak dependencies. All considered, it is necessary to use either Blomqvist's β or other correlation measures as demanded by the application.

Looking strictly at the financial field, copulas enable modeling of dependencies between different assets, capturing tail dependencies that traditional correlation measures might miss. This allows for a more accurate estimation of risk measures such as Value-at-Risk (VaR) and Conditional Value-at-Risk (CVar), leading to better-informed investment decisions. The flexibility of copulas allows for the modeling of complex, non-linear dependencies, and tail risks, which are critical in stress testing and scenario analysis. By selecting appropriate copula functions, risk managers can simulate various market conditions and assess potential impacts on portfolios. Some of the most commonly used copulas are the Gumbel copula for extreme distributions, the Gaussian copula for linear correlation, and the Archimedean copula and t-copula for dependence in tails.

Our goal is to model and examine the strength and direction the dependency structure(s) of concomitant random variables that are selected from the CAS datasets for an efficient way of creating guidelines for a better decisions pertaining to insurance pricing or so. This paper will use Pearson's correlation, Kendall's Tau, and Blomqvist's β to analyze the relationship between bivariate distributions in terms of the best fitted copula for each of those illustrative data sets and using copula based definition of those dependency measures. Choosing the most accurate statistical model is essential for extracting relevant information from the data. It is to be noted that the Pearson's correlation measure is best for identifying the strength of linear relationships, Kendall's τ is preferred for monotonic but nonlinear relationships, and the Blomqvist's β is suitable when median-based dependencies matter. Next, we provide some useful details on the real data sets selected from the CAS data sets in `R`.

3 Real Data Application

Swedish motor insurance data is provided by the CAS datasets repository on Github for public use (link: https://cas.uqam.ca/pub/web/CASdatasets-manual.pdf), especially with the Computational Actuarial Sciences textbook. This insurance data was collected in 1977 by the Swedish Committee on the Analysis of Risk premiums. The data of concern for this repository are *claims*, the number of claims by each policyholder; *insured*, the number of years of the policy; and *payment*, the sum of payments made by the policyholder. We will use bivariate copulas to model the bivariate dependency structures of both *claims* and *insured* to *payment*. We chose to model *claims* against *payment* to describe tail risks associated with the number claims made to payment, as this is a concern for financial risk purposes. We chose to model *insured* to *payment* to determine how payments toward the policy decide how long the policy is held, or vice versa.

Next, term life insurance data is also provided by the CAS datasets repository. This is data from the United States in a survey from survey of consumer finances. It is a nationally representative random sample of 500 households. The data of concern for this repository are *income*, the income of the family, and *face*, the amount of the payout in the event of death. We chose to model *income* to *face* to see how the choices of the value of the policy are decided by the income of the family.

Next, we consider Medicare Hospital Costs which is a dataset from the Health Care Financing Administration, Bureau of Data Management and Strategy. It is mainly for use with Regression Modeling with Actuarial and Financial Applications by Frees (2009). We look specifically at the Medical Expenditure Panel Survey (MEPS), which was a nationally representative survey conducted by the U.S. Agency of Health Research and Quality. The data of concern in this dataset are *income*, the median income of families, and *borrowed*, the amount borrowed on the life insurance policy. We

chose to model *income* to *borrowed* to determine the relationship between the income of the family and how much it borrows from its life insurance policy.

3.1 Variable Selection

To model the above mentioned example data sets via a bivariate copula, we first revisit the Swedish motor insurance data set and a US life insurance data set, the details on this data is given earlier in Section 2. This data set has several concomitant variables. However, we describe below (in the form of several models) variable selection from this data set and study the associated dependence structure via a best fitted copula.

1. **Model 1:** We consider Insured and Payment as two concomitant variables and we want to explore the strength and direction of the dependency between them. By implementing a copula based approach (via the CDA vine copula package in R), the best fitted bivariate copula in this case is found to be the following: BB6 copula with par1= 1.59, par2 = 2.81, τ = 0.73. Some useful structural details of this copula is provided in Section 4.
2. **Model 2:** Involving components Claims and Payment. Here, the best fitted bivariate copula is the bivariate t copula with par1= 0.98, par2 = 2, τ = 0.87. Some useful structural details of this copula is provided in Section 4.

After analyzing the individual scatter plots, we decided upon further analysis by comparing the contour plots, which can be found below in Figure 3. The left contour plot of Figure 3 shows very tight contour lines, once again suggesting a high dependence relationship between claims and payment. The lack of curvature in the corners of the contour graph suggests low-tail asymmetry. For the right contour plot, we see very tight contour lines in the top right and more loose in the bottom left, a clear sign of upper tail dependence. The sharper corner in the top right of the graph could suggest that the more extreme values of insured have a more closely associated extreme value of payment values, whereas lower values of insured may have more variability for the payment values. When modeling via the copulas, we found that the Insured and Payment variables had a Kendall's $\tau = 0.73$, which indicates a relatively strong correlation between the two variables. The best fitted copula is the BB6 copula, which is a member of the Archimedean family. The BB6 copula is particularly well suited for capturing strong tail dependencies, meaning that extreme values in one variable are likely to correspond to extreme values in the other. This makes it an appropriate choice for modeling insurance related datasets where large payments may be influenced by the number of insured individuals. The BB6 copula is parameterized by θ (par 1) and δ (par 2) which control the dependence structure. The θ parameter governs the overall dependence structure strength, while δ determines tail dependence, with larger values indicating stronger tail dependence. The CDF graph in figure 4 shows the cumulative probability over the unit square $[0, 1]^2$, based on the BB6 copula. The steep climb towards the corner suggests strong positive dependence, particularly in the upper tail. This kind of sharp slope usually indicates that more extreme high values of the variables tend to occur jointly more frequently than under independence, confirming our previous findings. As for the PDF graph, this shows the derivative of the CDF, which is a visualization of where the associated copula density is concentrated. There is a very sharp spike in the top right corner of (1,1), and almost flat everywhere else, confirming a strong upper tail dependence. This graph also captures the extreme co-movements, once again confirming that when one variable is high, the other variable will tend to be high too. For illustrative purposes, we provide the scatterplot between Claims and Payment in Figure 1 in the Appendix. From Figure 1, it appears that there exists a strong positive linear relationship. In Figure 4, we provide the pdf and the cdf for the BB6 copula with the estimated parameter values.

For the second model, the best fitted bivariate copula in this case is a Student's t-Copula, which is a member from the Elliptical family, and is particularly useful for capturing both linear and tail dependence. Using a Student's t-Copula allows for flexible correlation structures that better model financial and insurance datasets where extreme values can occur at the same time. The two key parameters for the Student's t-Copula are ρ (par 1) and ν (par 2). The ρ parameter represents the correlation between two variables, while ν, the degrees of freedom, controls the heaviness of tails, with smaller ν values indicating stronger tail dependence. The Student's t-Copula makes for a suitable choice for modeling extreme claims and payment behavior, as it allows for capturing risk concentration in the tails. In terms of the Claims and Payment copula, we found that the Kendall's τ value to be 0.87, which indicates a significantly strong correlation between the two variables. Figure 1 suggests a very strong, positive correlation, which confirms our high Kendall's τ of 0.87. As Claims increases, Payment tends to increase, with low variation. For lower claims, payment is more spread out, implying more noise or variability with how smaller claims are handled. For illustrative purposes, we provide the scatterplot between the two variables Insured and payment. Figure 2 exhibits a strong positive correlation. As Insured values increase, so do the payment values. These points are clustered very tightly in the top right corner, which is a sign of upper tail dependence, meaning high insured values are closely associated with high payment values. There is some higher vertical spread within the 0.25 to 0.75 range, suggesting some moderate variability within the dependence of the two variables. In Figure 7, we provide the pdf and the cdf of a bivariate t-copula with the estimated parameter values.

The tables below detail the summary of the goodness of fit, estimates of the model parameters etc. of each model. Note that all of these computations were performed in R using the `CDAvine package`.

Table 1: Dependence measures between variables for Swedish Motor data.

Swedish Motor/Model	X_1	X_2	Kendall's τ	Spearman's ρ
Model1	Claims	Payments	0.87	0.962
Model2	Insured	Payments	0.73	0.903

Table 2: Model diagnostics and goodness of fit statistics for the best fitted copula for the Swedish Motor data.

Swedish Motor Model	Best Fitted Copula	Parameter Estimates	AIC	BIC	Log Likelihood
Model1	T-Copula	$(0.98, 2)$	-7129.3	-7117.92	3566.65
Model2	BB6	$(1.59, 2.81)$	-4095.96	-4084.58	2049.98

3.2 Data Set 2

This data set, the US Term Life Insurance data set, was a survey with 500 household participants carried out in 2004 by the Survey of Consumer Finances (SCF) group. This included characteristics such as gender, age, marital status, education, etc. This data set is also publicly available on the

Wisconsin School of Business FreesBook-RMAFA website. Here, we consider two different models based on two sets of concomitant variables selected for this purpose.

1. **Model 3:** Involving components Income and Face. Here, the best fitted copula is given by Tawn type 1 with par1 = 1.84, par2 = 0.49, tau = 0.28.
2. **Model 4:** Involving components Income and Borrow CV Life Policies. Here, the best fitted bivariate copula is the bivariate Frank copula with par1= 1.59, par2 = 0.17, tau = 0.17.

As seen above in the Term Life Insurance copula between the variables Income and Face, we see the Kendall's Tau value is 0.28, which indicates a positive but very weak correlation, which at first glance indicates that the variables chosen have almost no impact on the life insurance policy these households would use/already have purchased. The Tawn type 1 copula, used to model this relationship, is an extension of the Gumbel copula that allows for asymmetric tail dependence. Unlike symmetric copulas such as the Gaussian or Clayton copulas, the Tawn type 1 copula introduces an additional asymmetry parameter, which enables it to better capture imbalances in dependency structure. This copula is characterized by θ parameter, which controls the overall strength of dependence, and is equal to 1.84. The second parameter, δ, which introduces asymmetry (one tail may be stronger than the other), is equal to 0.49. Given that the Kendall's Tau is very low, the copula structure confirms that there is minimal dependency between Income and Face in influencing life insurance policy choices. However, the choice of a Tawn type 1 copula suggests that there may still be some asymmetric tail dependence, meaning that individuals at extreme values of Income or Face might exhibit slightly different dependency patterns than those in the middle of the distribution.

Upon careful observation of the scatter plot and contour plots of Income and Face, which can be found in Figures 5 and 6 respectively, the scatterplot is much more diffuse than the previous ones. There is a very loose upward trend, which means that as the income values increase, the face values tend to increase, but it is not very tight and there is a large vertical spread across most income levels. The horizontal band of points at the bottom could suggest either a minimum face value for each policy given regardless of income, or a mass of ties in the data points when transformed into pseudo-observations. The contour plot also supports our previous summation of the scatter, as Figure 6 shows the contours are fairly elliptical and centered, which is a hallmark of weak symmetrical dependence. There is no strong skew or curvature, suggesting that there is no upper or lower tail dependence. The very slight elongation along the 45 degree line suggests a small positive correlation, meaning that increases in income tend to increase with face value, but very weakly. There are no strong outliers or asymmetry to speak of.

The tables below detail the results of each model.

Table 3: Dependence measures between variables for Term Life Insurance data

Term Life/Model	X_1	X_2	Kendall's τ	Spearman's ρ
Model1	Income	Face	0.28	0.365
Model2	Income	Borrow CV	0.17	0.210

3.3 Data Set 3

The third dataset we examined was Medicare Hospital Costs, where data was obtained from the Health Care Financing Administration, Bureau of Data Management and Strategy. This data set includes impatient hospital charges covered by the Medicare program for the years 1990-1995. This

Table 4: Model diagnostics and goodness of fit statistics for the best fitted copula for Term Life Insurance data

Term Life Model	Best Fitted Copula	Parameter Estimates	AIC	BIC	Log Likelihood
Model1	Tawn Type 1	$(1.84, 0.49)$	-112.75	-104.32	58.37
Model2	Frank	$(1.59, 0.17)$	-20.42	-16.21	11.21

included variables such as state name, total hospital charges, number of hospital stays, number of discharged, etc. This data set is available on the FreesBook-RMAFA website of the Wisconsin School of Business. This dataset involves components Charges and Number of Stays. On using the CDAVinepackage in R, the best fitted bivariate copula happens to be a Survival Gumbel with the following parameter choices: (par1 = 6.12, par2 = 0.84, τ= 0.84)

As indicated above in the Medicare data, the bivariate copula between the two seemingly concomitant variables, name ly the Charges and the Number of stays, we see the Kendall's τ= 0.84, which indicates a positive and strong correlation. This means that the total hospital charges covered by medicare has a strong, positive dependency with the number of hospital stays. The Survival Gumbel copula is a copula function derived from the Gumbel copula, which is known to capture a strong positive dependency in extreme values. Unlike the original Gumbel copula, which emphasizes the dependence on the lower tail, the survival version allows the copula to focus on the upper tail rather than the lower tail.

The tables below detail the results of each model.

Table 5: Dependence measures between variables for Medicare data.

Medicare Costs	X_1	X_2	Kendall's τ	Spearman's ρ
Medicare	Charges	Number of Stays	0.84	0.97

Table 6: Model diagnostics and goodness of fit statistics for the best fitted copula for Medicare data.

Medicare Costs	Best Fitted Copula	Parameter Estimates	AIC	BIC	Log Likelihood
Medicare	Survival Gumbel	$(6.12, 0.84)$	-895.61	-891.83	448.8

In the next, we discuss several useful structural properties related to each of the fitted bivariate copula beginning with the BB6 copula.

4 Structural properties of the fitted Copula

This section presents the analysis of certain structural properties of the copulas. We begin our discussion with the BB6 copula.

4.1 BB6 (Joe-Gumbel) Copula

The BB6 copula is an Archimedean copula, so that for generator $\phi(t)$,

$$C(u,v) = \phi^{-1}(\phi(u) + \phi(v)).$$

In this case, the generator and its inverse ϕ^{-1} are

$$\phi(t) =(-\log[1-(1-t)^{\theta}])^{\delta}, \theta > 0, \delta \geq 1;$$

$$\phi^{-1}(s) = 1-(1-\exp[-s^{1/\delta}])^{1/\theta}.$$

$$u \geq 0, v \leq 1, \theta \geq 1, \delta \geq 1.$$

Therefore, the expression of the BB6 copula is given by

$$C\left(u,v\right) = 1-\left\{1-\exp\left\{-\left[\left(-\log\left(1-\bar{u}^{\theta}\right)\right)^{\delta}+\left(-\log\left(1-\bar{v}^{\theta}\right)\right)^{\delta}\right]^{1/\delta}\right\}\right\}^{1/\theta}, \tag{1}$$

where $\bar{u} = 1-u$, and $\bar{v} = 1-v$.

- The BB6 copula is symmetric in the sense that for this copula $C\left(u,v\right) - C\left(v,u\right) = 0 \quad \forall (u,v) \in [0,1]^2$.
- Next, we determine the value of Blomqvist β for the BB6 copula based on the parameters of the Swedish Automobile insurance model with $\theta = 1.59$ and $\delta = 2.81$. Therefore, on using the formula for Blomqvist's β, the lower tail and upper tail dependence coefficients can be calculated using the same methodology that we used for the Frank copula. For the upper tail dependence coefficient, we obtain the following:

$$\begin{aligned}
\lambda_U &= \lim_{u\uparrow 1} \frac{1-2u+1-\left(1-\exp(-[2(-\log(1-\overline{u}^{\theta}))^{\delta}])\right)^{\frac{1}{\theta}}}{1-u} \\
&\overset{H}{=} \lim_{u\uparrow 1} 2-2^{\frac{1}{\delta}}(1-u)^{\theta-1}\exp\left[2^{\frac{1}{\delta}}\log(1-(1-u)^{\theta})\right](1-\exp\left[2^{\frac{1}{\delta}}\log(1-(1-u)^{\delta}\right])^{\frac{1}{\theta}-1} \\
&= 2-2^{\frac{1}{\theta\delta}}.
\end{aligned}$$

Similarly for the lower tail dependence coefficient:

$$\begin{aligned}
\lambda_L &= \lim_{u\downarrow 0} \frac{1-\left(1-\exp(-[2(-\log(1-\overline{u}^{\theta}))^{\delta}])\right)^{\frac{1}{\theta}}}{u} \\
&\overset{H}{=} \lim_{u\downarrow} 2^{\frac{1}{\delta}}(1-u)^{\theta-1}\exp\left[2^{\frac{1}{\delta}}\log(1-(1-u)^{\theta})\right](1-\exp\left[2^{\frac{1}{\delta}}\log(1-(1-u)^{\delta}\right])^{\frac{1}{\theta}-1} \\
&= 0,
\end{aligned}$$

which have been independently obtained in Ghosh et al. (2023).
Next, we consider the conditional copula of U given $V = v$ and vice versa.

– The conditional copula of U given $V = v$ will be

$$
\begin{aligned}
& C_1\left(u|V=v\right) \\
& = \frac{\partial C\left(u,v\right)}{\partial v} \\
& = \left\{1-(1-v)^{\theta}\right\}^{-1} \\
& \times \left[(1-v)^{\theta-1}\left(-\log\left(1-(1-v)^{\theta}\right)\right)^{\delta-1}\left(\left(-\log\left(1-(1-u)^{\theta}\right)\right)^{\delta}\right.\right. \\
& \left.\left.+\left(-\log\left(1-(1-v)^{\theta}\right)\right)^{\delta}\right)^{\frac{1}{\delta}-1}\right. \\
& \times \left(1-\exp\left(-\left(\left(-\log\left(1-(1-u)^{\theta}\right)\right)^{\delta}+\left(-\log\left(1-(1-v)^{\theta}\right)\right)^{\delta}\right)^{\frac{1}{\delta}}\right)\right)^{\frac{1}{\theta}-1} \\
& \left.\times \exp'\left(-\left(\left(-\log\left(1-(1-u)^{\theta}\right)\right)^{\delta}+\left(-\log\left(1-(1-v)^{\theta}\right)\right)^{\delta}\right)^{\frac{1}{\delta}}\right)\right]. \qquad (2)
\end{aligned}
$$

– Likewise, the The conditional copula of V given $U = u$ will be

$$
\begin{aligned}
& C_2\left(v|U=u\right) \\
& = \left[1-(1-u)^{\theta}\right]^{-1} \\
& \times (1-u)^{\theta-1}\left(-\log\left(1-(1-u)^{\theta}\right)\right)^{\delta-1}\left(\left(-\log\left(1-(1-u)^{\theta}\right)\right)^{\delta}\right. \\
& \left.+\left(-\log\left(1-(1-v)^{\theta}\right)\right)^{\delta}\right)^{\frac{1}{\delta}-1} \\
& \times \left(1-\exp\left(-\left(\left(-\log\left(1-(1-u)^{\theta}\right)\right)^{\delta}+\left(-\log\left(1-(1-v)^{\theta}\right)\right)^{\delta}\right)^{\frac{1}{\delta}}\right)\right)^{\frac{1}{\theta}-1} \\
& \times \exp'\left(-\left(\left(-\log\left(1-(1-u)^{\theta}\right)\right)^{\delta}+\left(-\log\left(1-(1-v)^{\theta}\right)\right)^{\delta}\right)^{\frac{1}{\delta}}\right). \qquad (3)
\end{aligned}
$$

Observe that, one may use the conditional copula of U given $V = v$, given in Eq. (2) and to simulate from the propose BB6 copula as given in Eq. (1) using the following steps:

– Simulate and v_i and u_i^* from a standard uniform distribution.
– If $v_i \leq 1$, then solve $C_1(u|v_i) = u_i^*$.
– Repeat the previous two steps, say, n times to obtain independence and identically distributed realizations (u_i, v_i), for $i = 1, 2, \cdots, n$ from the BB6 copula as given in Eq. (1).

A similar algorithm can be elaborated to simulate from the BB6 copula based on the conditional copula of V given $U = u$ as given in Eq. (3).

Next, we consider few other useful structural properties of the BB6 copula that has not been discussed as of yet to the best of the knowledge by the authors.

1. **Proposition 1.** The BB6 copula defined in Eq. (1) is decreasing with respect to its' dependence parameter θ, i.e., if $\theta_1 < \theta_2$ then
$$C_{\theta_2}(u,v) \leq C_{\theta_1}(u,v),$$
for all $(u,v) \in I^2 = [0,1] \times [0,1]$.
Proof.
Let us consider the partial derivative of Eq. (1) w.r.t. θ. We have
$$\frac{\partial C(u,v)}{\partial \theta} = \left\{ \frac{\frac{\delta(1-u)^\theta \log(1-u)\left(-\log\left(1-(1-u)^\theta\right)\right)^{\delta-1}}{1-(1-u)^\theta} + \frac{\delta(1-v)^\theta \log(1-v)\left(-\log\left(1-(1-v)^\theta\right)\right)^{\delta-1}}{1-(1-v)^\theta}}{\theta\left(\left(-\log\left(1-(1-u)^\theta\right)\right)^\delta + \left(-\log\left(1-(1-v)^\theta\right)\right)^\delta\right)} - \frac{\log\left(\left(-\log\left(1-(1-u)^\theta\right)\right)^\delta + \left(-\log\left(1-(1-v)^\theta\right)\right)^\delta\right)}{\theta^2} \right\} \times \left\{ \exp'\left(-\left(\left(-\log\left(1-(1-u)^\theta\right)\right)^\delta + \left(-\log\left(1-(1-v)^\theta\right)\right)^\delta\right)^{\frac{1}{\theta}}\right)\right\}. \quad (4)$$
Next, observe that for $\delta \geq 1$, and for $(u,v) \in [0,1]^2$,
- it is easy to see that $\left(-\log\left(1-(1-u)^\theta\right)\right)^\delta > 0$, and $\left(-\log\left(1-(1-v)^\theta\right)\right)^\delta > 0$.
- Also, $\log(1-u) < 0$, and $\log(1-v) < 0$.

Therefore, the numerator in the first two terms are negative, the second term is also negative. The third term involving $\exp'()$ is positive. Consequently, $\frac{\partial C(u,v)}{\partial \theta} < 0$, which completes the proof.
2. **Proposition 2.** The regression of U given $V = v$ is strictly decreasing in v.
Proof. The proof is left as an exercise to the reader.
3. **Proposition 3.** The BB6 copula as given in Eq. (1) is absolutely continuous. To establish the absolute continuity of the BB6 copula, we need to show that
$$\int_0^u \int_0^v \frac{\partial^2 C(s,t)}{\partial s \partial t} = C(u,v).$$
Proof. Simple and thus excluded.
It must be noted that these properties can also be derived for all the bivariate copulas described/utilized in this paper. However, for the sake of brevity, we have not considered them all.

4.2 Bivariate t Copula

The 2 dimensional unique t copula (see, Embrechts et al. (2001), McNeil, and Straumann (2001) or Fang & Fang (2002)) associated with a bivariate random vector $Y = (Y_1, Y_2)^T$, is given by
$$C_\delta^t(u,v) = \int_{-\infty}^{t_\delta^{-1}(u)} \int_{-\infty}^{t_\delta^{-1}(v)} \frac{\Gamma\left((\delta+2)/2\right)}{\Gamma(\delta/2)\sqrt{\left\{(\pi\delta)^2|\Sigma|\right\}}} \left[1 + \frac{y^T\Sigma^{-1}y}{\delta}\right]^{-\frac{\delta+2}{2}} dy_1 dy_2,$$

Table 7: Dependence Structures of the BB6 Copula based on Swedish Motor insurance data.

Generator Function	$\phi(t) = (-\log[1-(1-t)^{\theta}])^{\delta}$
Blomqvist β (General)	$4C(0.5, 0.5) - 1$
Blomqvist β (Swedish Auto)	0.7397
Upper Tail Dependence(General)	$2 - 2^{\frac{1}{\delta\theta}}$
Upper Tail (Swedish Auto)	0.8321
Lower Tail Dependence	0
Kendall's τ	0.73

where $t_{\delta}^{-1}(.)$ denotes the quantile function of a standard univariate $t_{\delta}(.)$ distribution. Furthermore, Σ is the correlation matrix given by

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

where ρ is the correlation coefficient between Y_1, and Y_2. The determinant of this matrix, denoted by $|\Sigma|$ is given by $|\Sigma| = 1 - \rho^2$. Next, one may verify the following regarding the dependence structure for a bivariate t copula

- This copula is also symmetric.
- Kendall's τ will be

  $$\tau = \frac{2}{\pi} \arcsin \rho,$$

  for the proof, see Fang & Fang (2002).
- Regarding the Spearman's correlation coefficient, there is no analytically tractable expression that are available. However, if ρ is the Spearman's correlation coefficient, and by denoting $\rho_t(\rho, v)$, and $\rho_N(\rho)$, as the Spearman's correlation coefficient for the bivariate t-copula and the bivariate Normal copula, one may show that:
  (i) when $\rho < 0$, $rho_t(\rho, v) > \rho_N(\rho)$; (ii) when $\rho > 0$, $rho_t(\rho, v) < \rho_N(\rho)$.
- The tail dependence coefficient (associated with a bivariate t copula as given earlier) λ is given by

  $$\lambda = 2t_{\delta+1}\left(-\frac{\sqrt{\{\delta+1\}}\sqrt{\{1-\rho\}}}{\sqrt{\{1+\rho\}}}\right),$$

  where $t_{\delta+1}$ is the univariate central student t distribution with $(\delta+1)$ degrees of freedom and ρ is the correlation coefficient. It is important to note that a student t-copula may exhibit both the positive tail-dependence although the "overall" association is negative $\rho < 0$. Furthermore, a student t-Copula with a large value of δ will tend to have a 0 tail-dependence even though the correlation is $.0$ The t-copula can capture the asymptotic dependence even when the variables are negatively (inversely) associated (see, Embrechts et al. 2001). In t-copula formula, as δ increases, the tail dependence weakens, and thus, the probability of occurrence of extreme values reduces

For illustrative purposes, we provide the following picture figure below (generated through the https://*copulatheque.shinyapps.io/copulas/* created by BenGraeler) shows a student t-copula with $\rho =$

0.995 and $\delta = 2$ which gives the value of Kendall's $\tau = 0.87$ and upper and lower tail-dependence of $\lambda = 0.87$.

The estimation of a student t-copula is quite difficult. Noticeably, the marginal tails (for bivariate and/or multivariate data distributions) of financial data are usually heavy tailed and hence this should be fitted by a t-distribution and not by a Gaussian distribution. In addition, the dependence in joint extremes of bivariate and/or multivariate financial data suggests a dependence structure allowing for tail dependence. Consequently, the use of t- copulas have become popular for modeling dependencies in financial data. Some recent applications have been: analysis of nonlinear and asymmetric dependence in the German equity market (Sun et al., 2008); estimation of large portfolio loss probabilities (Chan and Kroese, 2010); risk modeling for future cash flow (Pettere and Kollo, 2011). See also Dakovic and Czado (2011). Figure 7 represents the PDF and CDF of the bivariate t-copula with the estimated model parameters. For the CDF, there is a smooth increase from (0,0) to (1,1). The CDF has a little more curvature at the corners, which usually indicates tail dependence. The PDF confirms the tail dependence, as there is a spike both at (0,0) and (1,1) indicating both lower and upper tail dependence. It should be noted that due to the limitations of our R-Package, we could not directly graph our given parameters, our DF was equal to 2 in the t-copula we derived, and the package we are using must have DF greater than 2. This is as close as we could get it.

4.3 Tawn Type-1 copula

Here are some useful details on the Tawn Type-1 copula.

- The Tawn copula is a nonexchangeable extension of the Gumbel copula with three parameters (also known as the asymmetric logistic copula).
- Tawn copula's definition is based around so-called Pickands dependence functions, see Franc et al. (2011) for pertinent details. Eq. (4) in Franc et al. (2011) presents the way one can compute the density in the probability space using a Pickands function M:

$$C(u,v) = (u,v)^{\eta(w)},$$

  with $w = \frac{u}{uv}$.
- The Pickands dependence function derives Franc et al. (2011) for pertinent details. Equation (4) in Franc et al. (2011) presents the way one can compute the density in the probability space using a Pickands function M:

$$C(u,v) = (u,v)^{\eta(w)},$$

  with $w = \frac{u}{uv}$.
- The Tawn copula's Pickand function is

$$M(t) = (1-\psi_2)(1-t) + (1-\psi_1)t + \left[(\psi_1(1-t))^{\theta} + \psi_2^{\theta}\right]^{1/\theta}.$$

The CDF in Figure 8 (see, Appendix), appears to show one side rising steeper than the other, which hints at asymmetry, a mainstay feature of Tawn type-1 copulas. Specifically, Tawn type 1 copulas model upper tail asymmetry, which can be noticed in the CDF graph's upper right hand region spike. The PDF shows a spike in the upper right hand corner, and a much softer presence elsewhere, again indicating upper tail dependence. These graphs also do a good job depicting what the Tawn type 1 models, which is the co-movements of high extreme values only.

4.4 Frank copula

The Frank copula is the best-fitted copula for the US Term Life variables income and amount borrowed on the life insurance policy. The cdf in Figure 8 shows nice curvature but no tail bias, as it appears to be symmetric across the diagonal. This indicates a moderate to strong overall dependence between the studied variables, but no tail dependence. The PDF in Figure 8 shows a crater-like structure, where it peaks in the center, meaning it assigns higher density to the mid-ranges of the variables' dependence, but is lighter in the co-extremes in the lower left or upper-right corners.

4.5 Survival Gumbel Copula

First, it would be nice to have the basic preliminaries related to a bivariate survival copula. We begin by stating the Sklar's theorem for survival functions:

Let $S(t_1, t_2)$ be a bivariate distribution with marginal survivor functions defined by $S_1(t_1)$ and $S_2(t_2)$. Then, there exists a bivariate copula $\bar{C}$ such that for all $(t_1, t_2) \in \mathbb{R}^2$

$$S(t_1, t_2) = C(S_1(t_1), S_2(t_2)).$$

The survival copula C couples the joint survival function to its' univariate marginals in a manner completely analogous to the way a copula connects the joint distribution function to its marginals. There exists a link between the survival copula C and the copula C. In the bivariate case, it is

$$\bar{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2).$$

It should be pointed out that the survival copula is also a copula, i.e. $C(u, v)$ is also a proper distribution function on $[0, 1] \times [0, 1]$. It is well known that many dependence properties of a bivariate distribution are copula properties, and therefore, can be obtained by studying the corresponding copula. These properties, however, do not depend on the marginals.

Table 8: Dependence Structures of the Frank Copula.

Generator Function	$\phi(t) = -\log_\alpha(\frac{\alpha^{-t}-1}{\alpha-1})$
Blomqvist β (General)	$\beta = 4\log_\alpha[1 + \frac{(\alpha^{0.5}-1)^2}{\alpha-1}] - 1$
Blomqvist β (US Term Life)	-0.264
Upper Tail Dependence (General)	0
Lower Tail Dependence (General)	0
Kendall's τ (US Term Life)	0.17

Next, we discuss some useful structural properties of the bivariate Frank copula.

- The Frank copula is an important Archimedean copula. Archimedean copulas are commutative, meaning that the order of the variables does not change the nature of the copula. The Frank copula is asymptotically independent in both directions. This means that it will have upper and lower tail dependence 0 in all cases.
- Its generator function is

$$-\log_\alpha(\frac{\alpha^{-t} - 1}{\alpha - 1}), \text{ for } \alpha > 0, \alpha \neq 1.$$

- The copula is defined as

$$C(u,v) = \log_\alpha[1 + \frac{(\alpha^u - 1)(\alpha^v - 1)}{\alpha - 1}], \text{ for } \alpha > 0.$$

- Its Kendall's τ is

$$\tau = 4[\frac{3\log(\alpha)(-2\log(1-\alpha) + \log(\alpha) + 2) - 6Li_2(\alpha) + \pi^2}{6\log^2(\alpha)}] - 1,$$

- The conditional copula of U given $V = v$ will be

$$C_1\left(u|V=v\right) = \frac{\left(\alpha^u - 1\right)\alpha^v}{(\alpha-1)\left(\frac{(\alpha^u-1)(\alpha^v-1)}{\alpha-1} + 1\right)} \tag{5}$$

Likewise, the conditional copula of V given $U = u$ will be

$$C_2\left(v|U=u\right) = \frac{\alpha^u\left(\alpha^v - 1\right)}{(\alpha-1)\left(\frac{(\alpha^u-1)(\alpha^v-1)}{\alpha-1} + 1\right)}. \tag{6}$$

- The conditional mean and variance of V given $U = u$ will be

$$\begin{aligned}
&E\left[V|U=u\right] \\
&= \left(2\log^3(\alpha)\right)^{-1} \\
&\times \Big[-2\mathrm{Li}_3\left(\frac{\alpha^v - 1}{\alpha^v - \alpha}\right) + 2\mathrm{Li}_3\left(\frac{\alpha\left(\alpha^v - 1\right)}{\alpha^v - \alpha}\right) - 2\log(\alpha)\mathrm{Li}_2\left(\frac{\alpha\left(\alpha^v - 1\right)}{\alpha^v - \alpha}\right) + \log^2(\alpha) \\
&\times \left(\log\left(\alpha^v\right) - \log\left(-\frac{(\alpha-1)\alpha^v}{\alpha^v - \alpha}\right)\right)\Big].
\end{aligned} \tag{7}$$

$$\begin{aligned}
&Var\left[V|U=u\right] \\
&= \left(3\log^4(\alpha)\right)^{-1}\Big[6\mathrm{Li}_4\left(\frac{\alpha^v - 1}{\alpha^v - \alpha}\right) - 6\mathrm{Li}_4\left(\frac{\alpha\left(\alpha^v - 1\right)}{\alpha^v - \alpha}\right) - 3\log^2(\alpha)\mathrm{Li}_2\left(\frac{\alpha\left(\alpha^v - 1\right)}{\alpha^v - \alpha}\right) \\
&+ 6\log(\alpha)\mathrm{Li}_3\left(\frac{\alpha\left(\alpha^v - 1\right)}{\alpha^v - \alpha}\right) + \log^3(\alpha)\left(\log\left(\alpha^v\right) - \log\left(-\frac{(\alpha-1)\alpha^v}{\alpha^v - \alpha}\right)\right)\Big] \\
&- \left(4\log^6(\alpha)\right)^{-1}\Big[\left(-2\mathrm{Li}_3\left(\frac{\alpha^v - 1}{\alpha^v - \alpha}\right) + 2\mathrm{Li}_3\left(\frac{\alpha\left(\alpha^v - 1\right)}{\alpha^v - \alpha}\right) - 2\log(\alpha)\mathrm{Li}_2\left(\frac{\alpha\left(\alpha^v - 1\right)}{\alpha^v - \alpha}\right)\right. \\
&\left. + \log^2(\alpha)\left(\log\left(\alpha^v\right) - \log\left(-\frac{(\alpha-1)\alpha^v}{\alpha^v - \alpha}\right)\right)^2\Big].
\end{aligned} \tag{8}$$

Next, we define the following two propositions for the bivariate Frank copula.

Proposition 1. The bivariate Frank copula is subharmonic for $0 < \alpha < 1$.

Proof. Let us consider the following:

$$\begin{aligned}\nabla^2 C_\alpha(u,v) &= \frac{\partial^2 C_\alpha(u,v)}{\partial u^2} + \frac{\partial^2 C_\alpha(u,v)}{\partial v^2} \\ &= -\frac{\log(\alpha)(\alpha^u-1)(\alpha^u-\alpha)\alpha^v}{(\alpha-\alpha^u+\alpha^{u+v}-\alpha^v)^2} - \frac{\log(\alpha)\alpha^u(\alpha^v-1)(\alpha^v-\alpha)}{(\alpha-\alpha^u+\alpha^{u+v}-\alpha^v)^2} \\ &= -\frac{\log(\alpha)\left(\alpha^{u+1}-2\alpha^{u+v}-2\alpha^{u+v+1}+\alpha^{2u+v}+\alpha^{u+2v}+\alpha^{v+1}\right)}{(\alpha-\alpha^u+\alpha^{u+v}-\alpha^v)^2}. \end{aligned} \tag{9}$$

From Eq. (9), it appears that for all choices of $(u,v) \in [0,1]\times[0,1]$ and for $0<\alpha<1$, $\nabla^2 C_\alpha(u,v) \geq 0$. This completes the proof.

Proposition 2. The bivariate Frank copula is absolutely continuous.

Proof. Simple and thus excluded.

Proposition 3. The bivariate Frank copula is symmetric.

Proof. It is easy to observe that for the Frank copula, $C(u,v) = C(v,u)$ which implies the result.

Figure 9 represents the pdf and the cdf of a bivariate Frank copula with the estimated parameter value for the Income and BorrowCVLife poi. Figure 9 exhibits nice curvature but no tail bias, as it appears to be symmetric across the diagonal. This indicates a moderate overall dependence between the studied variables, but no tail dependence. The PDF in Figure 9 shows a crater-like structure, where it peaks in the center, meaning it assigns higher density to the mid-ranges of the variables' dependence, but is lighter in the co-extremes in the lower left or upper-right corners. Figure 10, however, highlights the importance of running multiple tests. As we can see, the ranked numerical data makes the scatterplot just be horizontal lines, so more analysis is needed other than the scatter plot, as we can draw no real conclusions from it. Figure 11's contour plot, shows roughly elliptical contours that are centered around the origin, which is a hallmark of weak/moderate dependence. The inner contours represent the highest area of density, which aligns with our scatter plot, Figure 10. The contours are not skewed, which confirms the symmetry in our data. Since it is also rounded and there are no corners formed, this confirms joint co-extremes are not likely, and more moderate values are going to move jointly.

4.6 Survival Gumbel Copula

The best-fitted copula of the Medicare data was the Survival Gumbel copula. The Survival Gumbel copula is the Gumbel copula under the survival transformation. Whereas the Gumbel copula is useful for strong upper tail dependence, this transformation rotates the copula so that lower tail dependence is strong instead. Looking below at the scatter, contour, PDF and CDF plots, which can be found in Figures 12, 13 and 14, respectively. Figure 12's scatterplot has a nearly perfect diagonal pattern, suggesting a very strong and monotonic dependence relationship. The relationship is also very symmetric, and shows very strong co-movement: as CONV_CHG increases, so does TOT_D. In Figure 13, the contours are thin and very narrow along the 45 degree line, which is textbook near perfect dependence. The density is very tightly concentrated along the diagonal, indicating very high correlation. There is no sign of tail dependence and the contours are symmetrically distributed along the diagonal, ruling out asymmetry. In Figure 14, the surface of the CDF of Figure 10 shows a sharp increase towards the lower left corner (0,0). That sharp slope indicates concentration in mass in the lower tail, meaning that both variables tend to be small together. Note that the regular Gumbel CDF tends to show more mass concentration in the upper right, meaning that both variables tend to

be larger together. The PDF also reinforces this assertion, as there is a steep spike in the bottom left corner, showing that the co-extremes are more closely associated in the lower tail.

[htp]

Table 9: Dependence Structures of the Survival Gumbel Copula for the Medicare Cost dataset.

Generator Function	$\phi(t) = (-\log t)^{\theta}$
Blomqvist β (General)	$\$\beta = 4C(0.5, 0.5) - 1$
Blomqvist β (Medicare)	0.79
Upper Tail Dependence (General)	0
Lower Tail Dependence (General)	$2 - 2^{1/\theta}$
Lower Tail Dependence (Medicare)	0.88
Kendall's τ (Medicare)	0.84

- The usual Gumbel copula has the generator function

$$\psi(t) = (-\log t)^{\theta}, \theta \geq 1.$$

- The Gumbel copula is then

$$C_{\text{Gumbel}}(u, v) = \exp[-((-\log(u))^{\theta} + (-\log(v))^{\theta})^{1/\theta}].$$

- The Survival Gumbel copula is defined as

$$C(u, v) = u + v - 1 + \exp[-((-\log(1-u))^{\theta} + (-\log(1-v))^{\theta})^{1/\theta}]$$

- Its Kendall's τ is then

$$\tau = 1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)}dt = 1 - \frac{1}{\theta}.$$

- The lower tail dependence of the Survival Gumbel copula is equivalent to the upper tail dependence of the Gumbel copula:

$$\lambda_L = \lambda_U^{\text{Gumbel}} = \lim_{u\uparrow 1} \frac{1 - 2u + C(u, u)}{1 - u}$$

$$= \lim_{u\uparrow 1} \frac{1 - 2u + \exp(-2(-\log u)^{\theta})}{1 - u} = 2 - 2^{1/\theta},$$

on applying L'Hôpital's rule.

5 Conclusion

The best fitted bivariate copulas in all the cases that we have studied, the Kendall's τ ranges from 0.17 to 0.87; and the Spearman's ρ ranges from from 0.21 to 0.97. The log-likelihoods, or model fitness, range from 11.21 to 3566.65. One of the possible reason(s) for the varying dependence and fitness is the use of ranked data. Copulas are valid for these types of data but tell a less interesting story when few ranks are available. In the code, we used the pseudo-observation function to approximate

the probability integral transform discussed in the introduction. This is most powerful when there are many ranks to analyze: when ranks are dense enough to be approximately continuous. For data with few ranks, copulas have less fit and seem to have less dependence.

In the Swedish Motor Insurance data, we see strong fit and dependence, particularly for *Claims* to *Payments*. *Insured* to *Payments* has high dependence and log-likelihood. This suggests copula modeling has value for this data. Inspecting the data, we have numerical values for *Claims*, *Payments*, and *Insured*. This gives significant evidence for the algorithm to select the best-fitted copula. The resulting copulas then showed high dependence.

The Medicare Cost provided a strong dependence from *Charges* to *Number of Stays*. This is an example of strong dependence that has little dependence at one tail. Specifically, the Survival Gumbel copula was the best-fitted. This copula has no dependence in the upper tail. This indicates that high charges to Medicare and high number of hospital stays do not have a particular dependence besides random chance.

In the US Term Life data, we see weaker dependence and fit, particularly for *Income* to *Borrow CV Life*. Upon inspection, *Income* is numerical, although clearly rounded compared to what you would expect for such data. *Face* is either rounded or companies require round values for the face value of the life insurance. *Borrow CV Life* is ranked from 0-5 by the amount borrowed from the policy. These create issues of fitness. This illustrates the importance of detailed data for copula modeling.

If the US Term Life data had more granularity, we may recover a better fit copula than the Tawn Type-1 and Frank copulas. This asserts that insights from the copulas generated require more care or more data.

We provide useful results of copula modeling for three major insurance institutions. Swedish Motor Insurance data exhibits strong dependence in the variables chosen. Medicare data does as well, with independence of large charges and numbers of stays. US Term Life suffers from constraints of granularity for copula modeling and could be studied further with more granular data. For the Swedish Motor and Medicare data, these considerations could be useful for pricing and risk mitigation.

References

[1] Tawn, J. A. (1990). Modelling multivariate extreme value distributions. Biometrika, 77(2), 245-253.

[2] A. Sklar (1959). Fonctions de répartition à N dimensions et leurs marges. Publications de l'Institut Statistique de l'Université de Paris, 8, 229–231.

[3] Blomqvist, N. (1950). On a measure of dependence between two random variables. The Annals of Mathematical Statistics, 593-600.

[4] Durante, F., & Sempi, C. (2016). Principles of copula theory (Vol. 474). CRC Press.

[5] Joe, H. (1997). Multivariate models and multivariate dependence concepts. CRC press.

[6] Joe, H. (2014). Dependence modeling with copulas. CRC press.

[7] Denuit, M., Dhaene, J., Goovaerts, M., & Kaas, R. (2005). Actuarial Theory for Dependent Risks: Measures, Orders and Models. John Wiley & Sons.

[8] Nadarajah, S., Afuecheta, E., & Chan, S. (2017). A compendium of copulas. Statistica, 77(4), 279-328.

[9] Nooraee, N., et al. (2014). Measuring agreement between raters in medical imaging: Blomqvist's β as a robust alternative. Journal of Biomedical Science and Engineering, 7(8), 601–609.

[10] Agresti, A. (2010). Analysis of Ordinal Categorical Data (2nd ed.). Wiley.

[11] Stigler, S. M. (1986). The History of Statistics: The Measurement of Uncertainty before 1900. Harvard University Press.

[12] Trivedi, P. K., & Zimmer, D. M. (2007). Copula modeling: an introduction for practitioners. Foundations and Trends® in Econometrics, 1(1), 1-111.

[13] Geenens, G. (2024). (Re-) Reading Sklar (1959)—A Personal View on Sklar's Theorem. Mathematics, 12(3), 380.

[14] Dorey, M., & Joubert, P. (2005). Modelling copulas: an overview. The Staple Inn Actuarial Society, 1-27.

[15] MacKenzie, D., & Spears, T. (2014). 'The formula that killed Wall Street': The Gaussian copula and modelling practices in investment banking. Social Studies of Science, 44(3), 393-417.

[16] Patton, A. J. (2009). Copula–based models for financial time series. In Handbook of financial time series (pp. 767-785). Berlin, Heidelberg: Springer Berlin Heidelberg.

[17] Kole, E., Koedijk, K., & Verbeek, M. (2007). Selecting copulas for risk management. Journal of Banking & Finance, 31(8), 2405-2423.

[18] Genest, C., & MacKay, J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. The American Statistician, 40(4), 280-283.

[19] Haugh, M. (2016). An Introduction to Copulas. IEOR E4602: Quantitative Risk Management. Lecture Notes. New York: Columbia University

[20] Tinungki, G. M., Siswanto, S., & Najiha, A. (2023). The Gumbel Copula Method for Estimating Value at Risk: Evidence from Telecommunication Stocks in Indonesia during the COVID-19 Pandemic. Journal of Risk and Financial Management, 16(10), 424.

[21] Venter, G. G. (2002, March). Tails of copulas. In Proceedings of the Casualty Actuarial Society (Vol. 89, No. 171, pp. 68-113).

[22] Nelsen, R. B. (2003, March). Properties and applications of copulas: A brief survey. In Proceedings of the first brazilian conference on statistical modeling in insurance and finance (pp. 10-28). Sao Paulo: University Press USP.

[23] Nelsen, R. B. (2005). Dependence modeling with archimedean copulas.

[24] Deng, L., Smith, M. S., & Maneesoonthorn, W. (2024). Large skew-t copula models and asymmetric dependence in intraday equity returns. Journal of Business & Economic Statistics, 1-17.

[25] Hintz, E., Hofert, M., & Lemieux, C. (2022). Computational Challenges of t and Related Copulas. Journal of Data Science, 20(1).

[26] Frees, E. W. (2009). Regression modeling with actuarial and financial applications. Cambridge University Press.

[27] Tawn, J. A. (1988). Bivariate extreme value theory: models and estimation. Biometrika, 75(3), 397-415.

Appendix

Appendix A

.On the R Package: Vine Copula
Here, we provide a generic R-code based on the Vine Copula package which is used in the main body of the text for selecting the best possible bivariate copula for the four different insurance datasets:

```
install.packages("copula")
library("copula")
m<-pobs(a)
n<-pobs(b)
 install.packages("VineCopula")
library("VineCopula")
selectedCopula<-BiCopSelect (m,n,familyset=NA)
summary(selectedCopula)
```

- **Remark 1** In the above code, a and b are the transformed (on a log (to the base e) scale) variable values corresponding to two components of the associated bivariate data.
- **Remark 2** The best-fitted bivariate copulas mentioned here do not possess a closed form of expression in terms of their density function (i.e., the p.d.f.). However, in order to obtain the p.d.f. of each of these copulas, one may use R. Next, we provide an example as to how one can simulate from the p.d.f. of a Survival BB1 copula with specific parameter choices in R.

Simulate from a bivariate BB6 copula:

```
 install.packages("VineCopula")
library("VineCopula")
BB6<-BiCop( family = 8 , par =0.25 , par2 = 0.75)
sim<-BiCopSim( 10000, BB6).
```

Appendix B

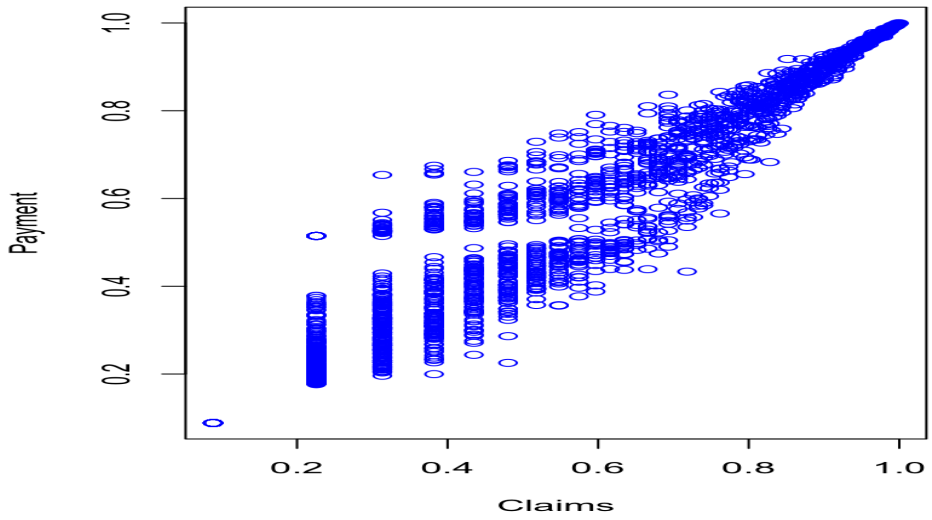


Figure 1: Scatter plot of Claims & Payment.

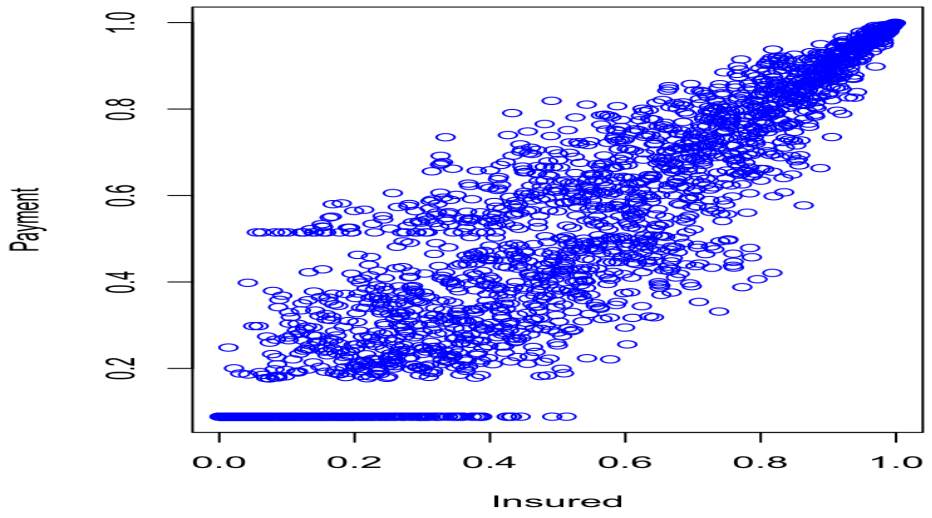


Figure 2: Scatter plot of Insured and Payment.

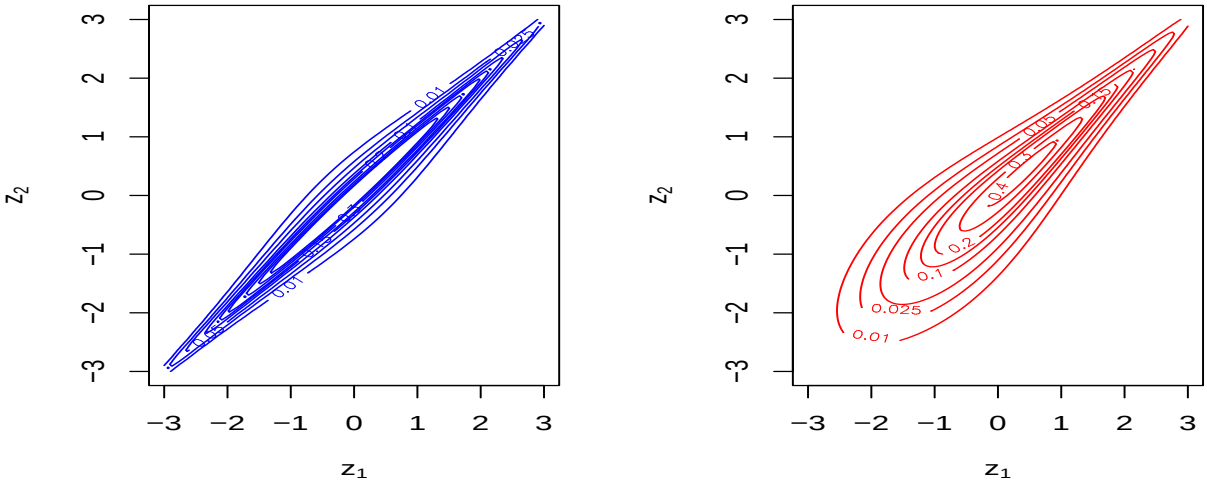


Figure 3: Contour plots of Claims & Payment (right panel) and Insured and Payment (left panel).

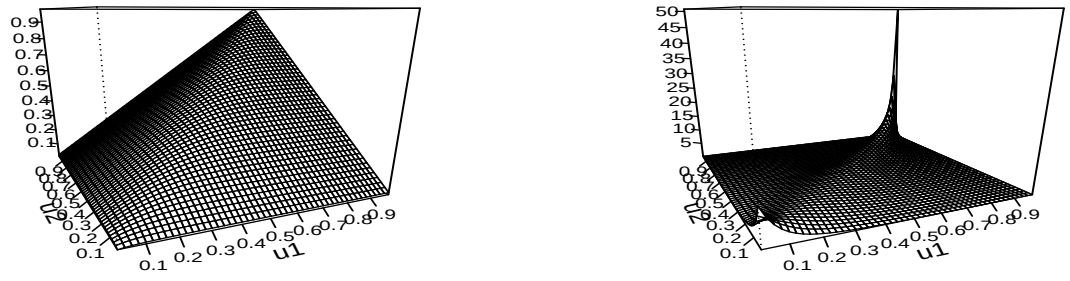


Figure 4: BB6 copula cdf and pdf with parameter values (1.59, 2.81) for Insured and Payment.

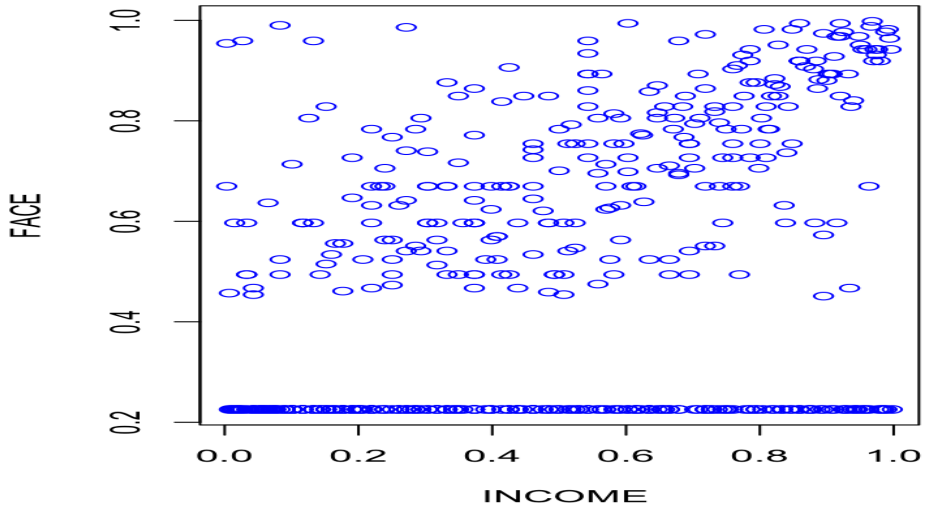


Figure 5: Scatter plot of Income & Face.

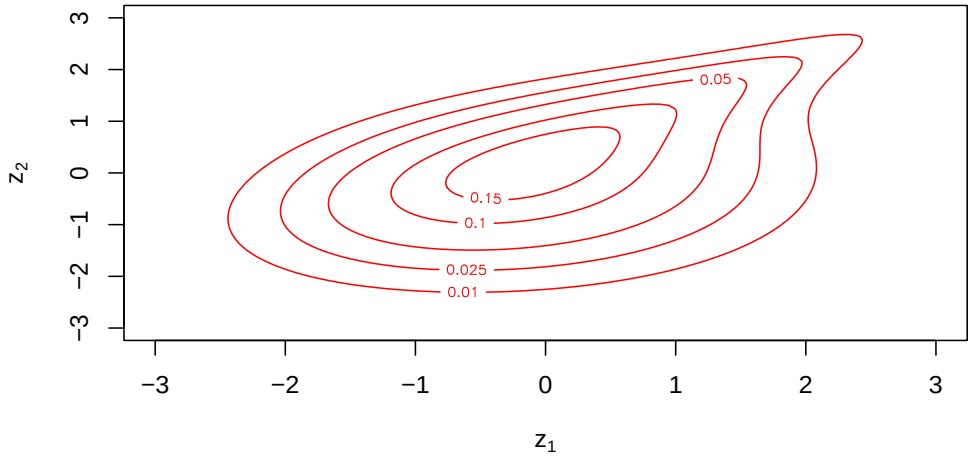


Figure 6: Contour plot of Income & Face.

Figure 7: Student t-copula pdf and cdf with parameter values (0.98, 2) for Claims and Payment.

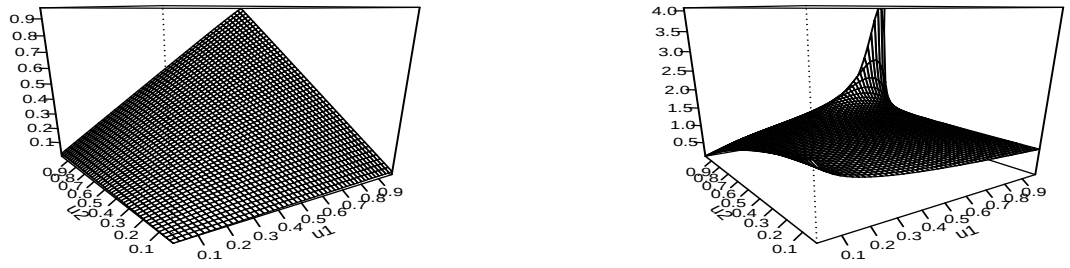


Figure 8: Tawn Type 1 Copula PDF and CDF with parameter values (1.84) and (0.49) for Income and Face.

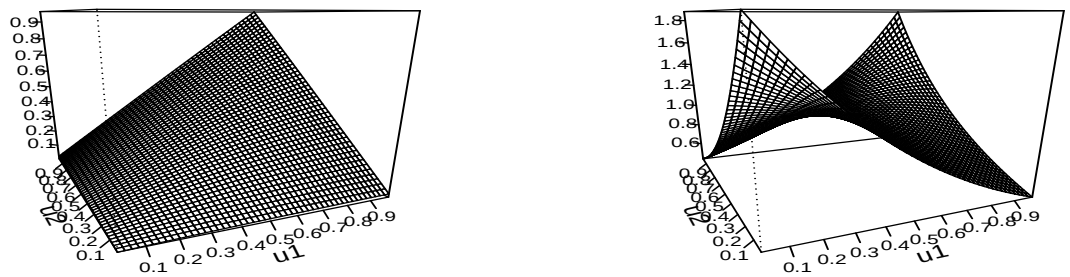


Figure 9: Frank Copula PDF and CDF with parameter value (1.59) for Income and BorrowCVLifePol.

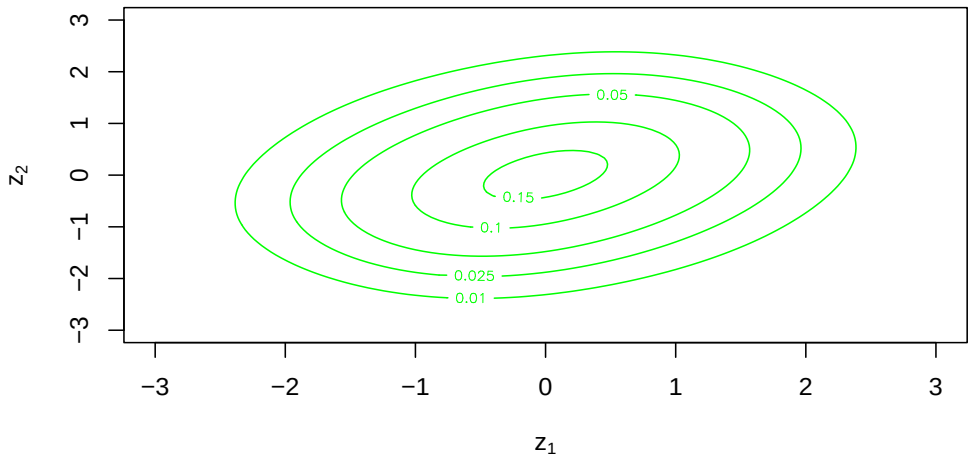


Figure 10: Frank Copula Scatter plot with parameter value (1.59) for Income and BorrowCVLifePol.

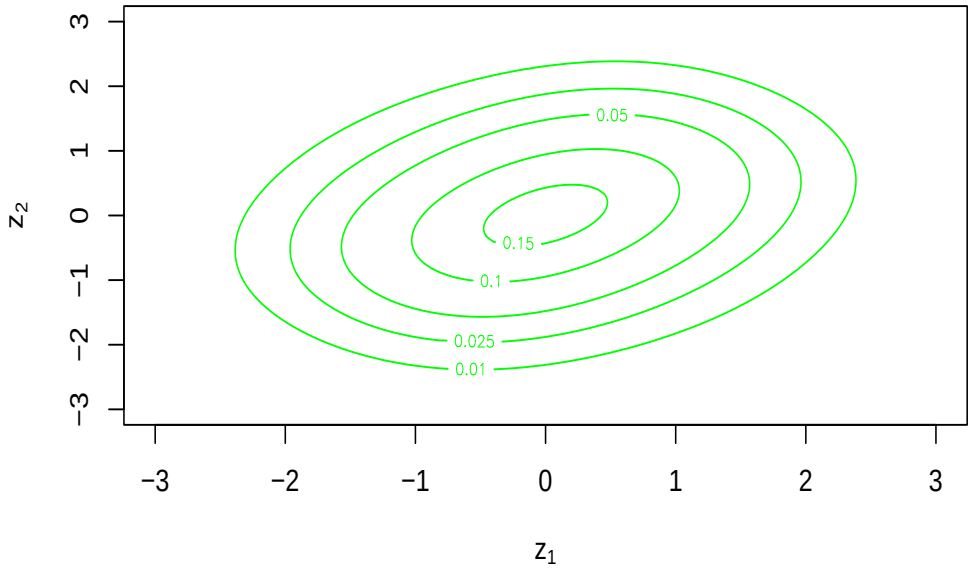


Figure 11: Frank Copula Contour plot with parameter value (1.59) for Income and BorrowCVLifePol.

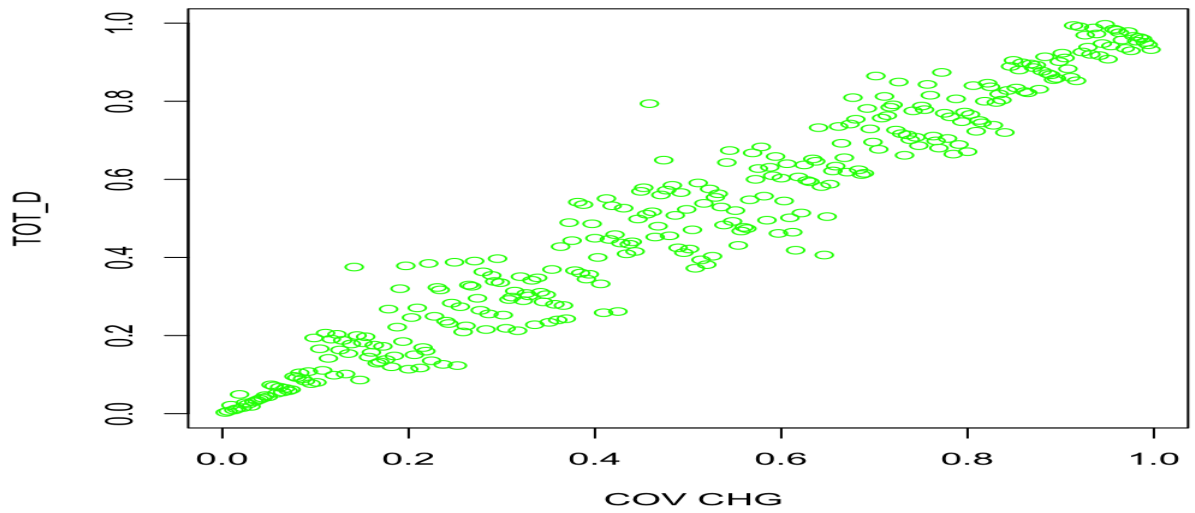


Figure 12: Survival Gumbel Copula Scatter plot for COV_CHG and TOT_D.

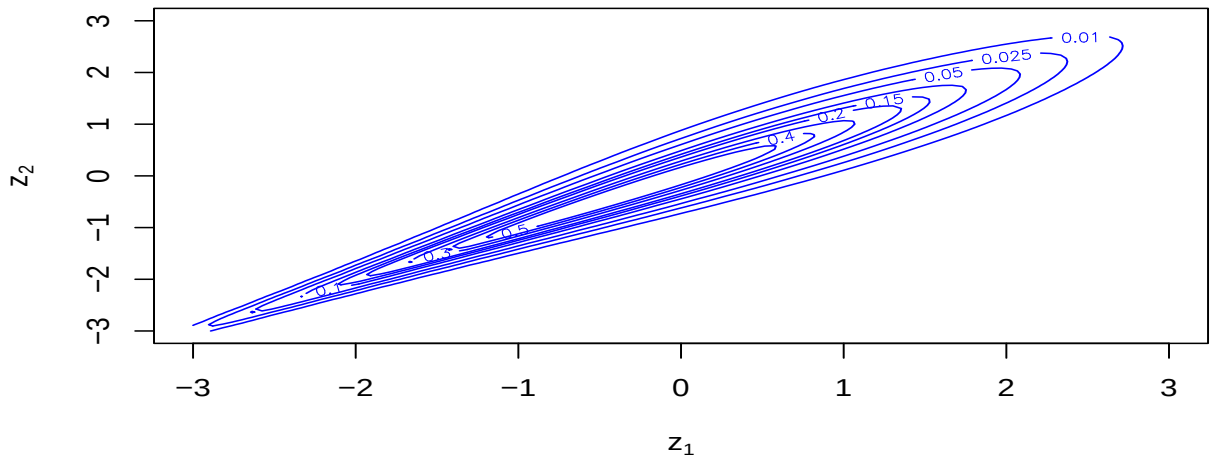


Figure 13: Survival Gumbel Copula Contour plot for COV_CHG and TOT_D.

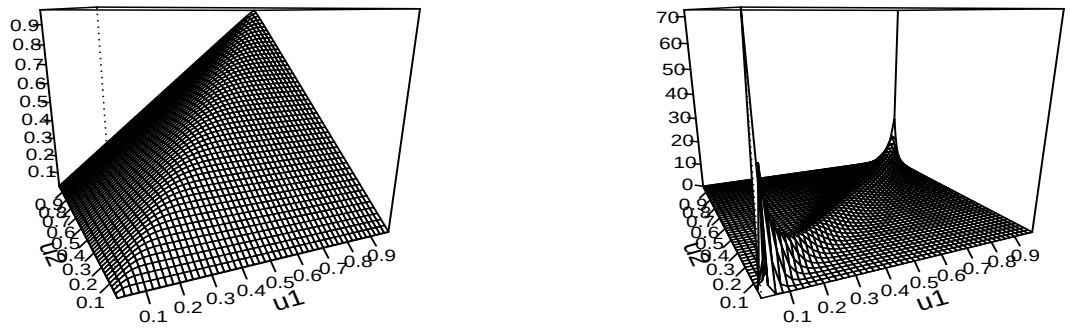


Figure 14: Survival Gumbel Copula PDF and CDF with parameter value (6.12) for COV_CHG and TOT_D.

Spanish Journal of Statistics
Vol. 7 No. 1 2025, Pages 39–73
doi:https://doi.org/10.37830/
SJS.2025.1.03

Regular Article

Gender gap and spatial disparities in the evolution of literacy in Spain, 1860-1910

José Manuel Gutiérrez

Universidad de Salamanca, Spain, jmgut@usal.es, ORCID iD: 0000-0002-2576-378X

Gloria Quiroga

Universidad Complutense de Madrid, Spain, mariagloria.quiroga@pdi.ucm.es, ORCID iD: 0000-0002-9825-0416

Received: January 12, 2025. Returned: June, 30, 2025 *Revised:* July, 21, 2025. *Accepted: October 3, 2025.*

Abstract: This article considers the dynamics of Spanish literacy in the period 1860-1910, characterized by local councils' responsibility of public elementary education. To this end, it is built a harmonized series of the literacy of the population aged ten or over, disaggregated by sex and province. Marked spatial differences and a very large gender gap can be observed. Five clusters are determined according to the male literacy rates of the provinces in 1860; these clusters prove to have explanatory power all along the period and for both sexes. A parsimonious statistical model of the evolution of male literacy during the period, introducing linguistic variables, shows a considerable temporal stability of the spatial distribution of male literacy. The model of the evolution of female literacy presents similarities with that of male literacy, although now the initial state (in 1860) is not described by female literacy, but yet by male literacy. All in all, the evolution of literacy in Spain between 1860 and 1910 did not follow the spatial pattern of the economic modernization process. Besides, there was no correlation between birth rates and literacy rates of children, for both sexes, and the same can be said of the correlation between urbanization and literacy. Considering the West European context, the Spanish literacy process during the period 1860-1910 was a failure, except for the geographical area of the top cluster.

Keywords: Historical censuses, literacy, nineteenth century, Spain, Official Statistics

MSC: 91F10, 62P25

1 Introduction

Human capital is a fundamental determinant of long-term economic and human development (Hanushek and Woessmann, 2010). In this sense, female illiteracy is a particular obstacle to economic

progress[1]. Sandberg (1982) argued that literacy rates, as a variable capturing the stock of human capital of the population, anticipate future increases in per-capita income. In general, all Western countries, except for Britain[2], greatly improved their literacy rates during the early stages of modern economic growth. In principle, industrialization requires previous physical and, above all, human capital (as Galor's unified growth theory points out[3]), and a fast accumulation of them. Literacy allows the accumulation of information.

We consider herein the dynamics of literacy in Spain during the period between mid-nineteenth century and the First World War. From 1860 the Spanish censuses report reliable literacy data. The first aim of the article is to build a harmonized series of the literacy of the population aged ten or over, disaggregated by sex and province (NUTS-3 level)[4], for the period 1860-1910. Vilanova Ribas and Moreno Julià (1992) compiled a harmonized series from 1887, but before 1887 the literacy data provided by the censuses are not broken down by age. In fact, for the period 1887-1910, we make use also of the censual literacy data not only disaggregated by sex and province, but also by age intervals.

This paper deals with the *evolution* of literacy between 1860 and 1910. Apart from the understanding that it provides *per se*, disaggregation by sex turns out to be a useful tool to study the dynamics of the literacy process during the period.

Not all historical data are of equal quality. Census data are usually the best when available and we have favoured them. For example, census data on the sectoral distribution of the labour force are the best available indicators of economic modernisation for the period under consideration. Exploratory data analysis allows for the consideration of variables that present problems of collinearity with other variables or whose available data are of inferior quality.

The period 1860-1910 is characterized by institutional factors that impinge on the funding of public elementary education. The comprehensive Public Instruction Law (1857) (known as Moyano Law after the incumbent minister), valid throughout this span of time, declared primary education compulsory (at least in theory) for all children between the ages of six and nine, and also free in public schools for the certified poor. The financing of public elementary education was left to municipalities. On the other hand, the *desamortización* of 1855 had confiscated the assets of (generally Catholic) educational foundations of all kinds[5] and most of the land belonging to the municipalities (until then a major source of the income of local councils). Near the end of the period, in 1901, the (central) state assumed the direct payment of teachers' salaries (although taking a percentage of municipal taxes in exchange), and in 1909 compulsory education was extended to the ages between six and twelve. Until 1910, the funding of public elementary education fell fully on the shoulders of municipalities[6].

[1] See Bowman and Anderson (1963), Núñez (1992, 2003a) and Sarasúa (2002, 2019).

[2] A paradox appears in the English case. Although there is no complete consensus on the level of English literacy rates before and during the early industrial revolution, it seems that they improved substantially between 1642 and 1750, and then stagnated until 1815-1830, especially in industrial centres. This evolution led many researchers to assume the hypothesis that education was not an essential requirement for modernization. All the same, that stagnation was a consequence of the industrialization process itself, which initially needed unskilled labour (also regrettably including children, making higher the opportunity cost of attending school), and of the intense urbanization process, which was not accompanied by a similar increase in the educational supply. Once this first phase of industrialization had been overcome, the need for a more qualified workforce and the extension of the franchise meant that from 1840 the literacy process became widespread in England (see West (1978); Schofield (1973); Mitch (1993, 2013); Pleijt et al. (2020)).

[3] Galor (2011)

[4] The 1887 census is the only one providing literacy data by age at a higher level of disaggregation, that of judicial districts (*partidos judiciales*).

[5] These were assets that had survived the *desamortizaciones* of 1836-1837 and 1841, which affected the properties of the Catholic Church.

[6] See Terrón Bañuelos (1997).

Only from that year public primary education was made free for all, and the state began to contribute to the financing of primary education[7].

The Moyano Law must be seen in the context of the laws which organised the alphabetization process in most Western European countries: the state establishes a national system of universal elementary education, imposing on local authorities the obligation to create and maintain schools. Financing is essentially dependent on local authorities and on the fees paid by parents[8], although it is free for the poor. The differences between countries lie in the time of commencement of the process and the extent to which the legal provisions were actually implemented.

C.E. Núñez[9] has carried out relevant studies on literacy and education in Spain, showing that, although Spanish literacy rates performed poorly at the national level in the period 1860-1910, large regional differences can be observed. According to Núñez, the northern half of the country, except for Galicia, was more literate than the south-eastern Mediterranean coast, and some areas had in 1860 male literacy rates similar to those of the most advanced countries of Europe, whereas others displayed levels among the lowest in Western Europe. Fifty years later, even when male literacy rates had improved, the differences between provinces persisted. For their part, female literacy rates were low and much more homogeneous in 1860, which entailed a high gender differential overall, much higher in the more literate areas.

The spatial distribution of Spanish literacy reflected by the censuses, in which the highest levels of literacy correspond to rural areas of the northern plateau, has led to attempts at explanation. Beltrán Tapia and Martínez-Galarraga (2018) builds a formal model in which the ratio of day labourers to the total agricultural population turns out to be significant to explain literacy. On the other hand, Reher (2023) stresses the existence of important regional differences in the perceived value of literacy and education and argues that these cleavages go beyond the importance of economic structures and have deep historical roots.

Beltrán Tapia et al. (2021) studies the spatial convergence of literacy in Spain with data disaggregated by municipality, showing a positive correlation between the (global) literacy rates in 1860 and the change in the literacy rates between 1860 and 1900, and a negative correlation between the literacy rates in 1900 and the change in the literacy rates between 1900 and 1930.

We focus here on the evolution of literacy in the period 1860-1910. As for any analysis of the dynamics of a system, an essential point is how much the initial state explains the final state. Five clusters are determined according to the similarity measure naturally provided by the male literacy rates of the provinces in 1860 (see Section 4). The resulting clusters turn out to be spatially contiguous to a high degree.

The study of the evolution of literacy is facilitated by considering its age structure (available after 1887), as low literacy rates in 1860 may obscure the literacy effort if only "biologically linked" literacy rates are used in later years. Two literacy rates may have a *biological link*, in so far as they share part of their underlying populations. For example, the male rate (for men 10 years old and over) in 1900 has a biological link with the same rate in 1910, because the men 10 years old and over in 1900 still surviving in 1910 are part of the men 10 years and over in 1910. In contrast, the male rate in 1900 has no biological link with the male rate for boys (between 11 and 15 years old) in 1910. Assuming that the acquisition of literacy for those older than a certain age becomes increasingly unlikely, a very

[7]Centralization in primary school systems fostered eventually literacy in Latin European countries like Portugal (see Reis (1993); Nunes (2003); Gomes and Machado (2020)), Italy (see Zamagni (1993) and Cappelli and Vasta (2020)) and Spain (see Núñez (1992)).

[8]Fees were eliminated later: in 1881 in France, in 1889 in Prussia and from 1910 in Spain. The Casati Law of Italy (1859) established full gratuity (for the first two years of education) upon its enactment.

[9]Núñez (1992, 1993a, 1997, 2003a,b, 2010).

low literacy rate in some year holds back the biologically linked literacy rates in later years (this is particularly the case with female rates).

The clusters considered above prove to have explanatory power all along the period and for both sexes. The 6 provinces of the first cluster, the "Castilian core", had surpassed the threshold of 75% male literacy in the 1870s, and still headed the list in 1910. All the 12 provinces of the fifth cluster, "South and East", were among the bottom 14 provinces by male literacy in 1910, with values between 30% and 45%. The statistical results show a considerable temporal stability of the spatial distribution of male literacy. We obtain a rather parsimonious model of the evolution of male literacy during the period, with coefficient of determination round 90% and only three regressors: the initial state (in 1860) and two linguistic variables. There is no influence of economic modernization: the proportion of the male active population working in agriculture, or the same proportion of those working in industry, turn out to be non-significant.

The harmonized literacy series for the period 1860-1910 facilitates especially the analysis of the female literacy development. The overall female rate in 1860 was low enough, 11.2%, in contrast with the certainly lacklustre male rate, 38.9%. Unlike male literacy, there is no clear spatial pattern of female literacy in 1860. Generally speaking, the main feature of the dynamics of female literacy in the period is that the greatest increase in female literacy did not occur in the most economically developed or urbanized areas, but rather in those provinces that in 1860 had the highest levels of male literacy[10]. In fact, the female literacy of 1910 is predicted quite well by the male literacy of fifty years before. The resulting model of the evolution of female literacy during the period presents similarities with that of male literacy, although now the initial state (in 1860) is not described by female literacy, but yet by male literacy. Indeed, female literacy in 1860 turns out to be non-significant for the evolution of female literacy in the period 1860-1910.

The evolution of literacy in Spain between 1860 and 1910 did not follow the spatial pattern of the economic modernization process, as it is usually the case[11]. On another note, birth rates and urbanization are two issues impinging on the financial constraints of parents and local councils, the essential funders of elementary education during the period. In this sense, there was no correlation between birth rates and literacy rates of children, for both sexes, and the same can be said of the correlation between urbanization and literacy (data of 1910). Certainly, the Moyano Law allowed small villages to maintain mixed-sex schools and opt for less paid (and qualified) teachers. But, ultimately, the list of per capita investors in public primary education was headed by rural provinces with high literacy (data of 1908).

All in all, the Spanish literacy process during the period 1860-1910 was a failure. If we consider eight West European countries able and willing to provide reliable censual literacy data during the 19th century, Spain had the fifth literacy level and the widest gender gap at the beginning of the period, and the eighth literacy level and still the widest gender gap at the end. Certainly, there were drastic regional differences, and, by 1910, literacy was almost universal among girls in the Castilian core, while only around one quarter of girls were literate in the South and East cluster.

The article is organized as follows. Section 2 examines the level of Spanish literacy in the period 1860-1910 within the European context. Section 3 deals with the methodology used to build a harmonized series of the literacy of the population aged ten or over; the series is provided in Appendix 1. Section 4 and Section 5 present the evolution of male and female literacy, respectively, and discuss

[10]See Núñez (1992) p.122: "the progress [of female literacy] was more intense where male literacy was more widespread". Cf. also Beltran Tapia et al. (2021) (footnote 43): "female literacy grew rapidly in municipalities where male literacy was already high".

[11]See Smith (1976. Book I-Chapter 10), Hanushek and Woessmann (2010) and Sandberg (1982). On the other hand, see Reher (1997).

the link between female literacy at the end of the period and male literacy 50 years earlier. Section 6 concludes.

2 The European context

Literacy concerns primarily communication. An individual who can communicate with another by means of written language is regarded as a "literate" person, and one who does not possess this ability is considered an "illiterate" (see UNESCO (1957)). Thus, literacy comprises both reading and writing[12].

In order to estimate the literacy level of a population, we face considerable problems, conceptual and practical, especially when we go back in time[13]. The advent of modern censuses in the mid-19th century opened new possibilities for the measurement of literacy. In modern censuses data were obtained on all individuals present in the household on the specified census day. Information was self-reported by the household heads through household forms (later individual forms). A field force of professional enumerators was employed to assist in the process from house to house (especially if there was no one in the house who could write) and collect the forms[14].

All modern censuses had a similar basic methodology, and thus comparisons between countries are made easier[15]. It must be considered when the data refer to literacy (ability to read and write) or semi-literacy (ability to read). We shall refer here to literacy data in Western Europe. Sometimes the first censuses gave literacy data without distinguishing ages, setting or not setting a minimum age to obtain the data (4, 5 or 6 years); in these cases, obtaining a true literacy rate (from 10, 11, 12 or 15 years old) requires estimation work (which can be quite precise if the necessary auxiliary data are available, as it is usually the case).

The first (modern) census in Western Europe with literacy data is that of Ireland in 1841. Then we have the censuses of Spain (1860), Italy (1861), Belgium (1866) and France (1866). More censuses are added later, carried out in tune with the political borders of their time. Some European countries have never included questions on literacy in their censuses (e.g., the United Kingdom (except Ireland) and Denmark), or have done so very late (e.g., Sweden, in 1930).

Taking into account the political entities in Western Europe where censual literacy data are available already in the nineteenth century (Ireland, Spain, Italy, Belgium, France, Prussia, Cisleithania[16] and Finland), there are three countries where the literacy rate had not reached the 75% level before

[12]The modern UNESCO proposed definition reads: "A person is considered *literate*, who can both read with understanding and write a short simple statement on his everyday life" (see UNESCO (1957); the proposal was made by a committee in 1951).

[13]As for the pre-statistical age, the main tool of analysis is considering who could sign and who could not sign in documents (such as marriage certificates, deeds, wills, etc.), and even the quality of the signatures. Apart from the issue of how representative of the population is the sample in each case, the ability of an individual to write his/her name does not entail, in principle, a general ability to read or write, although there can be statistical correlations (see Furet and Sachs (1974)).

[14]See Baffour et al. (2013) about modern censuses and their evolution.

[15]See UNESCO (1953) about problems arising in censual literacy data. Besides, when literacy is self-reported there are attendant issues of possible upward bias. A test was implemented in 1864 to check the accuracy of the literacy self-report of the conscripts in France, with the result that their statements were highly reliable (see Furet and Ozouf (1977)).

[16]The term "Cisleithania" denotes the northern and western part of Austria-Hungary, containing Austria proper, present-day Czechia and other *crown lands* ("Kronländer"). After the *Compromise* ("Ausgleich") of 1867, the Austrian Empire was transformed into the dual monarchy of Austria-Hungary, constituted by two parts, with their respective parliaments and governments: *Cisleithania* (the Austrian part) and *Transleithania* (lands of the "Archiregnum Hungaricum").

the First World War: Spain, Italy[17] and Finland[18]. In fact, the 75% threshold was reached in the three of them by the 1930-1940 interval, beyond the time span considered in this paper.

A comparative picture of the literacy processes in Western Europe on the eve of and during the Second Industrial Revolution is provided in Gutiérrez (2024), taking censual literacy rates as a yardstick to measure and compare literacy in different countries[19]. If only partial or insufficient censual data are available, literacy is assessed as if given by full censual data. A set of comparable (as far as possible) literacy data is built. The non-Spanish data in this section are in that paper or directly taken from the censuses referenced there.

It is worth highlighting the large gender gap in Spanish literacy rates. At the beginning of the period, in 1860, it is 27.7 points in Spain, while it is 15 points in Ireland (1861), 16.4 in Italy (1861), 7.5 in Belgium (1866) and 11.4 in France (1866). At the end of the period, in 1910, the gender gap is still 19 points in Spain, to be compared with 1.0 in Ireland (1911) or 3.4 in Belgium, but 12.4 in Italy (estimation, see Table 1 below).

On the other hand, Spain lost ground to Italy and Finland during the period 1860-1910. Table 1 focuses on the literacy rates of Spain, Italy and Finland. For each country and census, three percentages of literacy are stated: for men and women, separated by a hyphen, and the overall percentage (marked bold) in the bottom row. The minimum age limits are indicated in brackets under the name of every country (the figures from the 1861 census in Italy are for the population aged 12 or over, and those from the 1881, 1901 and 1911 censuses for the population aged 10 or over; data from the 1880 census in Finland are for the population aged 10 or over, and those from the 1900 and 1910 censuses for the population aged 15 or over). The dates of the censuses have been made homogeneous by linear interpolation. The problem arises that in Italy there are no data on literacy after 1881, but only on semi-literacy. In the table, a crude estimate (written in italics) of the Italian literacy rates in 1900 and 1910 has been given, subtracting from the (interpolated) Italian census percentages of semi-literates (out of the population aged 10 or more) the estimates of the percentages of those individuals able only to read (but not to write), assuming that the latter percentages are equal to those of Spain in the same year. These estimates are accurate for the purpose of the comparison between Spain and Italy (using them is equivalent to considering semi-literacy rates in both countries), but they are not so accurate beyond this comparison[20].

At the beginning of the period[21], Spain had the highest literacy level of the three countries. At the end of the period, it had the lowest. The comparison with Italy is particularly relevant (in the

[17]Spain and Italy were in this period predominantly agricultural economies and followed a *Latin pattern of modernization* (Tortella (1994), p. 5): relative backwardness in the nineteenth century and recovery in the twentieth century.

[18]Note that Finland followed the Swedish model (based on home instruction of the ability to read known texts: a set of selected religious texts, emphasizing submission to authority), with high *restricted semi-literacy* and low literacy until modern school systems were introduced. See Johansson (1977) and Tveit (1991).

[19]There is a typo in Table 1 (p. 43), in the last row of the "Slovenia" column: it says *88.5* and should read *81.5*.

[20]The percentage of semi-illiterates (people who can read but not write) was small in Italy (in 1861 it was 3.9% for men, 5.5% for women and 4.7% overall, and in 1881 it was 1.2% for men, 3.4% for women and 2.3% overall) and in Spain (in 1887 it was 2.2% for men, 4.5% for women and 3.4% overall, and in 1910 it was 1.0% for men, 2.3% for women and 1.7% overall (Vilanova Ribas and Moreno Julià, 1992).

[21]Prior to 1860, there are no safe data on literacy in Spain, but research based on the counting of signatures in various sources and regions (see Bennassar (1985), Rodríguez and Bennassar (1978) and Larquié (1981)) seems to show that the literacy level in Spain was similar to that in France during much of the Old Regime. At any rate, after the Spanish War of Independence against Napoleon (1808-1814), arguably the bloodiest event in Spain's modern history, it is clear that Spanish literacy entered a period of relative decline.

case of Finland it must be considered that in 1880 almost the whole Finnish population had at least restricted semi-literacy[22]).

	Spain (≥ 10)	Italy ($\geq 12, \geq 10$)	Finland ($\geq 10, \geq 15$)
1861	39.2-11.6 **25.1**	30.4-14.0 **22.17**	
1880	44.9-19.4 **31.7**	44.8-26.8 **35.8**	16.2-10.2 **13.1**
1900	52.7-30.5 **41.2**	*56.1-41.1* ***48.4***	41.1-36.5 **38.8**
1910	57.6-38.6 **47.7**	*66.4-54.0* ***60.0***	57.4-53.3 **55.3**

Table 1: Literacy rates in Spain, Italy and Finland.

However, as we shall see, in a predominantly rural area corresponding approximately to the original Castile, the threshold of 75% in the male literacy rate had been reached already in the interval 1871-1880 (in the 1877 census), whereas in a fifth of Spanish provinces it was less than 30%. At any rate, low female literacy was a burden spread throughout the country, to a greater or lesser degree, which meant that the literacy gender gap was very large (particularly in the most literate provinces), even when compared to Italy.

It is worth considering the internal spatial differences in literacy levels of other countries around 1870.

The most literate areas of France in 1866 were, apart from Paris and its surroundings, the north-eastern part of the country: Alsace, Lorraine, Franche-Comté and Champagne, with literacy levels above 75%. There was a gulf with the least literate areas, located in the eastern Pyrenees, much of Brittany and a strip in central France covering Perigord, Berry, Bourbonnais and the east of Limousin; in all of them, the percentage of literates did not reach a third of the population in 1866.

In Italy (1871 census, semi-literacy rates for the population aged 6 or over) literacy decreased from north to south, with the highest values in Piedmont (66.3% for males and 49.2% for females) and Lombardy (59.3% for males and 50.3% for females), and very low values in the south, with the smallest values in Basilicata (19.1% for males and 5.3% for females) and Calabria (20.9% for males and 5.3% for females)[23].

Despite the high literacy of Prussia already in 1871, a swathe of land along the far east of the country (Prussia proper, Posen and Upper Silesia) had literacy rates below 75%, with a minimum of 57.1% in Bromberg. This area corresponded to the districts with a sizeable Polish-speaking minority (Prussia proper) or a Polish-speaking majority (Posen and Upper Silesia). The rest of the country had literacy rates above approximately 90%, except for part of Pomerania (83.3% in Köslin and 84.1% in Stralsund). The highest values were in Berlin (97.4%) and the rural district of Sigmaringen (97.2%).

Unlike France and Prussia, in Italy there were still intense spatial differences at the end of the period (1911 census, semi-literacy rates for the population aged 6 or over): in Piedmont the literacy rates were 90.9% for males and 87.2% for females, whereas in Calabria they were 40.5% for males and 21.9% for females.

[22]Semi-literacy was high in Finland: in 1880 it was 81.0% for males, 87.6% for females and 84.4% overall, and in 1910 it was 41.3% for males, 45.7% for females and 43.6% overall (Myllyntaus, 1990). These Finnish figures should be understood in the sense of restricted semi-literacy (see the footnote above), as Finland followed the Swedish model.

[23]See Noble (1965) (p. 300).

3 Censuses and literacy rates

The first aim of this paper is to provide a census-based time series from 1860 to 1910 of the Spanish literacy rates, disaggregated by sex and province, for the population 10 years old and over (see Appendix 1)[24].

There are five "complete censuses" with literacy data in the period: 1860, 1877, 1887, 1900 and 1910 (the "incomplete censuses" of 1857 and 1897 are not relevant here)[25]. From 1887 onwards combined data of literacy and age appear in the censuses; in 1860 and 1877 these two kinds of information are given separately. There are 49 provinces in Spain throughout the period[26]. The city of Ceuta is always included in the province of Cádiz[27], and the city of Melilla is grouped with other small "plazas de soberanía" and is treated for census purposes as one more province, with which 50 divisions appear in the censuses. Colonial data are not considered in this paper.

Both the de *facto population* and the *de jure population* (*usually resident population*) are provided in all censuses (except in 1860, where only the de facto population is given). Literacy figures are taken from the de facto population. The number of individuals unspecified for literacy is also provided in all censuses (except in 1860), and it is always low (0.04% over the population aged 10 or over in 1877, 0.1% in 1887, 0.1% in 1900, 0.3% in 1910).

We have not excluded from the population those individuals unspecified for literacy when calculating the literacy rates (which is equivalent to considering them illiterate). Obviously, these literacy rates are lower than when individuals unspecified for literacy are excluded from the reference population (as it is done most frequently).

Now we set up some notation. Let us consider a certain group of people (Spain, a province, the women of that province, etc.), which is clear from the context. P_{10} is the population aged 10 or over in this group, and A_{10} is the literate population aged 10 or over. In general, P_k is the population aged k or over and A_k is the literate population aged k or over. We now define $T_k = \frac{A_k}{P_k}$, the literacy rate of individuals aged k or over. Similarly, P_{11-15} is the population aged 11-15 (inclusive) and A_{11-15} is the literate population aged 11-15. In general, P_{m-n} is the population between m and n years (inclusive) and A_{m-n} has the obvious meaning. The literacy rate for individuals aged between m and n (inclusive) is $T_{m-n} = \frac{A_{m-n}}{P_{m-n}}$.

Our purpose is to obtain T_{10} for each province, for men and women, in the 1860, 1877, 1887, 1900 and 1910 censuses. Also, the child literacy rates T_{11-15} in the 1887, 1900 and 1910 censuses are to be calculated.

The values of P_{10}, P_{11-15}, A_{10} and A_{11-15} can be found immediately, for each province, for both men and women, in the 1887, 1900 and 1910 censuses, and thus T_{10} and T_{11-15} can be calculated for these censuses. The rest of this section provides an exposition of how we estimate the value of T_{10} (through the values of P_{10} and A_{10}) in the 1860 and 1877 censuses, in each province, for men and for women.

In the 1877 census we can calculate P_{10} directly. On the other hand, in the 1860 census the only relevant values given are P_0, P_{0-5} and P_{6-10}. We estimate P_{10-10} (and, from there, P_{10}) by calculating

[24] Individuals who do not state their level of literacy are considered illiterate in this paper.

[25] See Melón (1951), Cusidó i Vallverdú and Gil-Alonso (2012), for comparative analyses of the Spanish censuses of the period.

[26] This geographical division has remained stable since its creation in 1833 until 1927, when the Canary Islands split into two provinces, *Las Palmas* and *Santa Cruz de Tenerife*, and has remained so until today. We shall use the official names of the provinces (as appearing in the censuses of the period) throughout.

[27] In the 1887 census the literacy data of Ceuta are not even provided separately and are subsumed into those of the judicial district (*partido judicial*) of Algeciras.

the coefficient $\frac{P_{10-10}}{P_{6-10}}$ from the 1877 census and using it as an estimate of the same coefficient in the 1860 census.

What is left is to estimate A_{10} in 1860 and 1877, in each province, for men and for women (as for literacy data, only A_0 is provided in these censuses). In order to do this, if we write $T'_{0-9} = \frac{A_{0-9}}{P_{6-9}}$, it suffices to estimate T'_{0-9}. Setting $T'_6 = \frac{A_0}{P_6}$ (a value known in 1860 and 1877), we now propose the simple regression model (SRM), separately for men and women:

$$T'_{0-9} = \beta_0 + \beta_1 T'_6 + \epsilon$$

where the variables run through the values of the provinces. The first census in which the values of T'_{0-9} are known is that of 1887. We now estimate the values of the parameters β_0 and β_1 with the data from 1887, intending to use them thereafter to "predict backwards" the values of T'_{0-9} in 1860 and 1877[28]. The results are:

- Men: $\beta_0 = -0.1039$, $\beta_1 = 0.6848$, with $R^2 = 0.8312$
- Women: $\beta_0 = -0.0364$, $\beta_1 = 0.7796$, with $R^2 = 0.8487$

We have now a strictly increasing functional relation. In order to assess the "backwards predictive" capacity of the model with these parameters, we apply the equation with the estimated parameters to predict the values of T'_{0-9} in 1900 and 1910, the two censuses after 1887. In these censuses we know the true values of T'_{0-9}, and we can see to what extent the prediction deviates from the true value. For 1900 the resulting coefficients of determination are as follows:

$$\begin{aligned} \text{Men} &: R^2 = 0.8432 \\ \text{Women} &: R^2 = 0.8248 \end{aligned} \tag{1}$$

For 1910 the R^2 values are 0.7476 for men and 0.7840 for women. As these coefficient of determination results are rather good at predicting the T'_{0-9} values in the two subsequent censuses (1900 and 1910), with the parameters calculated using the 1887 data, it can be plausibly concluded that the "backward prediction" in the two antecedent censuses (1877 and 1860) will also be good. It must be taken into account that, as it is discussed below, literacy changed less in the period 1860-1887 than in the period 1887-1910[29], and thus the backward predictions for 1860 and 1877 are not likely to be worse than the predictions for 1900 and 1910.

We have tried other methods to estimate the T'_{0-9} values in 1877, using two regressors, apart from the independent term. Specifically, we have considered the multiple regression model (MRM):

$${}^{n}T'_{0-9} = \beta_0 + \beta_1 \, {}^{n}T'_6 + \beta_2 \, {}^{n+1}T'_{0-9} + \epsilon$$

where ${}^{n}T'_{0-9}$ and ${}^{n}T'_6$ indicate values in the n-th census and ${}^{n+1}T'_{0-9}$ in the subsequent $(n+1)$-th census.

[28] Melilla presents atypical characteristics and is not included in the estimation of the model parameters, here and in subsequent regression models (and correlation coefficients) appearing in the paper. Accordingly, the T'_{0-9} values of Melilla in 1860 and 1877 are estimated, for men and women, by multiplying the $\frac{T'_{0-9}}{T'_6}$ ratio of Melilla in 1887 by the T'_6 value of Melilla in 1860 and 1877, respectively. On the other hand, the values of T'_6 for women in 1860 of five provinces are very low and far out of the range of the values of 1887, and this results in slightly negative estimations of T'_{0-9} which have been taken zero instead; at any rate, the affected figures are negligible.

[29] Although the T_{10} rates will be used later for this discussion (which for 1860 and 1877 are obtained using the parameters considered now), there is no circularity in the reasoning, since the result is the same using the cruder rates T'_6, which can be obtained directly from the census data.

We now estimate the values of the parameters β_0, β_1 y β_2 with the data from 1887 (n-th census) and 1900 ($(n+1)$-th census). The results are:

- Men: $\beta_0 = -0.0538$, $\beta_1 = 0.2056$, $\beta_2 = 0.7585$, with $R^2 = 0.9246$ and $\overline{R}^2 = 0.9214$
- Women: $\beta_0 = -0.0251$, $\beta_1 = 0.2937$, $\beta_2 = 0.5746$, with $R^2 = 0.9254$ and $\overline{R}^2 = 0.9221$.

The multiple regression model (MRM) certainly provides a better fit to the data (from 1887 and 1900) than the simple regression model (data from 1887), for both men and women. Another question is its ability to "predict backwards" the values of T'_{0-9} in 1877 (the values of the regressor ${}^{n+1}T'_{0-9}$ are known, as they correspond to the 1887 census). In order to compare this ability in the two models, we apply the multiple regression equation with the parameters now estimated to predict the values of T'_{0-9} in 1900, the post-1887 census. In 1900 we know the true values of ${}^{n}T'_{0-9}$ (as we know those of the regressor ${}^{n+1}T'_{0-9}$, now corresponding to 1910) and we can see how far the prediction deviates from the true value. The coefficients of determination are as follows:

$$\begin{aligned} \text{Men:} \quad & R^2 = 0.6806 \quad \text{and} \quad \overline{R}^2 = 0.6667, \\ \text{Women:} \quad & R^2 = 0.7704 \quad \text{and} \quad \overline{R}^2 = 0.7605. \end{aligned} \tag{2}$$

Comparing (1) and (2), the results of the multiple regression model (MRM) are worse than those of the simple regression model (SRM) at predicting the values of T'_{0-9} in 1900, for both men and women. Therefore, we choose the simple regression model to "predict backwards" the values of T'_{0-9} in 1877[30]. Even more so, we discard a similar multiple regression model (in which a regressor of the type ${}^{n+2}T'_{0-9}$ would have to appear) to estimate the values of T'_{0-9} in 1860, and we also maintain in this case the simple regression model.

4 The evolution of male literacy

4.1 The starting point

Spain in 1860 is a predominantly agrarian country, with little modern industry, except in the province of Barcelona. Politics is unstable, marked by *pronunciamentos* and uprisings, and it will continue to be so until 1876, with the end of the third Carlist (civil) War, at the beginning of the Bourbon Restoration. Spanish is the predominant language, although Basque, Catalan-Valencian and Galician are also spoken (see below).

Based on census data, we can estimate that the Spanish literacy rate for males (aged 10 and over) was 38.9% in 1860. This global level of literacy says nothing about an important part of reality: a heterogeneous and unusual spatial distribution of male literacy.

We introduce a clustering of the 49 Spanish provinces according to their male literacy rates in 1860 (i.e., the similarity measure is the modulus of the difference between the literacy rates). In fact, we proceed by ranking the provinces in descending order of their male literacy rates, assigning proportional indices to the male literacy rates of the provinces (with 100 corresponding to the mean

[30] Two alternative multiple regression models have also been considered, in which the second regressor is ${}^{n+1}T_{10-10}$ or ${}^{n+1}T_{16-20}$, instead of ${}^{n+1}T'_{0-9}$, but in them the coefficients of determination and the corrected coefficients of determination are worse than those of the multiple regression model (MRM), for both men and women, and thus they have been ruled out.

Spanish male literacy rate), and then taking the indices 160, 130, 100 and 70 as dividers between clusters[31]. The following five clusters are obtained (see Figure 1[32]):

- *Castilian core*, with male literacy rates above 65%. It is made up of 6 provinces, geographically contiguous, roughly corresponding to the County of Castile becoming a (more or less) independent entity in the mid-tenth century. By 1877, five of these provinces (Álava, Burgos, Palencia, Santander and Soria) had surpassed the threshold of 75% male literacy (the sixth province, Segovia, reached 72.1% in that census). It is a predominantly rural area, with a prevalence of small villages and no town with more than 20,000 inhabitants, except for Burgos with 25,000 and Santander with 30,000.
- *Northern Plateau*, with male literacy rates between 50% and 62%. It is made up of 8 provinces located in the Northern Plateau (*Meseta Norte*) or on its edge, around the Castilian core. Here is situated Madrid, the capital and largest city of Spain, with 300,000 inhabitants, but also the very rural region of León (provinces of León, Zamora and Salamanca; the university city of Salamanca has only 16,000 inhabitants).
- *Sundry North*, with male literacy rates between 40% and 49%. It is the only cluster without geographical unity, made up of 6 provinces located at different points in the northern half of Spain. Here is situated Barcelona, the second city in Spain, with almost 200,000 inhabitants. It is the most linguistically diverse cluster, including areas with a predominance of the Spanish, Catalan, Basque or Galician languages.
- *Transition*, with male literacy rates between 28% and 39%. It is made up of 17 provinces, which (except for three of them) constitute a continuous swathe of land from the eastern Pyrenees to the southwest coast in the Atlantic, separating the first three clusters from the fifth. Here is situated Seville, the third Spanish city, with almost 120,000 inhabitants, and three other cities with more than 50,000 inhabitants.
- *South and East*, with male literacy rates below 27%. It is made up of the Balearic and Canary Islands and 10 provinces that constitute a continuous strip along the Mediterranean coast, except for the northern part of it. Here you can find some of the richest agricultural areas and six of the twelve Spanish cities with more than 50,000 inhabitants, including Valencia, the fourth Spanish city, with almost 110,000 inhabitants.

Our clustering procedure is straightforward, dealing with attribute similarity, where the attribute is one-dimensional and naturally ordered. The resulting clusters turn out to be spatially contiguous to a high degree. Moreover, the spatial distribution of literacy shows a mostly concentric pattern with the province of Burgos at the centre. The whole first cluster is included in the first ring formed by Burgos and the bordering provinces. If we consider the 19 provinces of the second ring (the first ring and its bordering provinces), it turns out that it coincides with the 19 provinces with the highest male literacy rate (with only two exceptions).

The spatial pattern of the male literacy rate in 1860 is remarkable. There is no positive correlation between male literacy and the level of urbanization (measured by the percentage of the population living in the provincial capital or in towns with more than 30,000 inhabitants): the correlation coefficient is $\rho = -0.0096$ (we always consider disaggregation by province). There is also no appreciable correlation between male literacy and the level of industrialization, whether the latter is measured

[31] In this case, our procedure is equivalent to clustering by shading (see Johnson and Wichern 1998). The number of clusters have been determined by inspection of the distribution of the indices.

[32] The present official names of the provinces are used in the maps: "Oviedo" is now "Asturias", "Santander" is "Cantabria", "Logroño" is "La Rioja", "Guipúzcoa" is "Gipuzkoa", "Gerona" is "Girona", "Lérida" is "Lleida", "Coruña" is "A Coruña" and "Orense" is "Ourense".

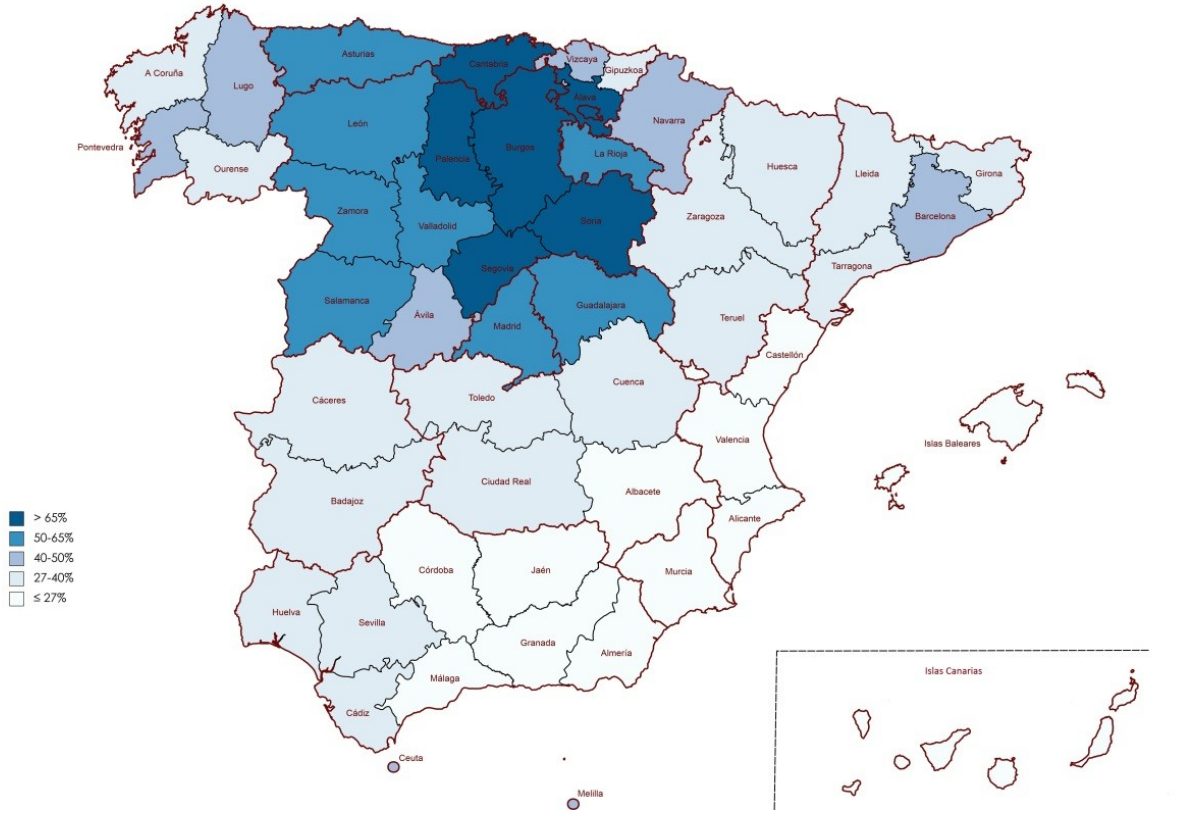


Figure 1: Literacy rates (males aged 10 and over) in 1860.

by the ratio of "workers in factories" ("jornaleros en las fábricas") to the total population, or if both miners and workers in factories are in the numerator: the correlation coefficients are $\rho = 0.0017$ and $\rho = -0.0784$, respectively.

Building explanatory models of literacy in 1860 would require the analysis of data much earlier than 1860. The purpose of this paper is rather to explain the *evolution* of literacy between 1860 and 1910, taking literacy rates in 1860 as initial conditions.

4.2 Evolution in the period 1860-1910

Literacy levels are the result of decisions made, on the one hand, by local authorities in the municipalities, where the provision of schools and teachers is established, and, on the other hand, by individuals, who decide whether they send their children to school, and for how long, or whether they try to become literate as adults. As in any decision problem, both types of decision makers, local authorities and individuals, have preferences and constraints.

The preferences of decision makers are part of their *mentality* and are marked above all by the value they give to education (see e.g. Reher (2023)).

The most important constraints are the economic ones, and particularly the financial resources available to the local councils, which are responsible for public primary education during this period. According to the *Moyano Law* (1857), the funding of primary education fell to the municipalities, which meant in practice a highly decentralized system. Irrespective of different historical contexts and levels of advancement in the literacy process, the Moyano Law shared the decentralized approach to the funding of public elementary education with the relevant laws in other Western European countries: the *Allgemeines Landrecht* (1794) of Prussia[33], the Guizot Law (1833) and the *Falloux Law* (1850) of France[34], the *Casati Law* of Italy (1859)[35].

In general, the resources available for primary education in Spain were conditioned throughout this period by the *desamortizaciones* of 1836-1837, 1841 and 1855 (see Bennassar (1985)), which confiscated the properties of the Catholic Church, the assets of educational foundations of all kinds, and most of the land belonging to the municipalities[36]. The impoverished Spanish local councils were saddled from 1857 with the obligation to support public primary education, in a context where educational charities were deprived of all their means.

Two literacy rates may have a *biological link*, in so far as they share part of their underlying populations[37]. Assuming that the acquisition of literacy for those older than a certain age becomes increasingly unlikely, a very low literacy rate in some year holds back the biologically linked literacy

[33]The previous Prussian *Generallandschulreglements* for Protestant schools (1763) and for Catholic schools (1765) were advanced for their time, but poorly enforced. In 1819/20 the (central) state participation in elementary education spending was 6.2 percent; it was 4.5 percent in 1861, when universal alphabetization of children had been reached already (see Zilch (2014)). Then the state participation began to grow (28.8% in 1911), especially when elementary education fees were eliminated in 1889.

[34]The Guizot Law required municipalities to establish elementary schools for boys, and the Falloux Law extended this obligation to schools for girls. See Diebolt et al. (2005) on the rise of mass schooling in France and its funding.

[35]See Cappelli and Quiroga (2020, 2021) and Bray (1991).

[36]The special public debt securities issued by the state as a partial compensation for the municipalities were nowhere near enough and became eventually worthless (the low yields were reduced even more and often not paid). See Moral Ruiz (1984) (p. 30-31, 106-107) and Comín (1996).

[37]For example, rate T_{10} in 1900 has a biological link with rate T_{10} in 1910, because the people 10 years old and over in 1900 still surviving in 1910 are part of the people 10 years and over in 1910. In contrast, T_{10} in 1900 has no biological link with T_{11-15} in 1910.

rates in later years[38]. In this connection, age-specific rates (whenever available) such as T_{11-15} may provide explanatory power.

There are three *res ipsa loquitur* features of Spanish male literacy in the period 1860-1910:
1. The global failure of the literacy process. The male literacy rate grew by only 18.7 percentage points over a 50-year period, from a comparatively modest 38.9% in 1860 to a comparatively low 57.6% in 1910. This slow growth occurred not only in the phase of political instability up to 1876 (the male literacy rate is still 43.5% in 1877), but also in the relatively stable span of the Bourbon Restoration (growth of 14.1 points in the 33 years between 1877 and 1910).
2. The large spatial differences. As we shall see later in more detail, the spatial structure of 1860 is maintained. The six provinces of the Castilian core are still the top six provinces by male literacy rate in 1910. The 19 provinces of the second ring around Burgos still coincide in 1910 with the 19 provinces with the highest male literacy rate (now with only one exception). All the 12 provinces of the fifth cluster are among the bottom 14 provinces by male literacy rate in 1910. The following table shows the evolution of male literacy in the five clusters[39] (for the last three censuses T_{10} is given in the top row and T_{11-15} in the bottom row; the rates higher than 75% are marked bold):

	1860	1877	1887	1900	1910
	Men T_{10}	Men T_{10}	Men T_{10}, T_{11-15}	Men T_{10}, T_{11-15}	Men T_{10}, T_{11-15}
Castilian Core	69.01	**76.63**	**80.84** **82.93**	**83.04** **82.43**	**87.93** **88.28**
Northern Plateau	56.57	64.54	69.39 66.94	74.58 71.17	**78.95** **76.41**
Sundry North	45.31	51.28	57.80 55.58	62.03 59.94	70.62 68.48
Transition	33.94	38.54	43.00 40.15	47.88 44.75	52.32 49.02
South and Est	23.68	25.67	30.18 24.22	34.38 28.89	38.91 31.56
SPAIN	38.90	43.51	48.18 44.33	52.69 49.11	57.55 53.43

Table 2: Spanish male literacy rates by cluster.

3. The appreciable percentage of men becoming literate after school age. Beginning in 1887, censuses provide data of literacy by age[40]. Despite the underlying trend of growing child literacy, in all censuses the maximum male literacy corresponds to the group of adults between 31 and 35 years old, with literacy rates approximately 10 points higher than those of boys between 11 and 15 years old[41]:

[38] For example, a very low T_{10} in 1900 makes impossible for T_{10} in 1910 to be very high.

[39] We can order the 1910 male literacy rates and construct 5 levels parallel to the 1860 clusters (i.e. the first 6 provinces would form level 1, as the first cluster has 6 elements, the next 8 would form level 2, as the second cluster has 8 elements, etc.). The 1910 level 1 provinces would be exactly those of the first 1860 cluster, and the changes between the other 1910 levels and their 1860 counterpart clusters would be minimal: 5 provinces would move up one level (all of them where Basque or Catalan-Valencian was spoken), and displace 5 others that would move down one level.

[40] The age intervals considered by the three censuses do not coincide, but a homogeneous series can be built with 5-year intervals for the ages between 11 and 50 years old. Gabriel (1997) conducts a similar study to ours for the ages between 6 and 30 years old, and in Gabriel (1998) 10-year intervals are considered for all ages from 11 years onwards. The conclusions are essentially in line with our own, both for males and females (see below).

[41] Cohort analysis filters out the trend of growing child literacy, even when it is still susceptible to other distortions: the cohorts decrease through mortality and change through migration, and both phenomena may affect the literate and the illiterate to different degrees (see Cipolla (1969)). At any rate, cohort analysis seems to confirm our conclusion. The cohort

Census	Men T_{11-15}	Men T_{16-20}	Men T_{21-25}	Men T_{26-30}	Men T_{31-35}	Men T_{36-40}	Men T_{41-45}	Men T_{46-50}	Men T_{51-60}	Men T_{61-70}
1887	44.33	49.99	52.06	51.40	54.31	49.82	51.27	47.16	45.65	42.33
1900	49.11	54.75	57.12	55.71	58.11	54.93	57.67	52.63	50.15	45.42
1910	53.43	59.89	62.40	61.07	63.28	59.64	61.70	57.16	56.23	50.01

Table 3: Spanish male literacy rates by age.[42]

As age increases, the significance of parents in the literacy process gives way to that of the concerned individual. The means to implement late literacy were varied. Village schools allowed the not-so-young to attend. Adult schools were segregated by sex[43] and in 1900 legislation was passed organizing night classes for workers, as an instrument to achieve "a solid knowledge leading to capable and intelligent workers and teachers, who contribute to the development and progress of the arts and industries of the country"[44]. It was also non-negligible the literacy work that the army carried out on recruits during their military service[45].

Figure 2 considers the male child literacy T_{11-15} at the end of the period (data in Appendix 2). This map shows how varying were the deeds of local authorities and parents in different areas of the country. It is a snapshot of the advancement of the modernization process in 1910, as far as male literacy is concerned. Besides, the correlation coefficient between T_{10} and T_{11-15} in 1910 is $\rho = 0.9782$.

Now we are to establish a model of the dynamics of male literacy in the period 1860-1910. Firstly, we must remark the high correlation between male literacy T_6 in 1860 and male child literacy T_{11-15} in 1910 (even though there is no biological link between the two rates): the correlation coefficient is $\rho = 0.8824$. The values of 1860 represent well the inherited historical substratum, particularly with regard to the mentality of the decision makers in each area of Spain. Secondly, during the period 1860-1910 Spain experienced social and economic changes, whose influence on the literacy process must be considered.

The linguistic factor may also be relevant in the literacy process. In several areas of Spain, the usual language of the majority of the population was not Spanish in 1860. Basque was spoken in three provinces (Vizcaya, Guipúzcoa and the north of the province of Navarra; by the mid-nineteenth century Basque was hardly used in Álava). Catalan-Valencian was spoken in 8 provinces (Barcelona, Gerona, Lérida, Tarragona, the Balearic Islands, Castellón and a large part of the provinces of Valencia and Alicante). Galician was spoken in four provinces (La Coruña, Pontevedra, Lugo and Orense). These languages have different degrees of closeness to Spanish. Basque is not even an Indo-European language. Catalan-Valencian is a Romance language and has a significant mutual intelligibility with Spanish in written form and has partial or low intelligibility in spoken form (the latter varies greatly according to dialect)[46]. Galician is very close to Spanish and both languages are mutually intelligible[47]. The Moyano Law of 1857, in force throughout the period, established the obligatory teaching

of those aged 11-15 in 1900 had a literacy rate, in that year, of 44.11, and in 1910 the members of that cohort recorded in the census (now aged 21-25) had a literacy rate of 62.40. This represents a variation of +18.29. An analogous calculation for the cohort aged 16-20 in 1900 gives a variation from 1900 to 1910 of +6.32. In turn, the variations from 1900 to 1910 for the 21-25, 26-30, 31-35 and 36-40 age cohorts in 1900 are, respectively: 6.16, 3.93, 3.59 and 2.23.

[42] The value of T_{15-15} is only available in 1887, and it is 46.96.

[43] Data of 1880 indicate that 91% of adult schools were for males and 94% of the students were male altogether (see Dirección General de Instrucción Pública (1883)).

[44] *Real Decreto* of 25 May 1900.

[45] See Quiroga (1999) and Dirección General del Instituto Geográfico y Estadístico (1914) (p. 366).

[46] See Juge (2007) for a general study of the distance of Catalan-Valencian to the other Romance languages.

[47] See Ramallo (2007) (p. 28); also Regueira (1999) about the closeness to Spanish of the different varieties of Galician.

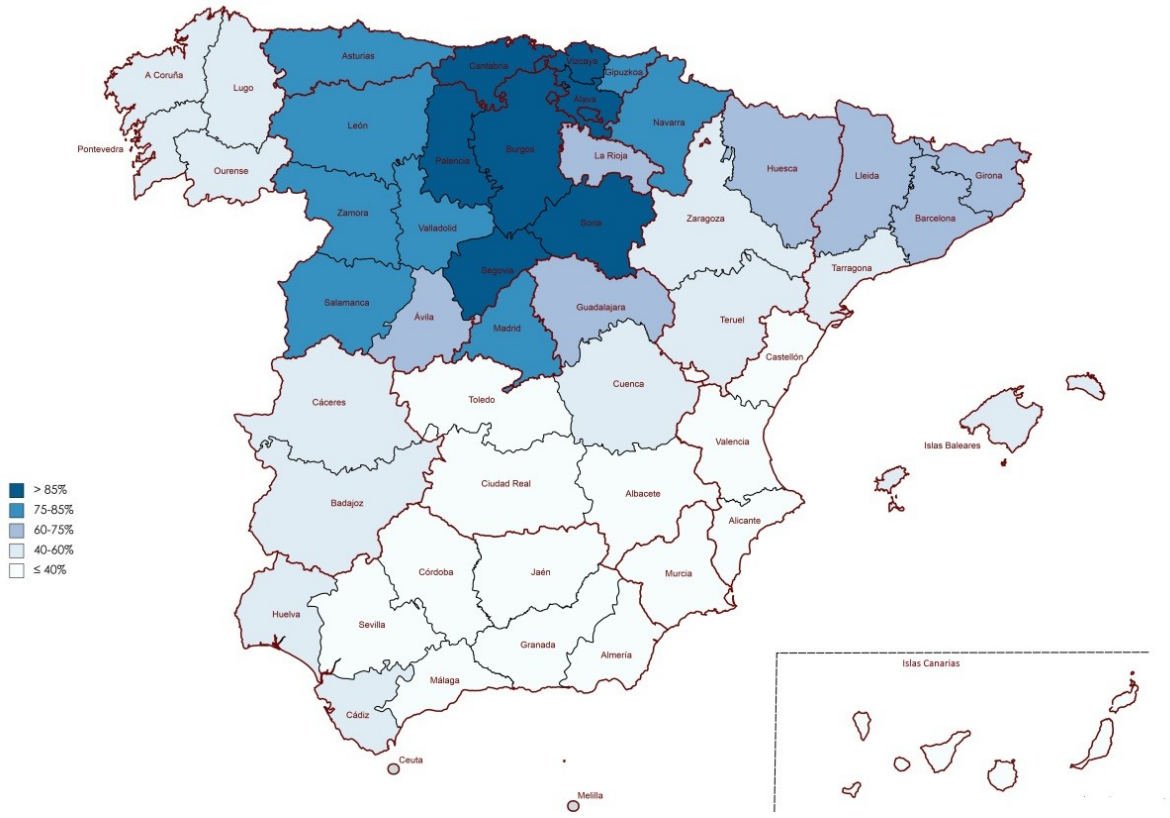


Figure 2: Literacy rates (males aged 11-15) in 1910.

of Spanish, but did not require it to be the compulsory language of instruction. In fact, it was not, in whole or in part, in quite a few municipalities of Spain, particularly in Catalonia[48]. In 1902, Spanish was imposed as the sole language of instruction by a decree[49], which was to be made ineffective in practice by a ministerial decree one month later[50].

The economic transformations, especially the industrialisation process, may affect literacy. Globally, the proportion of the agricultural active population to the total active population stayed at around 72% during the period (among men; it is difficult to estimate the composition of the female active population in predominantly agrarian economies)[51]. However, the process of establishing a modern industry advanced, with a marked tendency towards spatial concentration. Barcelona and Vizcaya became important industrial hubs[52].

We propose the following regression model, with the variables running through the values of the provinces:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

where Y is the male child literacy rate T_{11-15} in 1910 and ϵ is the error term. The regressors are:

- X_1: male literacy rate T_{10} in 1860.
- X_2, X_3, and X_4: dummy dichotomic variables, taking the value 1 if Basque (for X_2), Catalan-Valencian (for X_3), or Galician (for X_4) is spoken in the province, and the value 0 otherwise.
- X_5: proportion of the male active population working in agriculture (including forestry and fishing) in 1900[53].

Table 4 presents the estimates of four models (the corresponding p-values are shown in parentheses). The first model is the initial one, with 5 regressors. The values of the coefficient of determination R^2 and the adjusted coefficient of determination $\overline{R}^2$ are high. The regressors X_4 and X_5 may be successively removed as not at all significant. In the resulting second model, the coefficient of determination is hardly altered in relation to the first model, and all the regressors are significant.

The estimated model we arrive at is

$$Y = 0.0034 + 1.2929 X_1 + 0.2412 X_2 + 0.1436 X_3 + \epsilon$$

whose coefficient of determination is round 90%. On the other hand, the simple regression model

$$Y = 0.0811 + 1.1959 X_1 + \epsilon$$

has coefficient of determination $R^2 = 0.7786$. Consequently, male child literacy in 1910 is to a large extend explained by male literacy 50 years earlier. Basque and Catalan-Valencian being spoken are also two significant variables, where the former is more influential.

[48] See Diario de las sesiones de Cortes (1896, 1902); González Ollé (1985); Gabriel (2019).

[49] Decree of 21/11/1902. The incumbent minister was the Count of Romanones. Shortly after the publication of the decree, Sagasta's Liberal government fell and was replaced by Silvela's Conservative government.

[50] Ministerial decree of 19/12/1902, signed by the new Minister of Public Instruction in Silvela's government, Manuel Allendesalazar. Part of Romanones's decree was declared void as being *contra legem*, and the rest was reinterpreted so as to make it unenforceable in the relevant cases.

[51] The available rates are 72.14% for 1877, 72.26% for 1887, 72.20% for 1900 and 71.64% for 1910 (see the corresponding censuses and Nicolau (2005)).

[52] The proportion of the male active population working in industry in 1900 was 31.86% in Barcelona and 37.06% in Vizcaya. On the other hand, this rate was only 6.74% in Burgos, the province with the highest male literacy in Spain.

[53] An alternative model has also been considered, with X_5 standing for the proportion of the male active population working in industry (including mining, energy and construction) in 1900. There is no appreciable alteration in the results.

Regressors	β_0 intercept	β_1 M. 1860	β_2 Basque	β_3 Cat-V.	β_4 Galician	β_5 agric.	R^2	$\overline{R}^2$
X_1, X_2, X_3, X_4, X_5	0.0138 (0.883)	1.2915 (6E-22)	0.2387 (9E-6)	0.1423 (4E-5)	-0.0013 (0.974)	-0.0125 (0.909)	0.8979	0.8860
X_1, X_2, X_3	0.0034 (0.914)	1.2929 (4E-23)	0.2412 (7E-7)	0.1436 (9E-6)			0.8978	0.8910
X_1, X_2	0.0715 (0.046)	1.1867 (3E-19)	0.2181 (1E-4)				0.8407	0.8338
X_1	0.0811 (0.051)	1.1959 (5E-17)					0.7786	

Table 4: Regression models for male literacy.

Some explanations of the significance of these linguistic variables can be hypothesized. On the one hand, learning Spanish was perceived as increasingly important by local authorities and parents with the process of economic modernization, especially in the case of Basque; this learning went hand in hand with the acquisition of literacy. On the other hand, higher growth of literacy during the period for those not having Spanish as mother tongue might also be attributable to catching up from a relatively low level of literacy before 1860, caeteris paribus, even if the mother tongue was partially or totally the language of instruction. At any rate, further study of the influence of linguistic variables suggests itself, considering local data (e.g. on the language of instruction) and events such as the Carlist wars and the ensuing settlements. In contrast to political entities like Cisleithania (see Urbanitsch (2021)), France (see Furet and Ozouf (1977)) or Prussia (see Belzyt (1998)), not much attention has been paid to the influence of linguistic factors on the Spanish literacy process.

5 The evolution of female literacy

5.1 The starting point

Based on census data, we can estimate that the Spanish literacy rate for females (aged 10 and over) was 11.2% in 1860. Figure 3 displays the spatial distribution of female literacy in that year.

There is not as clear a spatial pattern for female literacy in 1860 as for male literacy. There are only 4 provinces above 20%. Half of the provinces have literacy rates that are concentrated between 7.1% and 11.1%. Nine provinces are below 7%, including the four provinces of Galicia.

Emigration to America is arguably the main reason for the low sex ratios (ratio of male to female for the population between 16 and 50 years old) in the Canary Islands and the provinces of the northwestern coast of Spain[54]. Here Galicia is included, but also Oviedo, Santander and Vizcaya[55]; the latter two provinces have relatively high female literacy rates (considering the dismal

[54]See Gozálvez Pérez and Martín-Serrano Rodríguez (2016). Internal migrations were mostly within the bounds of each province during the period, with exception of the immigration to Madrid and Barcelona (see Nicolau (2005) and Juif and Quiroga (2019)).

[55]The lowest sex ratios in 1860 were the following: Pontevedra 65.76%, Oviedo 71.29%, Coruña 71.53%, Canarias 71.83%, Lugo 78.43%, Santander 79.36%, Orense 82.37%, Vizcaya 85.87%; the national average was 95.69%. In the 1910 census the six lowest ratios were as follows: Pontevedra 59.95%, Coruña 64.65%, Canarias 72.03%, Oviedo 76.45%, Orense 76.66%, Lugo 76.70%; the ratio of Santander was 83.36% and that of Vizcaya 92.44%; the national average was 91.69%.

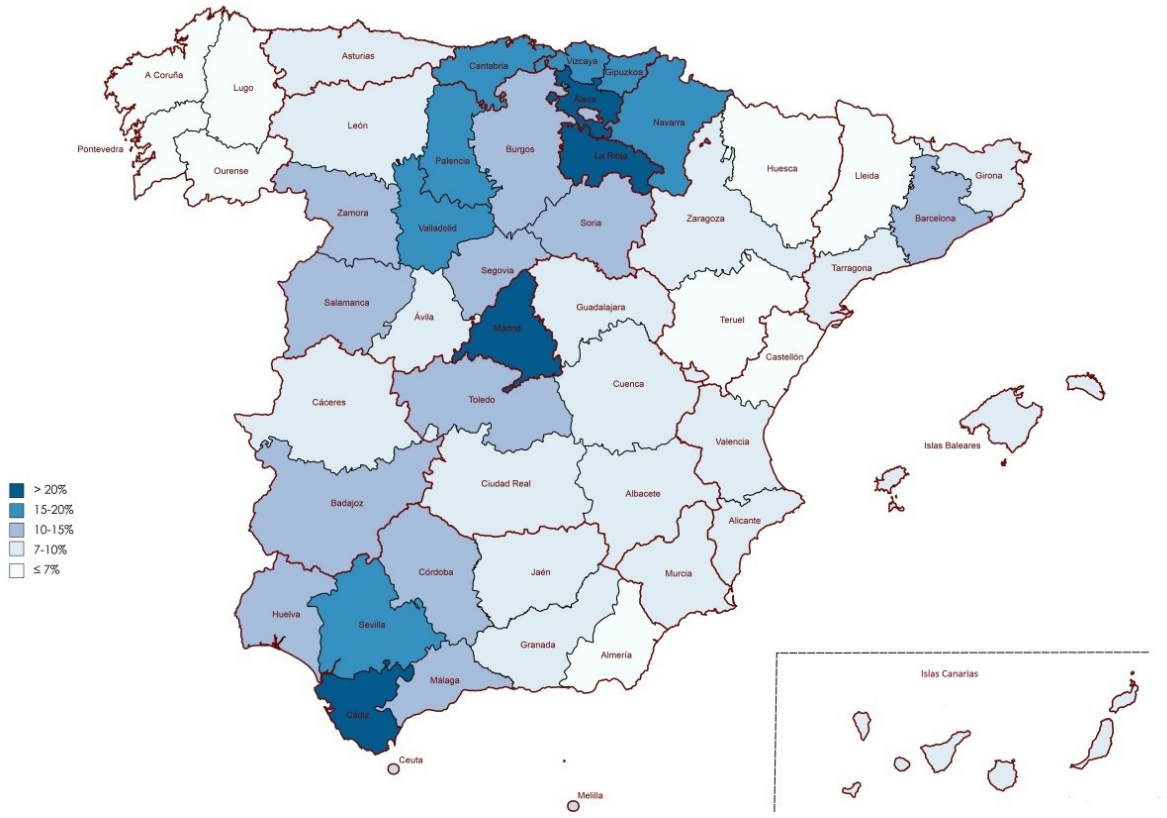


Figure 3: Literacy rates (females aged 10 and over) in 1860.

background)[56]. At any rate, low sex ratios tend to impair the position of women, although other factors may countervail the effect[57].

Contrary to what might be expected, there is not a high correlation between the male and female literacy rates in 1860, with $\rho = 0.5377$, $\rho^2 = 0.2891$ (we always consider disaggregation by province), a figure similar to that of the correlation between the female literacy rate and the urbanization rate ($\rho = 0.5444$, $\rho^2 = 0.2963$). The (squared) multiple correlation coefficient for the female literacy rate with respect to the male literacy rate and the urbanization rate is 0.5912.

There is hardly any correlation between female literacy and the level of industrialization, whether the latter is measured by the ratio of industrial workers to the total population ($\rho = 0.1970$), or if both miners and industrial workers are in the numerator ($\rho = 0.1490$).

5.2 Evolution in the period 1860-1910

In parallel to subsection 4.2, we consider these three features of the Spanish female literacy in the period 1860-1910:

1. As in the male case, the global failure of the literacy process. The female literacy rate grew by only 27.4 percentage points over a 50-year period, from a comparatively low 11.2% in 1860 to a comparatively very low 38.6% in 1910. This slow growth occurred not only in the phase of political instability up to 1876 (the female literacy rate is still 17.9% in 1877), but also in the relatively stable span of the Bourbon Restoration (growth of 20.7 points in the 33 years between 1877 and 1910).
2. The spatial distribution of female literacy gradually aproaches along the period the pattern of male literacy in 1860. On the one hand, the correlation coefficients of female literacy (in 1877, 1887, 1900 and 1910) with female literacy in 1860 decrease continuously along the period (in 1877 is $\rho = 0.9634$, in 1910 is $\rho = 0.7369$). This is not surprising. In contrast, the correlation coefficients of female literacy (in 1860, 1877, 1887, 1900 and 1910) with male literacy in 1860 increase steadily along the period (in 1860 is $\rho = 0.5377$, in 1877 is $\rho = 0.6543$, in 1910 is $\rho = 0.8134$). All in all, the spatial distributions of female and male literacy become closer over time ($\rho = 0.5377$ in 1860, $\rho = 0.8815$ in 1910).
   The clusters introduced in Section 4 continue to help us now. The following table, parallel to Table 2, shows the evolution of female literacy in the five clusters:

[56] Apart from the emigration to America, substantial in the aforementioned provinces, there was an important emigration from Almería to Argelia (the sex ratio of Almeria was 87.38% in 1860 and 79.41% in 1910). It was predominantly a temporary migration, with most emigrants returning eventually (see Nicolau (2005)).

[57] Out-of-wedlock birth rates are to be interpreted very cautiously, and considering the society of 1860, but some figures are worth noticing (data of the *Anuario Estadístico de España 1860-1861*). Canarias (20.4%) and Lugo (18.4%) had the highest rates in the country, whereas the rates of Santander (4.0%), Vizcaya (2.4%) and Almería (3.4%) were below the national average (5.6%).

	1860	1877	1887	1900	1910
	Women T_{10}	Women T_{10}	Women T_{10}, T_{11-15}	Women T_{10}, T_{11-15}	Women T_{10}, T_{11-15}
Castilian Core	15.92	27.81	37.47 57.91	51.42 67.42	66.80 **82.06**
Northern Plateau	16.03	25.55	32.18 42.52	43.32 52.66	54.05 65.69
Sundry North	10.86	19.54	26.24 36.75	35.14 46.13	45.63 58.12
Transition	10.01	15.85	20.14 26.32	26.47 34.02	33.64 41.59
South and Est	8.55	12.37	15.30 17.18	20.36 23.17	25.20 26.10
SPAIN	11.14	17.86	22.84 29.58	30.54 38.07	38.55 45.90

Table 5: Spanish female literacy rates by cluster.

As an exception to the dismal global evolution during the period, the female literacy process was successful in the Castilian core. There was also some closing of the gap with male literacy in the North of Spain, especially in the Northern Plateau. Certainly, the low initial female literacy T_{10} rates "burden" the later T_{10} rates through biological link, but not the later T_{11-15} rates; it is thus disappointing that T_{11-15} is almost as bad as T_{10} in the South and East cluster in 1910.
A measure of the level of improvement of female literacy during the period is given by the difference between T_{11-15} in 1910 and T_{10} in 1860: 66.14 percentage points for the Castilian core, 49.67 for the Northern Plateau, 47.26 for Sundry North, 31.57 for Transition and 17.56 for South and East, with 34.75 for Spain overall.

3. Less women than men became literate after (extended) school age. In all censuses the maximum female literacy corresponds to the interval of those between 16 and 20 years old; in this interval literacy rates are little different from the rates of girls between 11 and 15 years old[58] (these results are to be interpreted considering the trend of growing child literacy)[59]:

Census	W. T_{11-15}	W. T_{16-20}	W. T_{21-25}	W. T_{26-30}	W. T_{31-35}	W. T_{36-40}	W. T_{41-45}	W. T_{46-50}	W. T_{51-60}	W. T_{61-70}
1887	29.58	30.73	29.10	25.91	25.50	21.01	19.56	15.96	13.82	11.84
1900	38.07	39.77	38.17	33.88	33.42	28.80	28.71	24.13	19.98	16.10
1910	45.90	48.40	47.31	42.78	42.17	37.11	37.10	31.65	28.36	21.86

Table 6: Spanish female literacy rates by age.

Figure 4 displays the female child literacy T_{11-15} in 1910 (data in Appendix 2)[60]. This map shows the spatial distribution of a most relevant indicator of cultural modernization. Out of

[58] In 1887, disaggregated data for the 16-20 interval are available, and thus $T_{16-16} = 30.78$, $T_{17-17} = 32.13$, $T_{18-18} = 30.52$, $T_{19-19} = 31.96$ and $T_{20-20} = 28.79$.

[59] Cohort analysis seems to confirm our conclusion, comparing with the corresponding data for men (see above). The cohort of those aged $11-15$ in 1900 had a literacy rate, in that year, of 38.07, and in 1910 the members of that cohort recorded in the census (now aged $21-25$) had a literacy rate of 47.31. This represents a variation of $+9.24$. An analogous calculation for the cohort aged 16-20 in 1900 gives a variation from 1900 to 1910 of $+3.01$. In turn, the variations from 1900 to 1910 for the $21-25$, $26-30$, $31-35$ and $36-40$ age cohorts in 1900 are, respectively: 4.00, 3.23, 3.68 and 2.85.

[60] The correlation coefficient between T_{10} and T_{11-15} in 1910 is $\rho = 0.9543$.

49 provinces, there are 9 provinces above the threshold of 75%, 15 provinces between 50% and 70%, and 25 provinces under 44%.

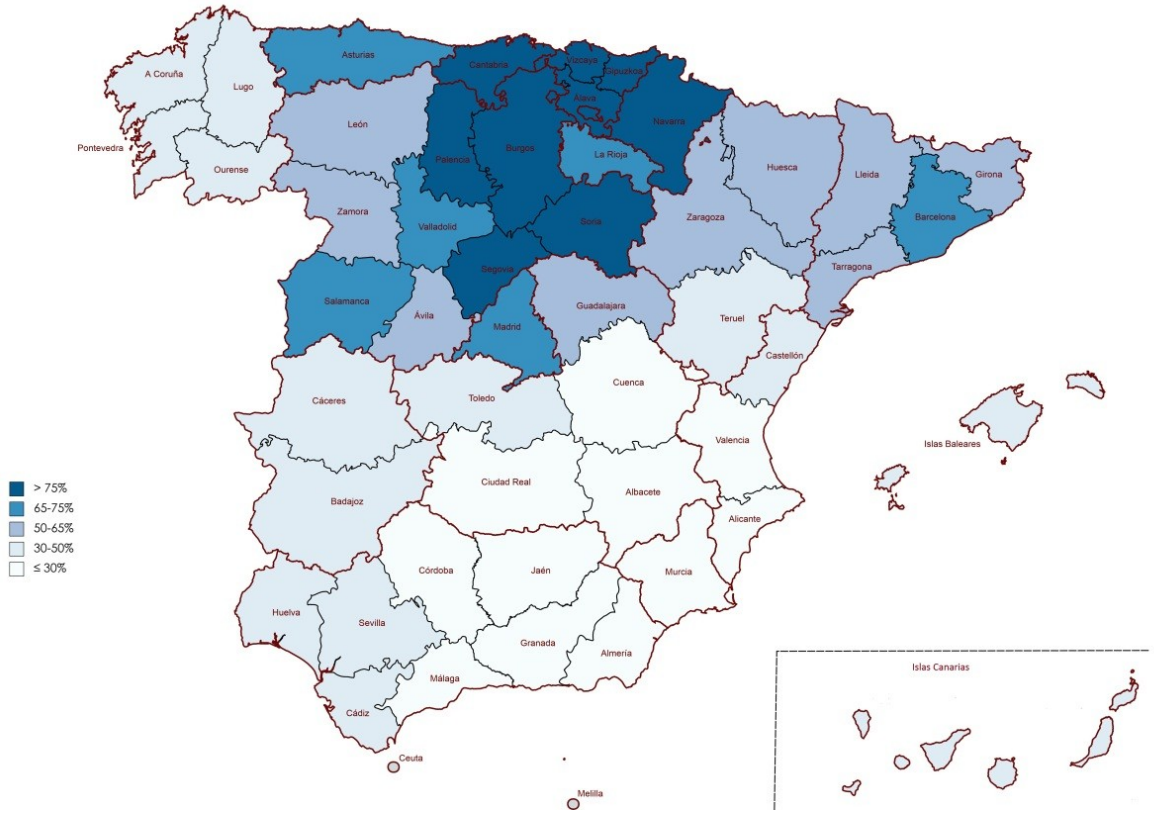


Figure 4: Literacy rates (females aged 11-15) in 1910.

Next, we establish a model of the dynamics of female literacy in the period 1860-1910. The same regressors as for male literacy (in Section 4) are considered, with the addition of a further regressor representing female literacy in 1860.
We propose the following regression model, with the variables running through the values of the provinces:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$$

where Y is now the female child literacy rate T_{11-15} in 1910, ϵ is the error term and X_1, X_2, X_3, X_4, X_5 are as in Section 4. We incorporate the regressor:

$$X_6 : \text{ female literacy rate } T_{10} \text{ in 1860.}$$

Table 7 presents the estimates of six models (the corresponding p-values are shown in parentheses). The first model is the initial one, with 6 regressors[61]. The values of the coefficient of

[61] An alternative initial model has also been considered, with X_5 standing for the proportion of the male active population working in industry (including mining, energy and construction) in 1900. There is no appreciable alteration in the results.

determination R^2 and the adjusted coefficient of determination $\overline{R}^2$ are high. The regressors X_6 and X_5 may be successively removed as non-significant. It is remarkable the non-significance of X_6, in view of the very strong significance of X_1. In the resulting model with 4 regressors, the coefficient of determination is little altered in relation to the first model. All the regressors are strongly significant, except for X_4 (Galician), whose p-value is between 0.01 and 0.05; on the other hand, the estimated coefficient of X_4 is negative, in contrast to the other linguistic regressors.

Regressors	β_0 intercept	β_1 M. 1860	β_2 Basque	β_3 Cat-V.	β_4 Galician	β_5 agric.	β_6 F. 1860	R^2	$\overline{R}^2$
$X_1, X_2, X_3, X_4, X_5, X_6$	0.2620 (0.13)	1.2651 (7E-16))	0.2602 (9E-6)	0.0982 (0.014)	-0.1063 (0.026)	-0.3361 (0.086)	-0.4891 (0.241)	0.8815	0.8646
X_1, X_2, X_3, X_4, X_5	0.0963 (0.347)	1.1872 (3E-19)	0.2647 (6E-6)	0.1207 (8E-4)	-0.0831 (0.054)	-0.1562 (0.190)		0.8775	0.8633
X_1, X_2, X_3, X_4	-0.0301 (0.400)	1.2029 (9E-20)	0.2936 (2E-7)	0.1334 (2E-4)	-0.0954 (0.026)			0.8725	0.8609
X_1, X_2, X_3	-0.0434 (0.241)	1.2105 (2E-19)	0.3037 (2E-7)	0.1444 (9E-5)				0.8570	0.8475
X_1, X_2	0.0251 (0.521)	1.1037 (3E-16)	0.2805 (1E-5)					0.7980	0.7892
X_1	0.0374 (0.432)	1.1155 (2E-13)						0.6929	

Table 7: Regression models for female literacy.

The estimated model we arrive at is

$$Y = -0.0301 + 1.2029X_1 + 0.2936X_2 + 0.1334X_3 - 0.0954X_4 + \epsilon$$

whose coefficient of determination is round 87%. On the other hand, the simple regression model

$$Y = 0.0374 + 1.1155X_1 + \epsilon$$

has coefficient of determination round 70%. Consequently, female child literacy in 1910 is to a considerable extend explained by male literacy 50 years earlier. The values of male literacy of 1860 represent well the inherited historical substratum. Catalan-Valencian being spoken is also a relevant variable, and Basque being spoken is very relevant.

6 Conclusions: towards an explanatory model of the evolution of Spanish literacy in the period 1860-1910.

During the second half of the 19th century, Spain was characterized by slow progress in literacy rates, with much worse results than other nearby European countries such as Italy, which had overcome Spain in this regard by 1880. A partial explanation for this poor performance is that the (central) state confiscated most of the land belonging to the municipalities by the *desamortización* of 1855. Yet the funding of public primary education was assigned to these municipalities by the Moyano Law of 1857. The lack of resources of Spanish local councils meant that aggregate spending on primary

education remained stagnant, while the school population grew. In contrast, Italy, which also had a decentralized system, increased the spending significantly[62].

As for male literacy, despite the slow progress in the country as a whole, there were great spatial disparities that remained essentially stable throughout the period, only with exceptions related to minority languages. In general terms, during these fifty years, provinces increased their rates by around 20 percentage points and thus those starting at the first positions reached almost complete universal male literacy by 1910, whereas some backward provinces ended up below 40% male literacy, the minimum threshold for sustained economic development to begin. The 6 top provinces in 1860 (still heading the list in 1910) coincided with the "Castilian core" of the country. Some explanations of their high male literacy values can be traced back to historical processes of medieval origin related to the *Reconquista*, including the model of land distribution and mental processes of imitation of the behaviour of the nobility and the clergy as means of social advancement, involving a high social valuation of education (see Pérez Moreda (1997)); at any rate these provinces tended to be those devoting more public resources to primary education.

Figure 5 shows the per capita public expenditure on primary education at the end of the period (1908). Certainly, the commitment of local authorities to popular education remained uneven across the country[63]. Rich Madrid and Barcelona were not at the top, but rural high literacy provinces.

In addition to this geographic disparity, Spain had large differences between male and female literacy rates. In 1860, just over 10% of women over the age of 10 could read and write compared to almost 40% of men. A partial explanation of this difference might be economic. The return on investment in education depended on its degree of use: if parents anticipated a different participation in the labour market according to sex, they would allocate family resources (including time) in a biased way; therefore, the family unit would be maximizing the return on its investment by having boys more educated than girls. All the same, the literacy gender gap was very large in comparative West European terms (see Section 2), especially in the more literate provinces, where it could exceed 50 percentage points (obviously, it was smaller in the little literate provinces). The abysmal literacy gender gap in the most literate Spain at the beginning of the period remains a puzzle, only to be addressed by analysing data prior to 1860.

Although the improvement of female literacy was globally less than mediocre during the period, female literacy somewhat reduced distances with male literacy, certainly in an uneven way. The literacy gender gap remained stable in the low literacy provinces. In contrast, female literacy grew steeply in the high literacy provinces, and so in 1910 girls were almost universally literate in the Castilian core. Statistical data show a lack of correlation between female literacy rates in 1910 and 1860 but a very high correlation between female literacy rates in 1910 and male literacy rates in 1860. At last literate fathers were willing to assume the economic cost of providing literacy to their daughters.

Certainly, the institutional framework eased the efforts of the overburdened small villages of rural Spain to fund primary education. The Moyano Law allowed the existence of mixed-sex "incomplete schools" and "seasonal schools" in villages with less than 500 inhabitants, and of incomplete schools for girls in villages with less than 2000 inhabitants. These schools had fewer subjects and less paid (and qualified) teachers, thus lowering the costs in education and making cheaper the extension of literacy to girls in small municipalities[65]. On the other hand, in small villages the school routine

[62] Cappelli and Quiroga (2020), Figure 3.

[63] Franchise in local elections was mostly selective (property based) until 1890 and universal (for males) afterwards. Suffrage was already universal after 1868 in villages under 100 inhabitants.

[64] Source: Dirección General del Instituto Geográfico y Estadístico (1913).

[65] Prima facie at the expense of quality, but the positive action of the bandwagon effect cannot be overlooked.

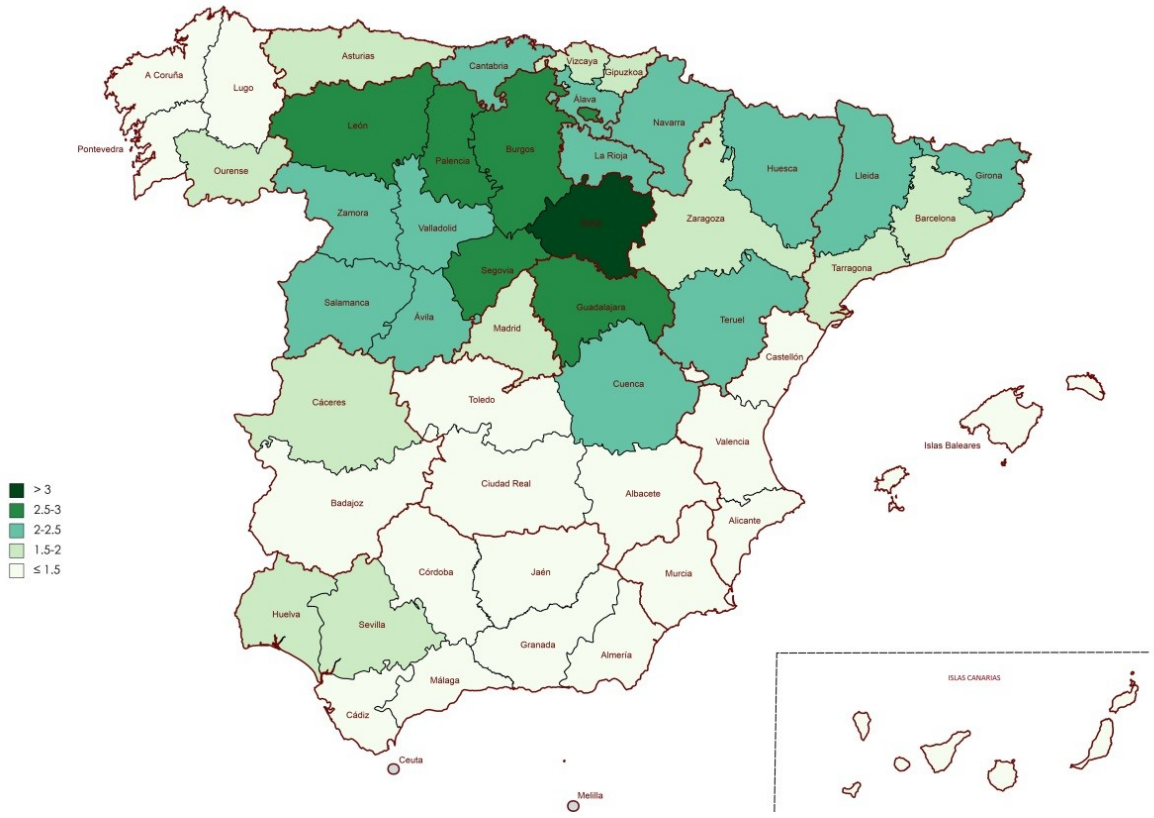


Figure 5: Per capita public expenditure in elementary education in 1908 (current pesetas).[64]

could be adapted to the seasonal work of children in agriculture, which was seen as an advantage by families; this was even more the case with seasonal schools[66]. As an example, in 1860, more than half of the population of Soria, one of the provinces with the highest literacy both for men in 1860 and for girls in 1910, lived in municipalities with less than 500 inhabitants, and over 70% of the schools were mixed-sex all along the period 1860-1910. At any rate, the local councils of very rural Soria were the top investors in public primary education in the entire country (data of 1908)[67].

The per capita public expenditure in elementary education in 1908 and the female child literacy T_{11-15} in 1910 are highly correlated: the correlation coefficient is $\rho = 0.7515$[68]. Figure 6 displays the corresponding scatter plot and regression line.

According to the legislation in force during the period, public primary education was free only for the certified poor. At any rate, the literacy gender gap was not only attributable to parents and local councils. As shown in Table 3 and Table 6, the maximum literacy of women was reached in the interval between 16 and 20 years old, whereas the peak in male literacy was in the interval between 31 and 35 years old. The requirements of the labour market provided inducements for men (more than for women) to acquire literacy (e.g., through adult schools, schools for workers or while doing military service). Available technology, economic backwardness and blatant prejudice limited the quantity and quality of jobs accessible to women.

It is to be pointed out that birth rates and literacy rates were uncorrelated in Spain, both in 1860 and 1910. Moreover, there was no correlation between birth rates and literacy rates for children (data of 1910)[69]. There was a positive correlation between literacy rates and the intensity of (Catholic) confessional allegiance (see Gutiérrez (2025)).

The evolution of literacy in Spain between 1860 and 1910 did not follow the spatial pattern of the economic modernization process. The high literacy rates of the Castilian core did not correspond to the relative income levels of the area at that time[70]. The most literate provinces were neither the most industrialized, nor the most urbanized[71]. The majority of them were agrarian provinces, with a predominance of small and medium property, and with the population living in small villages.

All in all, the dynamics of literacy during the period are predominantly explained by the initial *male* literacy (the initial female literacy turns out to be non-significant). To a lesser extent, Basque and Catalan-Valencian being spoken are also two significant variables, where the former is more influential. The corresponding parsimonious statistical models (only with three regressors) have

[66]Palencia, a province with high literacy, had the highest concentration of seasonal schools in Spain (see Núñez (1992) (p. 269)).

[67]Soria had an urbanization rate (percentage of the population living in the provincial capital or in towns with more than 30,000 inhabitants) of 4.75%, compared to 69.65% of Madrid or 50.54% of Barcelona (data of 1900). The financial effort of the municipalities of Soria has to be appreciated gauging the extent of the confiscation of municipal land by the *desamortización* of 1855 (see Marín Gutiérrez (2015)).

[68]The correlation for boys is a little higher than that for girls: the correlation coefficient between per capita public expenditure in elementary education in 1908 and the male child literacy T_{11-15} in 1910 is $\rho = 0.7839$.

[69]In 1860 the correlation coefficients were $\rho = -0.1321$, between birth rates and male literacy rates, and $\rho = 0.0915$, between birth rates and female literacy rates. In 1910 the correlation coefficients were $\rho = 0.0519$ and $\rho = 0.0027$, respectively; as for children, the correlation coefficients between birth rates and T_{11-15} were $\rho = -0.0544$ for boys and $\rho = -0.0659$ for girls. The sources for the birth dates are Junta General de Estadística (1861-1862) for 1860 (baptisms are taken instead of births), and Dirección General del Instituto Geográfico y Estadístico (1916) for 1910.

[70]There are no estimations of the GDP of this period disaggregated by province, but only by (present) region, and the latter estimations can be considered tentative. The majority of the provinces of the Castilian core and half of the provinces of the second cluster (Northern Plateau) are in the present region of Castilla y León. In 1900, the GDP per capita of Castilla y León was 91.2% of the Spanish average (see Álvarez Llano (1986) and Carreras et al. (2005)).

[71]In 1910 there was essentially no correlation between urbanization and literacy: the correlation coefficients were $\rho = -0.0030$, between urbanization rates and male literacy rates, and $\rho = 0.2065$, between urbanization rates and female literacy rates.

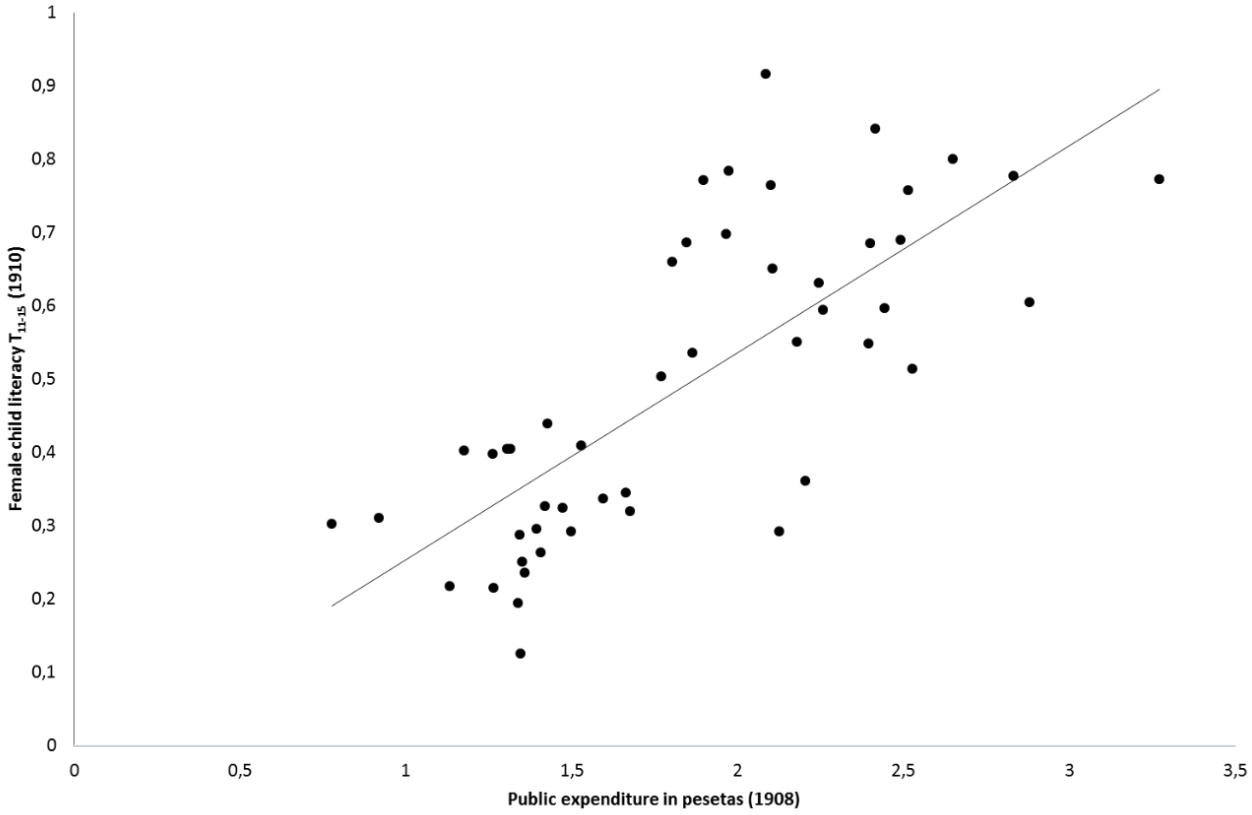


Figure 6: Public expenditure in elementary education in 1908 and female child literacy T_{11-15} in 1910.

coefficients of determination of round 90% for men and 86% for women. At any rate, further study, considering local data, of the influence of linguistic variables suggests itself.

References

Censuses

Junta General de Estadística: *Censo de la población de España, según el recuento verificado en 25 de diciembre de 1860 por la Junta General de Estadística*. Imprenta Nacional. Madrid (1863).

Dirección General del Instituto Geográfico y Estadístico: *Censo de la población de España, según el empadronamiento hecho en 31 de diciembre de 1877 por la Dirección General del Instituto Geográfico y Estadístico*. Imprenta de la Dirección General del Instituto Geográfico y Estadístico. Madrid (Tomo I, 1883; Tomo II, 1884).

Dirección General del Instituto Geográfico y Estadístico: *Censo de la población de España según el empadronamiento hecho en 31 de diciembre de 1887 por la Dirección General del Instituto Geográfico y Estadístico*. Imprenta de la Dirección General del Instituto Geográfico y Estadístico. Madrid (Tomo I, 1891; Tomo II, 1892).

Dirección General del Instituto Geográfico y Estadístico: *Censo de la población de España según el empadronamiento hecho en la península é islas adyacentes el 31 de diciembre de 1900*. Imprenta de la Dirección General del Instituto Geográfico y Estadístico. Madrid (Tomo I, 1902; Tomo II, 1903; Tomo III, 1907; Tomo IV, 1907).

Dirección General del Instituto Geográfico y Estadístico: *Censo de la población de España según el empadronamiento hecho en la península e islas adyacentes el 31 de diciembre de 1910*. Imprenta de la Dirección General del Instituto Geográfico y Estadístico. Madrid (Tomo I, 1913; Tomo II, 1916; Tomo III, 1917; Tomo IV, 1919).

Other statistical sources

Dirección General de Instrucción Pública: *Estadística General de Primera Enseñanza correspondiente al decenio que terminó el 31 de diciembre de 1880*. Imprenta y Fundición de M. Tello. Madrid (1883).

Dirección General del Instituto Geográfico y Estadístico: *Anuario Estadístico de España. Año 1912*. Imprenta de la Dirección General del Instituto Geográfico y Estadístico. Madrid (1913).

Dirección General del Instituto Geográfico y Estadístico: *Reseña Geográfica y Estadística de España. Tomo III*. Talleres del Instituto Geográfico y Estadístico. Madrid (1914).

Dirección General del Instituto Geográfico y Estadístico: *Movimiento Natural de la Población de España. Año 1910*. Talleres del Instituto Geográfico y Estadístico. Madrid (1916).

Junta General de Estadística: *Anuario Estadístico de España 1860–1861*. Imprenta Nacional. Madrid (1861-1862).

Further References

Álvarez-Llano, R. (1986): "Evolución de la estructura económica regional de España en la historia: una aproximación". *Situación*, 1, p. 5–61.

Baffour, B.; King, T.; Valente, P. (2013): "The Modern Census: Evolution, Examples and Evaluation". *International Statistical Review*, 81-3, p. 407–425.

Beltrán Tapia, F. J., Díez-Minguela, A., Martínez-Galarraga, J. & Tirado-Fabregat, D. A. (2021): "The uneven transition towards universal literacy in Spain, 1860–1930". *History of Education*, 50(5), p. 605–627.

Beltrán Tapia, F.J.; Martínez-Galarraga, J. (2018): "Inequality and education in pre-industrial economies: Evidence from Spain". *Explorations in Economic History*, 69, p. 81–101.

Bennassar, B. (1985): "Las resistencias mentales". In: B. Bennassar et al., *Orígenes del atraso económico español*, p. 147–163. Ariel. Barcelona.

Belzyt, L. (1998): Sprachliche Minderheiten im preußischen Staat 1815–1914. Die preußische Sprachenstatistik in Bearbeitung und Kommentar. Verlag Herder-Institut. Marburg.

Bowman, M.J.; Anderson, C.A. (1963): "Concerning the role of education in development". In: Geertz, C. (ed.), *Old Societies and New States*, p. 247–263. The Free Press. Glencoe, Illinois.

Bray, M. (1991): "Centralization versus decentralization in educational administration: regional issues". *Educational Policy*, 5, p. 371–385.

Cappelli, G.; Quiroga, G. (2020): "Literacy and schooling in Italy and Spain (1860–1921): a comparative analysis". *Rivista di storia economica*, XXXVI-1, p. 87–123.

Cappelli, G.; Quiroga, G. (2021): "Female teachers and the rise of primary education in Italy and Spain, 1861–1921: evidence from a new dataset". *Economic History Review*, 74-3, p. 754–783.

Cappelli, G.; Vasta, M. (2020): "Can school centralization foster human capital accumulation? A quasi-experiment from early twentieth-century Italy". *Economic History Review*, 73, p. 159–184.

Carreras, A.; Prados de la Escosura, L.; Rosés, J.R. (2005): "Renta y Riqueza". In: Carreras, A. and Tafunell, X. (eds.), *Estadísticas Históricas de España: Siglos XIX–XX*, 2nd edition, p. 1297–1376. Fundación BBVA. Bilbao.

Cipolla, C.M. (1969): *Literacy and Development in the West*. Penguin Books. Harmondsworth.

Comín, F. (1996): *Historia de la Hacienda Pública, II. España (1808–1995)*. Crítica. Barcelona.

Cusidó i Vallverdú, T.A.; Gil-Alonso, F. (2012): "Los censos en España: entre continuidad y cambio (1857–1970)". *Revista de Demografía Histórica*, 30-1, p. 29–67.

Diario de las sesiones de Cortes (1896): *Congreso de los Diputados. Sesión del viernes 19 de agosto de 1896*. Madrid.

Diario de las sesiones de Cortes (1902): *Congreso de los Diputados. Sesión del lunes 24 de noviembre de 1902*. Madrid.

Diebolt, C.; Jaoul, M.; San Martino, G. (2005): "Le mythe de Ferry: une analyse cliométrique". *Revue d'économie politique*, 115-4, p. 471–497.

Flora, P.; Alber, J.; Eichenberg, R.; Kohl, J.; Kraus, F.; Pfenning, W.; Seebohm, K. (1983): *State, Economy, and Society in Western Europe 1815–1975. A Data Handbook in Two Volumes. Volume I*. Campus Verlag. Frankfurt.

Furet, F.; Ozouf, J. (1977): *Lire et écrire*. Les Éditions de Minuit. Paris.

Furet, F.; Sachs, W. (1974): "La croissance de l'alphabétisation en France (XVIIIe–XIXe siècle)". *Annales. Économies, Sociétés, Civilisations*, 29-3, p. 714–737.

Gabriel, N. de (1997): "Alfabetización y escolarización en España (1897–1950)". *Revista de Educación*, 314, p. 217–243.

Gabriel, N. de (1998): "Literacy, Age, Period and Cohort in Spain (1900–1950)". *Paedagogica Historica*, 34-1, p. 29–62.

Gabriel, N. de (2019): "Nacionalismos y educación en España". *Historia de la Educación*, 37, p. 115–144.

Galor, O. (2011): *Unified Growth Theory*. Princeton University Press. Princeton.

Gomes, P.; Machado, M.P. (2020): "Literacy and primary school expansion in Portugal: 1940–62". *Revista de Historia Económica - Journal of Iberian and Latin American Economic History*, 38, p. 111–145.

González Ollé, F. (1985): "El uso de las lenguas regionales en la enseñanza primaria ante el Congreso de los Diputados (1896)". *Boletín de la Real Academia Española*, 65, p. 345–355.

Gozálvez Pérez, V.; Martín Serrano Rodríguez, G. (2016): "El censo de la población de España de 1860. Problemas metodológicos. Inicio de la aportación social en los censos". *Boletín de la Asociación de Geógrafos Españoles*, 70, p. 329–370.

Gutiérrez, J.M. (2024): "Census-based comparability of data on literacy processes in Western Europe". *Spanish Journal of Statistics*, 6, p. 39–59.

Gutiérrez, J.M. (2025): "A proxy variable built from a Kulturkampf". *Boletín de Estadística e Investigación Operativa*, 41, p. 47–53.

Hanushek, E.A.; Woessmann, L. (2010): "Education and economic growth". In: Baker, E.; McGaw, B. and Peterson, P. (eds.), *International Encyclopedia of Education*, p. 245–252. Elsevier. Oxford.

Johansson, E. (1977): *The History of Literacy in Sweden, in comparison with some other countries*. Umeå: Umeå University and School of Education, Educational Reports Umeå, 12. Partially reprinted as "The History of Literacy in Sweden", in H.J. Graff, A. Mackinnon, B. Sandin and I. Winchester (eds.), *Understanding Literacy in its Historical Contexts*, p. 28–57. Nordic Academic Press. Lund (2009).

Johnson, R.A.; Wichern, D.W. (1998): *Applied Multivariate Statistical Analysis*, 4th edition. Prentice Hall. Upper Saddle River, NJ.

Juge, M.L. (2007): "Catalan's Place in Romance Revisited". *Catalan Review*, 21, p. 257–277.

Juif, D.; Quiroga, G. (2019): "Do you have to be tall and educated to be a migrant? Evidence from Spanish recruitment records, 1890–1950". *Economics and Human Biology*, 34, p. 115–124.

Larquié, C. (1981): "L'alphabétisation à Madrid en 1650". *Revue d'Histoire Moderne et Contemporaine*, 28-1, p. 132–157.

Marín Gutiérrez, A. (2015): *La desamortización forestal en la provincia de Soria. La génesis de los "montes de socios"*. Diputación Provincial de Soria. Soria.

Melón, A. (1951): "Los censos de la población en España (1857–1940)". *Estudios Geográficos*, Vol. 12, No. 43, p. 203–281.

Mitch, D. (1993): "Educación y crecimiento económico: ¿Otro axioma de indispensabilidad? Del capital humano a las capacidades humanas". In: Núñez, C.E. and Tortella, G. (eds.), *La maldición divina. Ignorancia y atraso en perspectiva histórica*, p. 41–60. Alianza Editorial. Madrid.

Mitch, D. (2013): "The economic history of education". In: Parker, R.E. and Whaples, R.M. (eds.), *Routledge handbook of modern economic history*, p. 247–264. Routledge. London and New York.

Moral Ruiz, J. del (1984): *Hacienda central y haciendas locales en España, 1845-1905*. Instituto de Estudios de Administración Local. Madrid.

Myllyntaus, T. (1990): "Education in the Making of Modern Finland". In: Tortella, G. (ed.), *Education and Economic Development Since the Industrial Revolution*, p. 153–171. Generalitat Valenciana. Valencia.

Nicolau, R. (2005): "Población, salud y actividad". In: Carreras, A. and Tafunell, X. (eds.), *Estadísticas Históricas de España: Siglos XIX–XX*, 2nd edition, p. 77–154. Fundación BBVA. Bilbao.

Noble, F. (1965): "Istruzione scolastica". In: *Sviluppo della popolazione italiana dal 1861 al 1961*, Annali di Statistica, Serie VIII, Vol. 17, p. 295–317. Istituto Centrale di Statistica. Roma.

Nunes, A.B. (2003): "Government Expenditure on Education, Economic Growth and Long Waves: The Case of Portugal". *Paedagogica Historica*, 39-5, p. 559–581.

Núñez, C.E. (1992): *La fuente de la riqueza. Educación y desarrollo económico en la España contemporánea*. Alianza Editorial. Madrid.

Núñez, C.E. (1993): "Alfabetización y desarrollo económico en España: una visión a largo plazo". In: Núñez, C.E. and Tortella, G. (eds.), *La maldición divina. Ignorancia y atraso en perspectiva histórica*, p. 223–236. Alianza Editorial. Madrid.

Núñez, C.E. (1997): "La educación como fuente de crecimiento". *Papeles de Economía Española*, 73, p. 213–242.

Núñez, C.E. (2003a): "Literacy, Schooling and Economic Modernization: A Historian Approach". *Paedagogica Historica*, 39-5, p. 535–558.

Núñez, C.E. (2003b): "Within the European Periphery: education and labor mobility in twentieth-century Spain". *Paedagogica Historica*, 39-5, p. 621–649.

Núñez, C.E. (2010): "Sobre la escasez de capital social fijo y humano en la España contemporánea". In: Morilla, J. et al. (eds.), *Las claves del desarrollo económico y social. Ensayos en homenaje a Gabriel Tortella*, p. 241–270. LID Editorial - Universidad de Alcalá. Madrid.

Pérez Moreda, V. (1997): "El proceso de alfabetización y la formación de capital humano en España". *Papeles de Economía Española*, 73, p. 243–253.

Pleijt, A. de; Nuvolari, A.; Weisdorf, J. (2020): "Human capital formation during the first industrial revolution: evidence from the use of steam engines". *Journal of the European Economic Association*, 18-2, p. 829–889.

Quiroga, G. (1999): *Yo aprendí a leer en la Mili. El papel alfabetizador del ejército de tierra español, 1893-1954*. Ministerio de Defensa. Madrid.

Ramallo, F. (2007): "Sociolinguistics of Spanish in Galicia". *International Journal of the Sociology of Language*, 184, p. 21–36.

Regueira, X.L. (1999): "Estándar oral e variacion social da lingua galega". In: Álvarez, R. and Vilavedra, D. (eds.), *Cinguidos por unha arela común. Homenaxe ó profesor Xesús Alonso Montero*, 855–875. Universidade de Santiago. Santiago de Compostela.

Reher, D. (1997): "La teoría del capital humano y las realidades de la Historia". *Papeles de Economía Española*, 73, p. 254–261.

Reher, D.S. (2023): "Patterns of literacy in historic Spain: An interpretation". *Revista de Historia Económica / Journal of Iberian and Latin American Economic History*, 41-1, p. 83–117.

Reis, J. (1993): "El analfabetismo en Portugal en el siglo XIX: una interpretación". In: Núñez, C.E. and Tortella, G. (eds.), *La maldición divina. Ignorancia y atraso en perspectiva histórica*, p.237–269. Alianza Editorial. Madrid.

Rodríguez, M.C.; Bennassar, B. (1978): "Signatures et niveau culturel des témoins et accusés dans le procès d'Inquisition du ressort du Tribunal de Tolède (1525-1817) et du ressort du Tribunal de Cordoue (1595-1632)". *Cahiers du monde hispanique et lusobrésilienne*, 31, p. 17–46.

Sandberg, L.G. (1982): "Ignorance, Poverty and Economic Backwardness in the Early Stages of European Industrialization: Variations on Alexander Gerschenkron's Grand Theme". *Journal of European Economic History*, 11, p. 675–698.

Sarasúa, C. (2002): "El acceso de niñas y niños a los recursos educativos en la España rural del siglo XIX". In: Martínez Carrión, J.M. (ed.), *El nivel de vida en la España rural, siglos XVIII-XX*, p. 549–609. Universidad de Alicante. Alicante.

Sarasúa, C. (2019): "Women's work and structural change: occupational structure in eighteenth-century Spain". *The Economic History Review*, 72-2, p. 481–509.

Schofield, R. (1973): "Some dimensions of illiteracy, 1750-1850". *Explorations in Economic History*, 10-4, p. 437–454.

Smith, A. (1976 [1776]): *An Inquire into the Nature and Causes of the Wealth of Nations*, (2 vol.). Clarendon Press. Oxford.

Terrón Bañuelos, A. (1997): "La modernización de la educación en España (1900-1939)". In: Escolano Benito, A. and Fernandes, R. (eds.), *Los caminos hacia la modernidad educativa en España y Portugal (1800-1975)*, p. 101–121. Fundación Rei Afonso Henriques. Zamora.

Tortella, G. (1994): *El desarrollo de la España contemporánea. Historia económica de los siglos XIX y XX*. Alianza Editorial. Madrid.

Tveit, K. (1991): "The Development of Popular Literacy in the Nordic Countries. A Comparative Historical Study". *Scandinavian Journal of Educational Research*, 35-4, p. 241–252.

UNESCO (1953): *Progress of Literacy in Various Countries*. United Nations Educational, Scientific and Cultural Organization. Paris.

UNESCO (1957): *World Illiteracy at Mid-Century*. United Nations Educational, Scientific and Cultural Organization. Paris.

Urbanitsch, P. (2021): "Bildung und Bildungsinstitutionen zwischen Kulturförderung und Politik in Cisleithanien". In: Gottsmann, A. (ed.), *Die Habsburgermonarchie 1848-1918. Band X. Das kulturelle Leben. Akteure – Tendenzen – Ausprägungen*, p. 207–284. Verlag der Österreichischen Akademie der Wissensschaften. Wien.

Vilanova Ribas, M.; Moreno Julià, X. (1992): *Atlas de la evolución del analfabetismo en España de 1887 a 1981*. Ministerio de Educación y Ciencia. Madrid.

West, E.G. (1978): "Literacy and the Industrial Revolution". *The Economic History Review*, 31-3, p. 369–383.

Zamagni, V. (1993): "Instrucción y desarrollo económico en Italia, 1861-1913". In: Núñez, C.E. and Tortella, G. (eds.), *La maldición divina. Ignorancia y atraso en perspectiva histórica*, p. 181–222. Alianza Editorial. Madrid.

Zilch, R. (2014): *Finanzierung des Kulturstaats in Preußen seit 1800*, Acta Borussica, Reihe 2: Preußen als Kulturstaat. Abteilung II: Der preußische Kulturstaat in der politischen und sozialen Wirklichkeit. De Gruyter. Berlin.

Appendix 1 Literacy rates (male, female and total aged 10 years old and over) in Spain (1860, 1877, 1887, 1900 and 1910)

	1860	1877	1887	1900	1910	1860	1877	1887	1900	1910	1860	1877	1887	1900	1910
	Male T_{10} (estimation)	Male T_{10} (estimation)	Male T_{10}	Male T_{10}	Male T_{10}	Female T_{10} (estimation)	Female T_{10} (estimation)	Female T_{10}	Female T_{10}	Female T_{10}	Total T_{10} (estimation)	Total T_{10} (estimation)	Total T_{10}	Total T_{10}	Total T_{10}
Álava	69.54	76.27	80.19	84.36	87.20	26.28	38.82	50.22	62.95	70.53	48.74	57.93	65.32	73.71	78.79
Albacete	25.66	28.37	33.44	36.56	40.08	7.71	10.72	14.23	17.78	21.99	16.60	19.42	23.80	27.15	31.05
Alicante	20.70	25.40	30.57	35.36	42.25	7.11	11.71	15.13	20.30	27.32	13.76	18.31	22.64	27.61	34.48
Almería	21.05	24.13	24.55	32.33	35.17	5.70	9.04	8.83	17.16	19.91	12.99	16.09	16.31	24.41	26.85
Ávila	45.09	54.17	59.03	63.57	68.08	9.73	18.98	27.70	35.79	43.50	27.54	36.42	43.17	49.39	55.44
Badajoz	28.08	32.38	35.52	35.32	42.55	10.85	16.39	19.72	22.33	29.24	19.74	24.48	27.74	28.87	35.90
Baleares	25.38	30.57	31.88	33.81	44.89	7.64	13.56	15.72	19.92	31.24	16.32	21.66	23.54	26.57	37.69
Barcelona	44.52	52.92	60.76	62.25	73.19	14.98	27.24	35.76	42.53	54.70	29.80	39.87	47.91	52.03	63.54
Burgos	68.80	77.48	81.80	84.84	88.05	13.28	25.69	34.84	51.30	62.87	41.08	51.13	58.12	67.94	75.28
Cáceres	36.35	40.13	44.16	49.54	52.07	9.14	14.09	18.27	24.26	31.57	22.99	27.16	31.30	36.92	41.68
Cádiz	36.07	38.35	43.03	44.53	50.57	23.86	28.49	32.81	34.97	40.83	30.44	33.54	38.02	39.73	45.75
Canarias	18.53	21.37	23.57	31.51	32.13	8.94	12.95	17.66	24.39	28.10	13.09	16.71	20.19	27.58	29.86
Castellón	21.39	25.64	27.95	35.11	42.53	4.58	8.62	9.16	17.26	23.94	12.94	16.95	18.55	26.12	33.16
Ciudad Real	30.28	33.56	36.48	39.32	42.63	8.36	13.73	16.65	20.68	23.88	19.42	23.64	26.53	29.88	33.17
Córdoba	25.25	29.12	34.85	35.46	38.58	10.84	15.56	19.07	21.83	26.70	18.01	22.29	26.94	28.65	32.66
Coruña	37.78	41.94	47.15	51.50	58.37	6.52	10.15	13.18	17.92	26.15	20.00	23.56	27.61	32.04	39.41
Cuenca	38.00	41.95	46.10	47.46	50.90	7.97	13.27	17.53	21.41	25.66	22.92	27.41	31.73	34.37	38.23
Gerona	38.10	45.13	51.22	59.79	66.45	9.82	17.11	24.13	34.38	46.75	23.94	31.09	37.72	47.00	56.58
Granada	23.87	20.33	26.50	31.23	38.19	9.21	10.52	13.97	18.69	25.44	16.48	15.35	20.16	24.88	31.71
Guadalajara	52.99	58.01	59.44	66.24	69.25	9.67	16.90	22.04	31.75	39.12	32.01	37.40	40.87	49.04	54.31
Guipúzcoa	31.51	40.72	47.70	58.16	67.41	15.89	27.20	35.85	49.62	61.48	23.64	33.93	41.64	53.75	64.34
Huelva	30.92	36.05	33.48	47.01	48.96	12.90	20.34	20.50	32.40	36.32	22.02	28.24	27.24	39.66	42.64
Huesca	34.47	41.09	46.82	55.70	59.99	5.76	11.61	16.79	27.32	35.38	20.54	26.58	32.22	41.77	47.94
Jaén	23.85	26.69	30.36	29.91	34.60	9.81	13.42	15.95	17.03	22.12	17.01	20.18	23.23	23.53	28.45
León	60.10	67.44	73.05	76.53	81.31	8.91	15.08	19.42	30.30	43.02	33.27	39.71	44.78	52.08	60.71
Lérida	29.48	34.44	40.92	50.03	57.65	5.03	11.55	15.75	26.54	36.80	17.49	23.01	28.56	38.58	47.49
Logroño	59.14	63.62	67.43	70.73	74.58	22.35	30.95	37.84	46.49	56.03	40.23	46.73	52.23	58.28	64.87
Lugo	41.40	41.36	49.39	53.40	61.00	3.96	5.60	8.42	15.64	21.10	21.09	22.09	27.61	33.26	39.31
Madrid	61.53	72.25	74.49	80.51	81.49	32.33	45.28	49.99	61.31	63.79	47.58	58.59	61.95	70.37	72.14
Málaga	23.27	24.27	27.36	29.94	30.64	11.04	14.25	16.33	20.31	20.68	17.14	19.11	21.69	25.00	25.54
Murcia	24.12	27.26	29.97	35.12	39.57	8.38	12.25	14.45	21.20	22.78	16.22	19.64	22.14	28.08	31.04
Navarra	48.66	54.45	59.88	68.02	72.19	19.30	30.51	38.72	52.48	61.00	34.00	42.69	49.16	60.11	66.47
Orense	36.02	38.10	45.85	49.50	48.21	3.11	5.28	7.50	12.15	26.75	18.51	20.48	25.29	29.10	36.43
Oviedo	51.25	58.28	64.90	69.63	79.78	9.67	15.60	22.83	33.12	56.32	27.94	34.10	41.26	49.70	66.76
Palencia	68.27	77.19	80.45	83.12	86.92	15.26	28.40	36.14	48.57	61.12	41.92	52.46	58.12	65.62	73.71
Pontevedra	47.86	51.69	53.95	57.28	65.13	4.34	9.16	11.75	18.19	28.72	22.49	26.55	28.67	33.71	43.22
Salamanca	50.67	59.72	65.53	71.71	76.39	12.11	22.88	30.96	44.14	51.47	31.39	40.93	48.03	57.64	63.40
Santander	72.81	78.97	83.18	80.87	91.37	19.89	32.27	43.34	57.06	83.92	44.00	53.21	61.29	68.05	87.34
Segovia	65.27	72.10	78.18	82.37	87.26	14.92	26.81	37.68	50.07	62.68	40.55	49.44	57.95	66.19	74.83
Sevilla	31.29	36.43	39.39	47.76	47.65	16.93	23.84	26.47	34.65	34.88	24.33	30.15	32.92	41.13	41.20
Soria	68.67	75.50	78.74	82.32	83.94	10.80	18.86	27.03	39.02	50.60	38.75	45.60	51.65	59.73	66.53
Tarragona	31.64	36.20	42.24	47.82	54.37	8.56	15.76	22.02	29.31	38.49	19.99	25.84	32.14	38.49	46.36
Teruel	36.74	40.55	44.73	49.66	54.42	5.38	9.66	14.49	21.61	25.82	20.69	24.60	29.37	35.55	40.12
Toledo	36.23	39.37	43.81	44.57	49.40	10.97	17.50	23.58	26.29	31.82	23.96	28.58	33.80	35.45	40.61
Valencia	26.81	26.26	35.41	41.30	44.91	8.90	13.35	18.12	23.62	28.95	17.77	19.71	26.70	32.37	36.86
Valladolid	60.39	67.85	72.08	74.87	78.87	18.76	29.73	37.50	46.72	55.71	39.97	48.33	54.62	60.38	66.84
Vizcaya	46.89	55.64	63.75	72.64	80.59	19.77	30.00	39.33	52.37	63.84	32.66	42.59	51.24	62.51	71.87
Zamora	54.76	61.64	70.28	76.94	79.29	11.03	18.02	24.20	32.30	41.89	32.41	39.03	46.38	53.57	59.22
Zaragoza	33.02	40.99	46.33	49.84	55.50	9.39	16.80	22.77	29.77	38.09	21.40	28.74	34.53	39.64	46.59
Melilla y P.S.	40.28	49.42	57.46	53.79	64.12	44.95	45.81	50.42	50.31	46.07	40.74	48.72	56.22	52.77	60.15
ESPAÑA	38.90	43.51	48.18	52.69	57.55	11.14	17.86	22.84	30.54	38.55	24.78	30.27	35.14	41.24	47.68

Appendix 2 Literacy rates (male, female aged 11-15) in Spain (1887, 1900 and 1910)

	1887	1900	1910	1887	1900	1910
	Male	Male	Male	Female	Female	Female
	T_{11-15}	T_{11-15}	T_{11-15}	T_{11-15}	T_{11-15}	T_{11-15}
Álava	83.94	83.88	88.63	69.76	77.80	84.15
Albacete	27.21	27.99	31.06	16.84	19.85	25.04
Alicante	23.81	35.75	34.32	16.31	26.63	28.73
Almería	15.49	26.61	26.83	8.19	19.61	21.47
Ávila	58.94	60.61	65.25	45.36	50.55	59.61
Badajoz	32.15	27.90	42.67	25.21	23.26	40.49
Baleares	26.10	30.17	45.33	18.57	24.62	40.21
Barcelona	61.28	61.27	73.81	46.14	52.35	66.01
Burgos	84.57	83.40	87.18	61.26	71.76	80.05
Cáceres	39.88	40.81	43.67	23.24	29.47	34.43
Cádiz	35.03	37.94	37.87	31.94	34.54	40.46
Canarias	20.85	28.20	28.39	22.23	28.81	31.05
Castellón	21.05	32.41	39.91	10.67	24.19	32.65
Ciudad Real	30.87	34.52	31.43	19.13	23.96	23.55
Córdoba	29.82	26.91	30.94	22.71	22.90	29.22
Coruña	42.06	47.20	56.23	20.43	28.54	39.73
Cuenca	40.11	37.71	42.65	23.62	24.49	29.18
Gerona	55.71	63.96	71.26	34.69	49.12	63.15
Granada	19.05	24.61	32.56	14.93	20.37	26.37
Guadalajara	55.87	61.67	64.79	31.39	43.38	51.39
Guipúzcoa	54.47	67.49	75.87	48.87	66.72	77.11
Huelva	25.74	38.00	44.97	22.08	33.47	40.90
Huesca	50.40	65.25	64.26	27.15	49.24	54.79
Jaén	27.25	22.66	23.92	20.28	18.23	12.54
León	66.08	69.86	77.75	28.09	41.54	60.46
Lérida	43.05	57.50	62.96	25.05	45.91	55.03
Logroño	67.27	69.02	73.62	54.04	59.99	69.00
Lugo	39.93	43.11	51.55	12.27	17.45	30.25
Madrid	72.48	77.82	77.06	58.01	72.01	69.78
Málaga	22.86	24.54	21.28	17.00	22.57	19.48
Murcia	23.92	28.06	28.51	15.85	21.33	21.72
Navarra	66.15	76.42	76.00	55.99	72.47	76.38
Orense	37.94	39.65	43.66	12.13	18.34	32.01
Oviedo	62.60	66.13	79.92	32.42	36.64	68.67
Palencia	80.86	80.56	85.47	53.54	63.26	75.69
Pontevedra	44.25	49.55	59.94	17.29	26.78	43.95
Salamanca	68.40	72.40	76.22	49.87	60.80	68.53
Santander	85.36	82.91	93.88	59.89	66.12	91.56
Segovia	80.74	81.19	86.28	57.21	66.35	77.73
Sevilla	32.96	46.40	38.03	28.82	39.67	33.62
Soria	80.01	82.16	85.38	47.96	60.39	77.21
Tarragona	44.53	47.67	59.39	31.26	39.45	53.54
Teruel	45.81	49.45	51.91	23.93	34.91	36.09
Toledo	38.21	35.36	38.06	27.79	27.90	32.46
Valencia	29.35	35.94	38.36	19.41	26.71	29.58
Valladolid	72.28	71.32	75.66	52.54	57.12	65.02
Vizcaya	67.76	76.96	85.80	52.48	68.61	78.41
Zamora	67.74	75.03	75.38	37.14	47.83	59.42
Zaragoza	49.87	50.02	56.71	33.51	39.24	50.35
Melilla y P.S.	63.00	62.93	60.05	65.69	57.61	49.46
ESPAÑA	44.33	49.11	53.43	29.58	38.07	45.90

SPANISH JOURNAL OF STATISTICS
Vol. **7** No. 1 2025, Pages 75–89
doi:https://doi.org/10.37830/
SJS.2025.1.04

REGULAR ARTICLE

Feasibility of Implementing Accelerometers in the Spanish Health Survey

Borja del Pozo Cruz[1], Rosa M. Alfonso Rosa[2], and Jesús del Pozo-Cruz[3]

[1]1Department of Sport Sciences, Faculty of Medicine, Health, and Sports, Villaviciosa de Odón, Madrid, Spain, borja.delpozo@universidadeuropea.es
[2]Epidemiology of Physical Activity and Fitness across Lifespan Research Group (EPAFit), Department of Human Motricity and Sport Performance, University of Seville, Sevilla, Spain, roalrosa@us.es
[3]Epidemiology of Physical Activity and Fitness across Lifespan Research Group (EPAFit), Department of Physical Education and Sports, University of Seville, Sevilla, Spain, jpozo2@us.es

Received: July 28, 2025. Returned: -. Revised: -. Accepted: October 23, 2025.

Abstract: Accurately measuring physical activity, sedentary behavior, and sleep is vital for public health monitoring, but self-reported data are often biased. Accelerometers offer objective data, yet their feasibility within the Spanish Health Survey (ESdE) has not been assessed. This study evaluated the integration of thigh-worn accelerometers in ESdE by analyzing participant compliance, device usability, data return, and comparisons with self-reported measures. A total of 100 adults aged 30–90 were recruited through five provincial delegations of the National Statistics Institute (INE), with each delegation enrolling 20 participants equally split between home-based collection and prepaid return groups. All participants wore a thigh-mounted SENS accelerometer continuously for 7 to 10 days using a water-resistant patch, with two patches provided in case of replacement. INE staff administered the ESdE questionnaire and coordinated device logistics. Valid accelerometry data were obtained from 98 participants, with excellent compliance. Device return rates were 100collection and 85Comparison with self-reported data was only possible for sedentary behavior, where participants consistently underestimated sitting time. Agreement between self-reports and accelerometry was low (ICC = -0.05 to 0.43), and Bland-Altman plots revealed a clear negative bias. These findings demonstrate the feasibility of incorporating accelerometry into national surveys like ESdE, with high participant adherence and minimal operational issues. The objective data provided by accelerometers can complement self-reported measures and capture domains like sleep and incidental activity, which are often missed. Their inclusion in future surveys may enhance the accuracy and utility of lifestyle surveillance in Spain.

Keywords: accelerometry, physical activity, sedentary behavior, sleep, health survey, feasibility, self-report, objective measurement, public health surveillance

MSC: 62P10, 62C05, 62M20, 68T09

1 Introduction

A healthy lifestyle, characterized by regular physical activity, reduced sedentary behavior, and sufficient sleep, plays a crucial role in reducing morbidity and mortality DelPozoCruz2022.JAMAInternalMedicine, McGregor2021.EuropeanJournalOfPreventiveCardiology. The World Health Organization (WHO) and national guidelines recommend at least 150 minutes of moderate-intensity or 75 minutes of vigorous-intensity aerobic physical activity per week, along with muscle-strengthening exercises twice a week (Bull et al., 2020), to promote health (López-Bueno et al., 2023). Additionally, limiting sedentary time (Bull et al., 2020) and ensuring 7–9 hours of sleep per night (Watson et al., 2015) are associated with a lower risk of chronic disease and improved quality of life.

Despite their well-documented benefits, accurately assessing these behaviors at the population level remains challenging. National health surveys, such as the Spanish Health Survey (ESdE), primarily rely on self-reported questions, which are currently the only widely available tool for assessing lifestyle behaviors at the population level. While these self-reported measures are cost-effective and provide valuable information, it is important to recognize their inherent limitations, such as recall bias, social desirability bias, and limited ability to capture incidental or unstructured movement. These limitations may lead to some degree of misclassification and potential underestimation of associations between lifestyle behaviors and health outcomes. Therefore, complementary objective measures, such as accelerometry, can enhance the accuracy and depth of data collected in national health surveys, improving public health monitoring and research (Pedišić and Bauman, 2015).

Accelerometers, commonly embedded in smartwatches, activity trackers, and smartphones, capture continuous movement data with high temporal resolution, enabling more precise detection of physical activity intensity, sedentary patterns, and sleep behaviors. Large-scale epidemiological studies, such as NHANES in the United States (Matabuena et al., 2022), the UK Biobank (del Pozo Cruz et al., 2022,?), and the SHARE study in Europe (del Pozo Cruz et al., 2023), have successfully incorporated accelerometry to enhance population-level lifestyle surveillance, demonstrating its feasibility and scientific utility.

However, the feasibility of integrating accelerometry into ESdE remains unexplored, limiting the expansion of objective lifestyle data at the population level. Given that self-reported data inherently depend on respondents' recall and perception, complementary objective measurements could provide valuable additional insights. This study assessed the feasibility of implementing accelerometers in ESdE to objectively measure physical activity, sedentary behavior, and sleep patterns. Specifically, we evaluated participant compliance, device usability, and adherence to wear protocols, while also comparing accelerometer-derived data with self-reported measures to identify potentential discrepancies between subjective and objective assessments. Additionally, we examined operational and logistical challenges, including device distribution, retrieval methods, and data processing, to assess the practicality of large-scale implementation. By addressing these objectives, this study provides insights into the feasibility of using accelerometers in national health surveys, identifying potential barriers and facilitators for their future application in large-scale public health monitoring.

2 Methods

2.1 Study Design and Participants

This feasibility study was conducted within the framework of ESdE to assess the integration of accelerometers into routine population health data collection. The study was carried out in collaboration with the Spanish National Statistical Institute (INE) and implemented across five provincial INE delegations. Each delegation recruited 20 volunteers, resulting in a total sample of 100 participants. Participants were recruited through the INE network and volunteers from each delegation. While this was a convenience sample, efforts were made to ensure an even distribution of participants across age groups (30–90 years) and sex. The study was conducted in accordance with the principles of the Declaration of Helsinki. Participants signed an informed consent form before taking part in the study and were informed verbally about its details.

To evaluate different data collection approaches, the first 10 participants recruited in each delegation were assigned to the home-based collection group, while the remaining 10 participants were assigned to the prepaid return group. In the home-based collection group, trained INE personnel visited participants' homes, administered the ESdE questionnaire, provided the accelerometer, and returned after the monitoring period to collect the device. In the prepaid return group, participants received the accelerometer during the ESdE visit, along with a prepaid envelope to return the device by mail after the monitoring period. All participants were instructed to wear a SENS accelerometer on their thigh continuously for 7 to 10 consecutive days, day and night, without removal.

2.2 Training and Data Collection Procedures

A standardized training session was conducted at the central INE offices to ensure adherence to study protocols across all delegations. Fieldworkers from the five provincial delegations participated in face-to-face hands-on training, where they received all necessary materials and devices. The session covered proper accelerometer placement, participant guidance, device retrieval logistics for home visits and prepaid returns, and quality control measures for data validation. Additionally, participants attended a remote session on administering the ESdE questionnaire.

Participants were instructed to wear the device continuously without removal, as the patch was water-resistant, allowing them to wear it while showering and engaging in daily activities. To account for potential adhesion issues, each participant received two patches, enabling them to replace the device if detachment occurred. Compliance was monitored through self-reported adherence logs, and participants were encouraged to report any issues with device wear.

2.3 Accelerometer Device and Data Processing

The SENS accelerometer was selected for its validated capability to measure physical activity, sedentary behavior, and sleep patterns. This waterproof triaxial accelerometer (45 × 23 × 5 mm, 6 g) was designed to be worn on the thigh, approximately 10 cm above the lateral epicondyle, using a skin-attached patch specifically designed for the device. It recorded acceleration at 12 Hz, capturing orientation and movement intensity. Data were transmitted wirelessly to a smartphone application every 10 minutes when within range or stored for later transmission. Anonimized raw

data were automatically uploaded to a secure web server for processing. A previously validated rule-based, activity pattern recognition algorithm was used to process the data (Pedersen et al., 2022; Milther et al., 2023; Bartholdy et al., 2018; McGrosky et al., 2025). Recorded data were analyzed in 5-second epochs, categorizing each interval into one of nine predefined activity categories based on movement frequency, intensity, and sensor orientation using a validated pattern-recognition algorithm. These categories included Resting (lying or sitting rest), Lying or Sitting Movement, Standing, Sporadic Walking, Walking, Moderate Intensity, High Intensity or Running, Cycling, and Steps Taken. Step detection was based on FFT frequency-domain analysis, differentiating between sporadic, continuous, and high-intensity steps. The intensity count, ranging from 0 to 100, was derived from high-pass filtered accelerometer data, subtracting noise to ensure accurate classification. Standing was identified with intensity values below 2, sporadic walking fell between 2 and 10, continuous walking was detected between 10 and 50, moderate-intensity activities such as slow running ranged between 50 and 75, and high-intensity activities such as fast running exceeded 75.

Bedtime was detected within a predefined nighttime window between 6:00 PM and 1:00 AM, ensuring a focus on nocturnal rest while excluding daytime naps. Sleep onset was estimated as the last walking episode lasting more than one minute between 6:00 PM and 1:00 AM, while wake time was identified as the first walking episode lasting at least 30 seconds between 5:00 AM and 12:00 PM. Once these parameters were established, sleep was further classified into three categories based on movement patterns: Sleep No Movement, Sleep Movement, and Sleep Active. Sleep No Movement was characterized by a horizontal body position within ±45 degrees, with an intensity count below 2, indicating minimal or no movement. Sleep Movement occurred when the participant remained in a lying position but exhibited low-to-moderate movement, classified by an intensity count above 2. To improve classification accuracy, detected movements were assigned an additional 30 seconds before and after the event. This category primarily represented lighter sleep stages, where movement was more frequent. Sleep Active identified episodes in which the participant was standing upright within ±45 degrees during sleep periods, regardless of whether movement occurred. This classification was indicative of brief awakenings or nighttime restlessness.

2.4 Feasibility Assessment

Feasibility was evaluated by assessing compliance rates, return rates, participant burden, and operational challenges. Compliance was monitored through self-reported adherence logs, and participants were encouraged to report any issues with device wear. Compliance rates were determined by the percentage of participants who wore the accelerometer continuously for at least 7 valid days (or 10 days when applicable), without non-wear periods. Return rates were measured as the percentage of devices successfully retrieved. Participant burden was assessed through self-reported feedback on ease of use, comfort, and willingness to participate in future similar studies. Operational challenges were documented based on fieldworker feedback regarding recruitment, device management, and logistical difficulties. A focus group was conducted with the interviewers involved in data collection to discuss practical challenges and perceptions of feasibility.

2.5 Comparison with Self-Reported Measures

To evaluate potential discrepancies between subjective and objective assessments, participants completed a questionnaire as part of the ESdE survey, which inquired about their physical activity and sedentary behavior. The questionnaire included questions about the number of days per week they engage in moderate-to-vigorous physical activities (MVPA) such as sports, gymnastics, cycling, or brisk walking for at least 10 consecutive minutes, as well as the total amount of time dedicated to these activities in a typical week, reported in hours and minutes. Sedentary behavior was measured by asking participants to report the total amount of time they spend sitting on a typical day, including time spent at work, home, studying, reading, commuting, or engaging in leisure activities such as watching television.

In this study, we focused on comparing sedentary time between self-reported and accelerometer-derived data, as sedentary behavior represents a comparable construct across both methods. Sedentary time was estimated from accelerometry using the sum of time spent resting while sitting or lying down and moving while sitting.

In contrast, accelerometer-derived physical activity and self-reported MVPA represent different but complementary constructs. Accelerometry captures a continuous and comprehensive spectrum of movement intensity and patterns throughout the day, including incidental and unstructured activities that are not fully captured by questionnaire items focused on specific types of physical activity performed in bouts of at least 10 minutes. Therefore, direct comparison between accelerometer-measured physical activity and selfreported MVPA is not appropriate. Instead, these methods provide complementary insights that together offer a more complete understanding of participants' activity behaviors.

2.6 Statistical Analysis

Descriptive statistics were used to summarize participant characteristics, wear-time compliance, and device return rates. Compliance and return rates were reported separately for each data collection. Agreement between self-reported and accelerometer-derived sedentary time was assessed using Bland-Altman plots, intraclass correlation coefficients (ICC), and Spearman's correlation coefficients. All analyses were conducted in R (v4.3.1), and statistical significance was set at $p < 0.05$.

3 Results

3.1 Feasibility and Implementation Outcomes

The study initially enrolled 100 participants, all of whom were given the option to wear the device for 7 to 10 days. Compliance was excellent, with no data loss except for one participant who only wore the device for 2 days, leading to their exclusion from the final dataset and reducing the number of valid participants to 99. Additionally, one participant lost the device before wearing it, meaning they were recruited but did not contribute usable data, further reducing the dataset to 98 participants with valid measurements. Table 1 and Table 2 shows the participants characteristics and accelerometer-derived lifestyle behaviors measurements, respectively.

For data collection, devices were either picked up at home or returned via mail. While no devices were lost during home-based pickups (100returned via mail (85devices were lost, all data were successfully recovered. As a result, the final dataset consisted of 98 participants, all with complete and usable data for analysis.

A virtual focus group was conducted with the interviewers involved in device placement and retrieval to assess feasibility. Interviewers reported that, overall, the study procedures were feasible and well-received by participants. However, several operational challenges emerged. Some participants, particularly older adults, required in-person assistance for device placement despite receiving written instructions. Participants with hairy thighs reported difficulties with adhesion, which sometimes led to early detachment or discomfort. In some cases, individuals initially hesitated to wear the device due to concerns about having an unfamiliar object attached to their body, but most reported that they stopped noticing it after a few days. Some participants also requested reminders to remove the device at the end of the monitoring period, although this was not strictly necessary since the device automatically stopped recording on the programmed date.

3.2 Device Placement and Technical Issues

During the implementation of the preference-based device placement method, several minor incidents were reported. Three participants required assistance with device attachment due to physical limitations or lack of confidence with the device, but the remaining 95 participants were able to self-administer it successfully.

Adhesion issues were reported in two cases, particularly after prolonged wear or water exposure, requiring device replacements. Although the device was water-resistant, some detachment incidents (e.g., swimming, clothing changes) suggest that improved guidance on device maintenance and reattachment could enhance compliance.

Additionally, some participants reported skin-related or usability concerns, including mild irritation, difficulties in device removal due to strong adhesion, and interference caused by body hair affecting proper placement. Furthermore, some older participants expressed reluctance to use the device due to perceived discomfort. In general, however, most participants reported that, after the initial adjustment period, wearing the accelerometer was not bothersome.

Lastly, two MRI scans had to be repeated because participants did not remove the device, highlighting the need for clearer instructions regarding MRI compatibility.

3.3 Comparison with Self-Reported Measures

For sedentary time, Bland-Altman plots revealed a systematic underestimation of sitting time in self-reported measures, with the mean difference significantly below zero (Figure 1). This suggests that individuals tend to report less sitting time compared to accelerometer-derived estimates. The limits of agreement were wide, reinforcing the high variability of self-reported sedentary time (Figure 1). Intraclass correlation coefficients (ICC) were notably lower than for physical activity, with ICC values ranging from -0.05 (single absolute agreement) to 0.43 (average fixed raters), indicating poor agreement (Table 3). Spearman's correlation coefficient (rho = 0.39, $p < 0.001$) further supported this

finding, suggesting that self-reported sedentary time is a weak proxy for objectively measured sitting time.

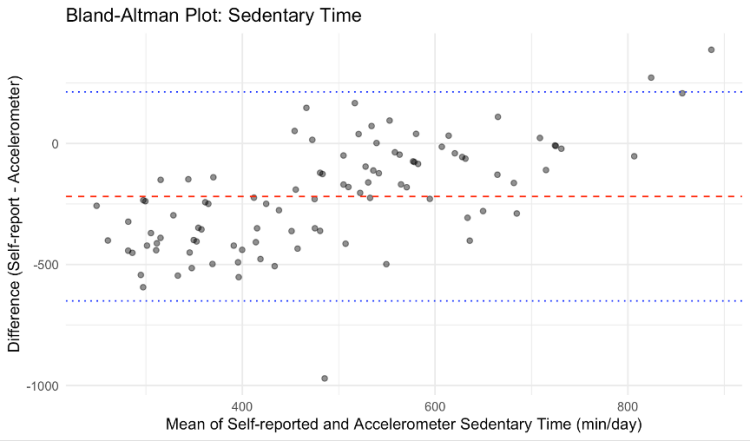


Figure 1: Mean of self-reported and accelerometer sedentary time (min/day)

4 Discussion

This study support preliminary evidence of feasibility of integrating thigh-worn accelerometers into ESdE, with high compliance, minimal data loss, and effective logistics for device placement and retrieval. However, certain practical challenges emerged that should be considered for future large-scale implementation.

One of the main challenges identified was related to device placement and participant adherence. While most participants self-administered the device, some required assistance, particularly older adults. This aligns with previous research, where older individuals have shown greater difficulty following placement instructions, often requiring additional guidance to ensure compliance (Alaqil et al., 2024). Similarly, past studies have highlighted adhesion issues, particularly with prolonged wear and exposure to water (Alaqil et al., 2024). Although providing two patches in this study helped address some of these challenges, improved guidance on reattachment and skin preparation could further reduce these issues.

Initial reluctance to wear the device was reported by some participants, primarily due to concerns about having an unfamiliar object attached to their body for an extended period. However, most reported that these concerns diminished after a few days, which is consistent with findings from previous feasibility studies on wearable accelerometers (Hamer et al., 2020). Furthermore, several participants requested reminders about when to remove the device at the end of the study period. While this did not impact data collection due to the device's pre-programmed recording period, it suggests that additional communication strategies could enhance participant engagement and adherence in future large-scale implementations.

The high compliance and feasibility outcomes observed in this study are in line with previous research examining accelerometer wear time and usability (Milther et al., 2023; Alaqil et al., 2024;

Hamer et al., 2020). Studies evaluating accelerometer feasibility in population-based surveys have consistently shown high adherence, particularly when devices are comfortable, waterproof, and accompanied by clear instructions. Research on thigh-worn accelerometers has demonstrated their capacity to reliably capture physical activity and sedentary behavior, reinforcing their suitability for large-scale public health surveillance.

Our results suggest that some discrepancies may exist between self-reported and accelerometer-derived sedentary time. We found that our participants tended to self-report less sedentary time compared to the sedentary time recorded by the accelerometer. This finding is consistent with previous evidence showing a systematic underestimation of sedentary time in undergraduate students (Nelson et al., 2019), adults (Arango Vélez et al., 2020), or older women (Shiroma et al., 2015), where self-reported sedentary time was underestimated when compared with accelerometer measurements.

However, it is important to recognize that self-reported questionnaires and accelerometry capture different yet complementary aspects of physical activity and sedentary behavior. Questionnaires, such as those used in ESdE, provide valuable contextual information about the type, setting, and subjective experience of physical activity, which cannot be fully captured by accelerometers. Conversely, accelerometers offer continuous, objective data on movement patterns, intensity, and duration, including incidental and unstructured activities often missed by self-report. Integrating self-reported and accelerometer-derived data can enrich our understanding of lifestyle behaviors. This combined approach enhances the precision and contextualization of physical activity and sedentary behavior assessment, providing a more comprehensive foundation for public health monitoring of lifestyles.

Implications for Future Research and Policy

If confirmed in larger studies, our results may support the potential scalability of accelerometry in national health surveys. However, several refinements should be considered for future implementation. First, improving instructions on device placement—particularly for older adults and those with long body hair—may enhance adherence. Additionally, the study highlights the importance of participant education to address trust issues and initial hesitation regarding the device.

From a methodological perspective, integrating accelerometers into national surveys presents clear advantages over including self-reported measures only. The current approach in ESdE relies on self-reported assessments of physical activity and sedentary behavior. Although the physical activity questions included in the ESdE are standardized within the European Health Interview Survey (EHIS), they do not originate from a full validated questionnaire—i.e., a tool rigorously tested for reliability and accuracy across populations but rather consist of a set of independent questions designed to capture selected aspects of activity patterns. This limits their ability to comprehensively assess movement behaviors and introduces challenges when comparing results with studies using validated instruments.

One of the key features of the current ESdE approach is its use of a 10-minute bout criterion for physical activity reporting, which aligns with the standardized methodology agreed upon by EHIS and member states. While this approach provides a consistent framework for data collection and comparison across countries, recent evidence suggests that accumulating shorter bouts of movement throughout the day may also contribute significantly to health benefits (Stamatakis et al., 2025; Ahmadi et al., 2022). Incorporating accelerometry into the survey offers an opportunity to capture these shorter, incidental, and unstructured activities, thereby enhancing both the quantity and

quality of information on real-world movement patterns and overall energy expenditure.

Additionally, current self-reported measures primarily capture structured physical activities, such as sports participation and active transportation, which provide valuable information within these domains. However, a wide range of daily movements—including household chores, occupational movement, and spontaneous short bouts of physical activity—may not be fully captured by questionnaires. These types of activities can be objectively recorded using accelerometers (Stamatakis et al., 2025). Given the growing evidence supporting the health benefits of lightintensity activities (del Pozo Cruz et al., 2021),integrating objective measures alongside self-reports can provide a more complete picture of overall physical activity levels.

Another important consideration is the inherent limitations of self-reported methods in accurately assessing sedentary behavior. While the ESdE includes questions on total sitting time, self-reported sedentary behavior is subject to recall bias and may underestimate true sedentary time. Accelerometry offers continuous, objective data on sedentary patterns, including prolonged bouts of uninterrupted sitting, which are particularly relevant due to their association with adverse health outcomes (Saunders et al., 2020).

Beyond physical activity and sedentary behavior, sleep duration and patterns are not currently assessed in the ESdE, despite being a key component of lifestyle and health. Accelerometers offer the advantage of capturing both movement and sleep parameters, providing a more complete picture of daily behaviors that impact health. The inclusion of sleep monitoring in future national health surveys could offer valuable insights into its interaction with physical activity and sedentary time, further strengthening public health surveillance.

Despite these advantages, it is important to acknowledge that accelerometers cannot measure all aspects of physical activity, particularly muscle-strengthening exercises, which are included in the current survey, albeit only the number of days dedicated to it and not time. Resistance training activities often involve minimal movement at the acceleration level detectable by these devices, making it necessary to maintain a complementary role for selfreported information in certain domains. However, combining accelerometry with contextual self-reported data would enhance the interpretability of results, allowing for a better understanding of when, where, and why individuals engage in specific movement behaviors.

Given its ability to objectively and comprehensively assess movement behaviors, policymakers should consider the long-term integration of accelerometry into national health surveillance. This would improve the precision of lifestyle behavior monitoring, provide stronger epidemiological evidence, and support the development of more effective public health strategies to promote physical activity and reduce sedentary time at the population level.

4.1 Conclusions

This study provides preliminary evidence of the feasibility of using thigh-worn accelerometry for objective lifestyle monitoring within ESdE. Despite minor adhesion and usability challenges, compliance was excellent, and no data loss occurred among participants who met the minimum wear-time criteria. Most participants successfully self-administered the device, reinforcing its scalability for large population-based surveys. Self-reported measures tend to overestimate physical activity

and underestimate sedentary time, reinforcing the importance of integrating accelerometry into future national surveys. Moving forward, a combined approach using both self-reported and objective measures may enhance the accuracy of population-level lifestyle assessments.

Acknowledgments

We sincerely thank the interviewers from the participating delegations of the National Department of Statistics (Susana María Morales Jiménez, Guadalajara; Antonio Luna Gomez, Madrid; Carmen Soriano Martín, Granada; Raúl Sala Gómez, Alicante; Angel Manuel Estévez Rivera, Alicante; and Rafael Ferrer Chamorro, Málaga) for their support in data collection and participant engagement. We also extend our gratitude to all participants for their commitment to this study.

References

Ahmadi, Matthew N., Philip J. Clare, Peter T. Katzmarzyk, Borja del Pozo Cruz, I-Min Lee, and Emmanuel Stamatakis (2022). Vigorous physical activity, incident heart disease, and cancer: How little is enough? *European Heart Journal 43*(46), 4801–4814. DOI: doi:10.1093/eurheartj/ehac572.

Alaqil, Abdulrahman I., Borja del Pozo Cruz, Shaima A. Alothman, Matthew N. Ahmadi, Paolo Caserotti, Hazzaa M. Al-Hazzaa, Andreas Holtermann, Emmanuel Stamatakis, and Nidhi Gupta (2024). Feasibility and acceptability of a cohort study baseline data collection of device-measured physical behaviors and cardiometabolic health in saudi arabia: Expanding the prospective physical activity, sitting and sleep consortium (propass) in the middle east. *BMC Public Health 24*(1), 1379. DOI: doi:10.1186/s12889-024-18438-1.

Arango Vélez, Elkin Fernando, Andrés Mauricio Echavarría Rodríguez, Fabián Alexander Aguilar-González, and Fredy Alonso Patiño Villada (2020). Validación de dos cuestionarios para evaluar el nivel de actividad física y el tiempo sedentario en una comunidad universitaria de colombia. *Revista Facultad Nacional de Salud Pública 38*(1), 1–11. DOI: doi:10.17533/udea.rfnsp.v38n1e334156.

Bartholdy, Cecilie, Henrik Gudbergsen, Henning Bliddal, Morten Kjæ rgaard, Kasper Lundberg Lykkegaard, and Marius Henriksen (2018). Reliability and construct validity of the sens motion® activity measurement system as a tool to detect sedentary behaviour in patients with knee osteoarthritis. *Arthritis 2018*, 6596278. DOI: doi:10.1155/2018/6596278.

Bull, Fiona C., Salih S. Al-Ansari, Stuart Biddle, Katja Borodulin, Matthew P. Buman, Greet Cardon, Catherine Carty, Jean-Philippe Chaput, Sebastien F. M. Chastin, Roger Chou, Paddy C. Dempsey, Loretta DiPietro, Ulf Ekelund, Joseph Firth, Christine M. Friedenreich, Leandro Garcia, Muthoni Gichu, Russell Jago, Peter T. Katzmarzyk, Estelle Lambert, Michael Leitzmann, Karen Milton, Francisco B. Ortega, Chathuranga Ranasinghe, Emmanuel Stamatakis, Anne Tiedemann, Richard P. Troiano, Hidde P. van der Ploeg, Vicky Wari, and Juana F. Willumsen (2020). World health organization 2020 guidelines on physical activity and sedentary behaviour. *British Journal of Sports Medicine 54*(24), 1451–1462. DOI: doi:10.1136/bjsports-2020-102955.

del Pozo Cruz, Borja, Matthew Ahmadi, Sharon L. Naismith, and Emmanuel Stamatakis (2022). Association of daily step count and intensity with incident dementia in 78,430 adults living in the united kingdom. *JAMA Neurology 79*(10), 1059–1063. DOI: doi:10.1001/jamaneurol.2022.2672.

del Pozo Cruz, Borja, Matthew N. Ahmadi, I-Min Lee, and Emmanuel Stamatakis (2022). Prospective associations of daily step counts and intensity with cancer and cardiovascular disease incidence and mortality and all-cause mortality. *JAMA Internal Medicine 182*(11), 1139–1148. DOI: doi:10.1001/jamainternmed.2022.3446.

del Pozo Cruz, Borja, Rosa M. Alfonso-Rosa, Rubén López-Bueno, Stuart J. Fairclough, Alex Rowlands, and Jesús del Pozo-Cruz (2023). Associations between hospitalization and device-assessed physical activity in a representative sample of older adults. *Gerontology 69*(4), 506–512. DOI: doi:10.1159/000527377.

del Pozo Cruz, Borja, Stuart J. H. Biddle, Paul A. Gardiner, and Ding Ding (2021). Light-intensity physical activity and life expectancy: National health and nutrition survey. *American Journal of Preventive Medicine 61*(3), 428–433. DOI: doi:10.1016/j.amepre.2021.02.011.

Hamer, Mark, Emmanuel Stamatakis, Sebastien Chastin, Natalie Pearson, Matt Brown, Emily Gilbert, and Alice Sullivan (2020). Feasibility of measuring sedentary time using data from a thigh-worn accelerometer. *American Journal of Epidemiology 189*(9), 963–971. DOI: doi:10.1093/aje/kwaa075.

López-Bueno, Rubén, Matthew Ahmadi, Emmanuel Stamatakis, Lin Yang, and Borja del Pozo Cruz (2023). Prospective associations of different combinations of aerobic and muscle-strengthening activity with all-cause, cardiovascular, and cancer mortality. *JAMA Internal Medicine 183*(9), 982–990. DOI: doi:10.1001/jamainternmed.2023.2613.

Matabuena, Marcos, Paulo Félix, Ziad Akram Ali Hammouri, Jorge Mota, and Borja del Pozo Cruz (2022). Physical activity phenotypes and mortality in older adults: A novel distributional data analysis of accelerometry in the nhanes. *Aging Clinical and Experimental Research 34*(12), 3107–3114. DOI: doi:10.1007/s40520-022-02178-z.

McGrosky, Jared, Matthew N. Ahmadi, Borja del Pozo Cruz, Jesús del Pozo Cruz, and Emmanuel Stamatakis (2025). Energy expenditure and obesity across the economic spectrum: A global analysis of accelerometry data. *Nature Human Behaviour 9*(3), 455–463. DOI: doi:10.1038/s41562-025-01985-3.

Milther, Camilla, Lærke Winther, Michelle Stahlhut, Derek John Curtis, Mette Aadahl, Morten Tange Kristensen, Jette Led Sørensen, and Christian Have Dall (2023). Validation of an accelerometer system for measuring physical activity and sedentary behavior in healthy children and adolescents. *European Journal of Pediatrics 182*(8), 3639–3647. DOI: doi:10.1007/s00431-023-04947-4.

Nelson, Megan C., Katie Taylor, and Chantal A. Vella (2019). Comparison of self-reported and objectively measured sedentary behaviour and physical activity in undergraduate students. *Measurement in Physical Education and Exercise Science 23*(3), 237–248. DOI: doi:10.1080/1091367X.2019.1610765.

Pedersen, Britt Stævnsbo, Morten Tange Kristensen, Christian Ohrhammer Josefsen, Kasper Lundberg Lykkegaard, Line Rokkedal Jønsson, and Mette Merete Pedersen (2022). Validation of two activity monitors in slow and fast walking hospitalized patients. *Rehabilitation Research and Practice 2022*, 9230081. DOI: doi:10.1155/2022/9230081.

Pedišić, Željko and Adrian Bauman (2015). Accelerometer-based measures in physical activity surveillance: current practices and issues. *British Journal of Sports Medicine 49*(4), 219–223. DOI: doi:10.1136/bjsports-2013-093407.

Saunders, Travis J., Travis McIsaac, Kevin Douillette, Nick Gaulton, Stephen Hunter, Ryan E. Rhodes, Stephanie A. Prince, Valerie Carson, Jean-Philippe Chaput, Sebastien Chastin, Lora Giangregorio, Ian Janssen, Peter T. Katzmarzyk, Michelle E. Kho, Veronica J. Poitras, Kenneth E. Powell, Robert Ross, Amanda Ross-White, Mark S. Tremblay, and Genevieve N. Healy (2020). Sedentary behaviour and health in adults: An overview of systematic reviews. *Applied Physiology, Nutrition, and Metabolism 45*(10 Suppl 2), S197–S217. DOI: doi:10.1139/apnm-2020-0272.

Shiroma, Eric J., Nancy R. Cook, JoAnn E. Manson, Julie E. Buring, Eric B. Rimm, and I-Min Lee (2015). Comparison of self-reported and accelerometer-assessed physical activity in older women. *PLOS ONE 10*(12), e0145950. DOI: doi:10.1371/journal.pone.0145950.

Stamatakis, Emmanuel, Matthew N. Ahmadi, Raaj Kishore Biswas, Borja del Pozo Cruz, Cecilie Thøgersen-Ntoumani, Marie H. Murphy, Angelo Sabag, Scott Lear, Clara Chow, Jason M. R. Gill, and Mark Hamer (2025). Device-measured vigorous intermittent lifestyle physical activity (vilpa) and major adverse cardiovascular events: Evidence of sex differences. *British Journal of Sports Medicine 59*(5), 316–324. DOI: doi:10.1136/bjsports-2024-108484.

Watson, Nathaniel F., M. Safwan Badr, Gregory Belenky, Donald L. Bliwise, Orfeu M. Buxton, Daniel Buysse, David F. Dinges, James Gangwisch, Michael A. Grandner, Clete Kushida, Jennifer L. Martin, Sanjay R. Patel, Stuart F. Quan, and Esra Tasali (2015). Recommended amount of sleep for a healthy adult: A joint consensus statement of the american academy of sleep medicine and sleep research society. *Sleep 38*(6), 843–844. DOI: doi:10.5665/sleep.4716.

Table 1. Characteristics of participants in the study (n=100)

Variable	Value
Age (yrs.)	53.48 (14.78)
Female, n, %	56 (56.0%)
Employed, n, %	70 (70.0%)
University graduate or above, n, %	53 (53.0%)
Spanish, yes, %	96 (96.0%)
Married, n, %	61 (61.0%)
Self-reported health status, n, %	
Very good	21 (21.0%)
Good	48 (48.0%)
Fair	27 (27.0%)
Bad	3 (3.0%)
Very bad	1 (1.0%)
Body Mass Index (Kg/m²)	25.40 (4.04)
No smoker, n, %	55 (55.0%)
Alcohol intake, n, %	
5-6 days per week	14 (14.0%)
3-4 days per week	4 (4.0%)
1-2 days per week	10 (10.0%)
1-2 days per week	31 (31.0%)
2-3 days per month	12 (12.0%)
Once per month	6 (6.0%)
Less than once per month	11 (11.0%)
None in the last 12 months	4 (4.0%)
Never	8 (8.0%)
Chronic disease, yes, n, %	65 (65.0%)
No limitations, n, %	66 (66.0%)
Difficulty walking 500 meters, no difficulty, n, %	94 (94.0%)
Difficulty climbing up or down 12 steps, no difficulty, n, %	91 (91.0%)
Physical activity main occupation, n, %	
Sitting most of the time	63 (63.0%)
Standing up most of the time	21 (21.0%)
Walking most of the time	10 (10.0%)
Engaging in tasks that require high physical effort	2 (2.0%)
Not applicable	2 (2.0%)
Do not know/do not answer	2 (2.0%)
Frequency of leisure time physical activity, n, %	
Never	14 (14.0%)
Physical activity occasionally	19 (19.0%)
Physical activity several times per month	17 (17.0%)
Physical activity several times per week	50 (50.0%)
Walking for transport 10 minutes or more, days, n, %	
0	4 (4.0%)
1	0 (0%)
2	4 (4.0%)
3	9 (9.0%)
4	6 (6.0%)
5	6 (6.0%)
6	4 (4.0%)
7	67 (67.0%)
Cycling for transport 10 minutes or more, days, n, %	
0	84 (84.0%)
1	5 (5.0%)
2	3 (3.0%)
3	2 (2.0%)
4	1 (1.0%)
5	2 (2.0%)
7	3 (3.0%)
Moderate-to-vigorous physical activity (min/day)	36.62 (37.95)
Sitting time (min/day)	379.32 (233.50)
Values are mean (SD) unless otherwise stated	

Table 2. Accelerometer-derived descriptive statistics (n=98)

Valid days	8.39 (1.22)
Sitting (min/day)	569.05 (114.95)
Sitting with movement (min/day)	29.03 (14.24)
Standing (min/day)	106.52 (40.10)
Sporadic walking (min/day)	104.14 (25.92)
Walking (min/day)	126.00 (35.28)
Moderate-intensity activity (min/day)	21.82 (20.10)
High-intensity activity-running (min/day)	3.48 (7.17)
Cycling (min/day)	3.39 (6.72)
Restorative sleep (min/day)	389.91 (66.14)
Restless sleep (min/day)	76.67 (33.07)
Active while sleep (min/day)	3.20 (4.21)
Sit-to-stand transitions (count)	72.95 (21.01)
Daily steps, brisk walking	10710.13 (4281.88)
Daily steps, slow walking	2465.07 (608.54)
Daily steps, sporadic walking	3339.82 (905.53)
Valid days, weekdays	6.15 (1.17)
Sitting (min/day), weekdays	572.54 (118.31)
Sitting with movement (min/day), weekdays	29.47 (14.33)
Standing (min/day), weekdays	107.75 (41.92)
Sporadic walking (min/day), weekdays	104.41 (27.70)
Walking (min/day), weekdays	127.04 (38.12)
Moderate-intensity activity (min/day), weekdays	23.32 (21.46)
High-intensity activity-running (min/day), weekdays	3.55 (7.31)
Cycling (min/day), weekdays	2.79 (4.51)
Restorative sleep (min/day), weekdays	382.64 (63.47)
Restless sleep (min/day), weekdays	73.96 (33.44)
Active while sleep (min/day), weekdays	2.39 (2.65)
Sit-to-stand transitions (count), weekdays	74.58 (22.13)
Daily steps, brisk walking, weekdays	10922.88 (4478.43)
Daily steps, slow walking, weekdays	2465.87 (648.95)
Daily steps, sporadic walking, weekdays	3365.33 (972.98)
Valid days, weekend days	2.23 (0.55)
Sitting (min/day), weekend days	555.67 (150.46)
Sitting with movement (min/day), weekend days	28.02 (17.31)
Standing (min/day), weekend days	105.16 (46.60)
Sporadic walking (min/day), weekend days	105.31 (35.08)
Walking (min/day), weekend days	125.94 (44.60)
Moderate-intensity activity (min/day), weekend days	17.35 (19.46)
High-intensity activity-running (min/day), weekend days	3.28 (8.45)
Cycling (min/day), weekend days	4.80 (14.08)
Restorative sleep (min/day), weekend days	409.16 (93.77)
Restless sleep (min/day), weekend days	84.88 (40.51)
Active while sleep (min/day), weekend days	4.89 (12.83)
Sit-to-stand transitions (count), weekend days	69.78 (25.35)
Daily steps, brisk walking, weekend days	10247.68 (4919.97)
Daily steps, slow walking, weekend days	2509.35 (837.73)
Daily steps, sporadic walking, weekend days	3351.33 (1121.57)
Values are mean (SD) unless otherwise stated	

Table 3. Intraclass correlation coefficients, Sitting time (n=98)

Type	ICC	Value	p-value	Lower Bound	Upper Bound
Single raters, absolute	ICC1	-0.055	0.70	-0.248	0.14
Single random raters	ICC2	0.163	<0.01	-0.055	0.37
Single fixed raters	ICC3	0.277	<0.01	0.086	0.45
Average raters, absolute	ICC1k	-0.117	0.70	-0.658	0.25
Average random raters	ICC2k	0.280	<0.01	-0.116	0.54
Average fixed raters	ICC3k	0.434	<0.01	0.159	0.62

SPANISH JOURNAL OF STATISTICS
Vol. **7** No. 1 2025, Pages 91–108
doi:https://doi.org/10.37830/
SJS.2025.1.05

REGULAR ARTICLE

Semiparametric von Mises kernel circular density estimator

Yasmina Ziane[1], Nabil Zougab[2], Kahina Bedouhene[3], and Smail Adjabi[4]

[1]Research unit LaMOS, Faculty of Exact Sciences, Bejaia University, 06000 Bejaia, Algeria,
yasmina.ziane@univ-bejaia.dz
[2]Research unit LaMOS, Faculty of Exact Sciences, Bejaia University, 06000 Bejaia, Algeria,
nabil.zougab@univ-bejaia.dz
[3]Department of Mathematics, University of Tizi-Ouzou, 15000 Tizi-Ouzou, Algeria,
kahina.bedouhene@yahoo.fr
[4]Research unit LaMOS, Faculty of Exact Sciences, Bejaia University, 06000 Bejaia, Algeria,
smail.adjabi@univ-bejaia.dz

Received: July 13, 2025. Returned: -. Revised: -. Accepted: October 20, 2025.

Abstract: In this paper, we propose to estimate the circular density function by the semiparametric bias-corrected circular kernel method using the particular von Mises kernel. This method consists to apply a multiplicative bias correction for the initial parametric model in order to improve the quality of the estimator as well as the bias. Two semiparametric estimators Hjort and Glad (1995) (HG) and Jones, Signorini, and Hjort (1999) (JSH) for probability density estimation are applied on circular data with support $[0, 2\pi)$. The properties of the latter are reported such as the bias, the variance and the mean square error integrated (MISE). A comparative study is performed to evaluate the performance of the semiparametric estimator (HG and JSH). The popular cross validation technique is adapted for bandwidth selection. A simulation and a real data application for circular data illustrate in terms of integrated squared bias (ISB) and integrated squared error (ISE) that the semiparametric estimators JSH and JLN with the von Mises kernel perform better than the classical and HG estimators.

Keywords: Bandwidth selection, Circular data, Cross validation, Multiplicative bias correction MBC, von Mises kernel

MSC: 60E05, 62P99

1 Introduction

In recent years, circular data analysis has received a lot of interest in the statistical literature, because they appear in several areas, such as biology (Batschelet, 1981), ecology (Jammalamadaka and Lund, 2006), meteorology (Bowers et al., 2000), sociology (Brunsdon and Corcoran, 2006), medicine (Mooney et al., 2003) and biomechanics (Mann et al., 2003). Given a random sample $\Theta_1, \Theta_2, \ldots, \Theta_n$

which take values in $[0, 2\pi)$ from some unknown density f, the nonparametric kernel density estimator of circular data proposed by (Hall et al., 1987), is given by $\hat{f}(\theta) = (1/n)\sum_{i=1}^{n} k_\nu(\theta - \Theta_i)$, where $\theta > 0$ is the estimation point and ν is the bandwidth parameter, this estimator is applied by several authors: (Bai et al., 1988), (Klemelä, 2000), (Taylor, 2008), (Marzio et al., 2009), (Marzio et al., 2011), (Oliveira et al., 2012), (Amiri et al., 2017), (Tsuruta and Sagae, 2017) and (Bedouhene and Zougab, 2019).

The multiplicative bias correction (MBC) technique improved the kernel estimation by reducing the order of magnitude of the bias from $O(\nu^{-1})$ to $O(\nu^{-2})$. Notice that the MBC techniques were originally proposed in linear symmetric and asymmetric kernel density estimation by (Terrell and Scott, 1980) (TS estimator) and (Jones et al., 1995) (JLN estimator). This techniques are later used by several authors: (Hirukawa, 2010; Hirukawa and Sakudo, 2014; Zougab and Adjabi, 2016; Funke and Kawka, 2015; Zougab et al., 2018) in continuous situation. (Harfouche et al., 2018) and (Harfouche et al., 2020) in discrete case. Recently, (Bedouhene and Zougab, 2019) applied the MBC techniques on the circular data which made it possible to improve the quality of the estimate by reducing the order of magnitude of bias from $O(\nu^{-1})$ to $O(\nu^{-2})$.

An alternative to nonparametric method, is the semiparametric estimate of the probability density proposed by (Hjort and Glad, 1995) (HG), using symmetric kernels for linear case. The idea of the latter, is to develop an estimator composed of two parts, the first one is the parametric, and the second represents the non-parametric correction function, this allows to modify the bias and keeps the variance; see also (Hagmann and Scaillet, 2007) using asymmetric kernels and (Kokonendji et al., 2009) using discrete kernels.

The present paper mainly focuses on two objectifs. The first objective is to investigate the semiparametric estimator on circular data and to develop the associated properties by using the Taylor series approximations. This has not yet been done in the literature. This work will give an idea on the efficiency of the semiparametric estimator compared to the nonparametric estimator on circular data. The semiparametric estimator given by

$$\tilde{f}(\theta) := g(\theta)\hat{r}(\theta), \theta \in [0, 2\pi), \tag{1}$$

where $g(\theta)$ is an initial density estimator and $\hat{r}(\theta) = \hat{f}(\theta)/g(\theta)$ is the correction factor. Based on the estimator defined by (1), several cases can be obtained by changing the function g, this allows us to extend the study to a more general study that includes several modes of estimating the probability density in the case of circular data. The function g can takes several forms: the uniform distribution $g \equiv 1/2\pi$ which gives a classical kernel estimator, the kernel estimator $g = \hat{f}$, which gives a bias-corrected estimator JLN (Jones et al., 1995), and a parametric model $g = f_{par}$ which gives a semiparametric estimator defined by $\tilde{f}_{HG}(\theta) = f_{par}(\theta)\hat{r}(\theta)$ (see,(Hjort and Glad, 1995),(Hirukawa and Sakudo, 2019) in the linear case).

To improve the quality of the semiparametric estimators HG, (Jones et al., 1999) proposed the so-called JSH estimator using the classical symmetric kernels. (Jones et al., 1999) are based on the same idea as (Hjort and Glad, 1995) but in the total nonparametric mode, that we correct the parametric part by a nonparametric correction what gives generally a same variance and a smaller bias. The principle of this estimator is to replace the function g by a semiparametric estimator $g = \tilde{f}_{HG}$. More recently, (Hirukawa and Sakudo, 2019) extends this idea to the asymmetric kernel density estimator with support $\mathbb{R}^+$. This study has shown that the JSH estimator reduces the order of magnitude of the bias from the $O(h^{-1})$ to $O(h^{-2})$ (h is the bandwidth parameter), which can even become unbiased under the right conditions.

The second objective of the paper is to improve the semiparametric HG ($\tilde{f}_{HG}$) estimator with von Mises kernel in the case of circular data by applying the JSH estimator, which will improve the order

of magnitude of the bias from $O(\nu^{-1})$ of the HG estimator to $O(\nu^{-2})$ in the case of the JSH estimator and will keeps the order of magnitude of the variance. The asymptotic and global properties for the proposed JSH estimator are established.

This paper is organized as follows. Section 2, briefly recalls the classical von Mises kernel density and its properties. Section 3, presents the semiparametric von Mises kernel density estimator with different forms of the parametric part. In Section 4, we introduce the HG estimator and its properties for circular data. Section 5, develops the JSH estimator and its properties for circular case. The unbiased cross validation (UCV) procedure is adapted for choosing the optimal bandwidth. We examine the performance of the JSH estimator using data generated from known circular distributions via the integrated squared error (ISE) and integrated squared bias (ISB) criterions in section 6. Section 7, illustrates an application of real data. Section 8, concludes the paper.

2 von Mises kernel density estimators (A brief review)

Let $\Theta_1, \Theta_2, \ldots, \Theta_n$ be independant observation from a count distribution with unknown probability density function (pdf) f defined on the support $[0, 2\pi)$. A von Mises circular kernel density estimator for f can be expressed as (see for example (Taylor, 2008)):

$$\begin{aligned}\hat{f}_{vM}(\theta;\nu) &= \frac{1}{n}\sum_{i=1}^{n} k_\nu(\theta-\Theta_i) \\ &= \frac{1}{n2\pi I_0(\nu)}\sum_{i=1}^{n}\exp\{\nu\cos(\theta-\Theta_i)\}, 0\le\theta<2\pi, \end{aligned} \tag{2}$$

where $\theta > 0$ is the estimation point (angle where the density is estimated), $\nu = \nu(n) > 0$ is the smoothing parameter that fulfills $\lim_{n\to\infty}\nu(n) = 0$, and $I_z(.)$ denotes the modified bessel function of order z. The asymptotic formulas of bias and variance are given by (Taylor, 2008)):

$$Bias(\hat{f}_{vM}(\theta;\nu)) = \frac{1}{4\nu}f''(\theta) + o(\nu^{-1}),$$

and

$$Var(\hat{f}_{vM}(\theta;\nu)) = \frac{\nu^{\frac{1}{2}}}{2n\sqrt{\pi}}f(\theta) + o\left(\frac{\nu^{\frac{1}{2}}}{n}\right).$$

The next section discuss the semiparametric approach for circular kernel density estimation.

3 Semiparametric von Mises kernel density estimator

In this section, we present the semiparametric estimator which is composed of two parts parametric ($g(x)$) and non parametric ($\hat{r}(x)$). Let $\Theta_1, \Theta_2, \ldots, \Theta_n$ be independant observation from a count distribution with unknown probability density function (pdf) f defined on the support $[0, 2\pi)$. The semiparametric kernel density estimator $\tilde{f}$ is given by:

$$\tilde{f}(\theta) := g(\theta)\hat{r}(\theta) := g(\theta)\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{k_\nu(\theta-\Theta_i)}{g(\Theta_i)}\right\}, \tag{3}$$

where g is the function that can be take many forms (constant, estimator density, parametric and semiparametric model) and $\hat{r}(x)$ serves as a correction factor. The form of the semiparametric von

Mises kernel density estimator that depends on a design point θ and smoothing parameter ν is as following:

$$\tilde{f}_{vM}(\theta) := g(\theta)\hat{r}(\theta) := g(\theta)\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{1}{2\pi I_0(\nu)}\exp\{\nu\cos(\theta-\Theta_i)\}}{g(\Theta_i)}\right\}, \tag{4}$$

The estimator $\tilde{f}$ changes with the change of the function g, several cases of estimators can be derived according to the function g form.

3.1 Function g is an uniform density

When the function g is an uniform density i.e $g(\theta) \equiv 1/2\pi$, the estimator $\tilde{f}_{vM}$ is reduced to a classical estimator $\tilde{f}_{vM,Classic}$ defined by (2), as shown below:

$$\begin{aligned}
\tilde{f}_{vM}(\theta) &= g(\theta)\hat{r}(\theta), \\
&= g(\theta)\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{1}{2\pi I_0(\nu)}\exp\{\nu\cos(\theta-\Theta_i)\}}{g(\Theta_i)}\right\}, \\
&= \left\{\frac{1}{2\pi}\right\}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{1}{2\pi I_0(\nu)}\exp\{\nu\cos(\theta-\Theta_i)\}}{\frac{1}{2\pi}}\right\}, \\
&= \frac{1}{n2\pi I_0(\nu)}\sum_{i=1}^{n}\exp\{\nu\cos(\theta-\Theta_i)\}, \\
&= \tilde{f}_{vM,Classic}(\theta).
\end{aligned} \tag{5}$$

The properties of $\tilde{f}_{vM,Classic}$ given by (5) are developed by (Taylor, 2008), (Marzio et al., 2009, 2011).

3.2 Function g is a kernel density estimator

When the function g is a kernel density estimator i.e $g(\theta) = \hat{f}_{vM}(\theta)$, the estimator $\tilde{f}_{vM}$ becomes the (Jones et al., 1995), type fully nonparametric MBC estimator (JLN) $\tilde{f}_{vM,JLN}$ given by:

$$\tilde{f}_{vM,JLN}(\theta) = \hat{f}_{vM}(\theta)\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{1}{2\pi I_0(\nu)}\exp\{\nu\cos(\theta-\Theta_i)\}}{\hat{f}_{vM}(\Theta_i)}\right\}. \tag{6}$$

This estimator is introduced by (Jones et al., 1995), and considered by (Hirukawa, 2010), (Hirukawa and Sakudo, 2014) for asymmetric kernel density estimation, (Zougab and Adjabi, 2016) for heavy tailed data using generalized Birnbaum-Saunders kernels, (Harfouche et al., 2018) and (Harfouche et al., 2020) for discrete situations. More recently, (Bedouhene and Zougab, 2019) have examined the JLN-von Mises kernel estimator for circular data.

3.3 Function g is a parametric model

When g belongs to a parametric family, the function g can takes different circular parametric models f_{par} (von Mises, Projected Normal, Wrapped distributions, ...), then $\tilde{f}_{vM}$ reduces to the (Hjort and Glad, 1995) type semiparametric MBC estimator $\tilde{f}_{vM,HG}$, studied by (Hagmann and Scaillet, 2007) using asymmetric kernel and by (Kokonendji et al., 2009) for discrete data. Note that the HG-von Mises circular kernel estimator and its properties will be presented in section 4.

3.4 Function g is semiparametric estimator $\tilde{f}_{vM,HG}$

In this case, (Jones et al., 1999) proposed a new JSH estimator which improves the quality of the estimator as well as the bias compared to the HG estimators already mentioned.

The JSH estimator principle, is to consider the semiparametric $\tilde{f}_{vM}$ estimator and replace the function g by the $\tilde{f}_{vM,HG}$ estimator, which improves the convergence of the bias towards $o(\nu^{-2})$ whatever the scenario. The section 5, discusses in details the estimator and its properties.

4 The HG estimator and its properties

The semiparametric HG von Mises kernel estimator $\tilde{f}_{vM,HG}$ is given as follows:

$$\tilde{f}_{vM,HG}(\theta) = f(\theta;\hat{\vartheta}) \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\frac{1}{2\pi I_0(\nu)} \exp\{\nu \cos(\theta - \Theta_i)\}}{f(\Theta;\hat{\vartheta})} \right\}. \tag{7}$$

Where $f(\theta;\vartheta)$ is the pdf of the von Mises distribution $vM(\vartheta)$ with $\vartheta = (\mu, k)$. the von Mises distribution is expressed as:

$$f(\theta;\mu,k) = \frac{1}{2\pi I_0(k)} \exp\left(k\cos(\theta-\mu)\right), \;\; 0 \le \theta < 2\pi,$$

and $\hat{\vartheta} = (\hat{\mu}, \hat{k})$ is the estimated parameter of the $\vartheta = (\mu, k)$ by the maximum likelihood (ML) method.

4.1 Asymptotic properties

To approximate the bias and variance of HG-vM kernel estimator, we assume that:
Set $\vartheta_0 = (\mu_0, k_0)$ the value which minimizes the Kullback–Leibler distance of $f(\theta;\vartheta)$ from the true $f(\theta)$, we also denote $f_0(\cdot) = f(\cdot,\vartheta_0)$,

A1. f has four continuous and bounded derivatives.
A2. The sequence of bandwidths $\nu = \nu(n)$, satisfies $\nu \to \infty$ and $\nu^{1/2}/n \to 0$ when $n \to \infty$

Theorem 4.1. *Under the Assumptions 1-2, the bias and variance of* HG-vM *kernel estimator defined by (7), for a given $\theta \in [0, 2\pi)$, are given by:*

(i) The bias of HG-vM *kernel estimator is given by:*

$$Bias\left(\tilde{f}_{vM,HG}(\theta)\right) = \frac{1}{4\nu} f(\theta;\vartheta_0) r''(\theta) + o(\nu^{-1}), \tag{8}$$

(ii) The variance of HG-vM *kernel estimator is given by:*

$$Var\left(\tilde{f}_{vM,HG}(\theta)\right) = \frac{\nu^{1/2}}{2n\sqrt{\pi}} f(\theta) + o\left(\frac{\nu^{1/2}}{n}\right). \tag{9}$$

Proof. *A Taylor expansion gives*

$$\begin{aligned}
\frac{f(\theta;\widehat{\vartheta})}{f(\Theta_i;\widehat{\vartheta})} &= \exp\{\log f(\theta;\widehat{\vartheta}) - \log f(\Theta_i;\widehat{\vartheta})\} \\
&\doteq \frac{f_0(\theta)}{f_0(\Theta_i)} + \frac{f_0(\theta)}{f_0(\Theta_i)} \{u_0(\theta) - u_0(\Theta_i)\}^{\intercal} \left(\widehat{\vartheta} - \vartheta_0\right) \\
&= \frac{f_0(\theta)}{f_0(\Theta_i)} \left[1 - \{u_0(\Theta_i) - u_0(\theta)\}^{\intercal} \left(\widehat{\vartheta} - \vartheta_0\right)\right],
\end{aligned}$$

where $u_0(\theta) = \partial \log f(\theta;\vartheta_0)/\partial\vartheta$. *The semiparametric estimator (7) can be approximated as follows:*

$$\widehat{f}_{\nu}^{HG-VM}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2\pi I_0(\nu)} \exp\{\nu\cos(\theta-\Theta_i)\}\frac{f_0(\theta)}{f_0(\Theta_i)}\left[1-\{u_0(\Theta_i)-u_0(\theta)\}^{\intercal}\left(\widehat{\vartheta}-\vartheta_0\right)\right]. \quad (10)$$

Using the representation (10), the properties of the von Mises random variable and the Taylor development when $\nu \to \infty$ *(see (Mardia and Jupp, 2009)) and the same computational steps of (Hjort and Glad, 1995), we obtain the results given in the theorem 4.1.* □

Note that the asymptotic variance of the HG-VM estimator is similar to that of the classical von Mises estimator.

Corollary 1. *The criterion to use for the global property is the mean integrated squared error* (MISE) *defined as:*

$$\begin{aligned} MISE\left(\tilde{f}_{vM,HG}(\theta)\right) &= \int_0^{2\pi} \mathbb{E}\left(\tilde{f}_{vM,HG}(\theta)-f(\theta)\right)^2 d\theta \\ &= \int_0^{2\pi} MSE\tilde{f}_{vM,HG}(\theta)d\theta \\ &= \int_0^{2\pi} bias^2\left(\tilde{f}_{vM,HG}(\theta)\right)d\theta + \int_0^{2\pi} Var\left(\tilde{f}_{vM,HG}(\theta)\right)d\theta \\ &= \frac{1}{16\nu^2}\int_0^{2\pi}\{f_0(\theta)r''(\theta)\}^2 d\theta + \frac{\nu^{1/2}}{2n\sqrt{\pi}} + o\left(\nu^{-2}+\frac{\nu^{1/2}}{n}\right). \end{aligned} \quad (11)$$

By minimizing (11) in the bandwidth ν, *we obtain the optimal value:*

$$\nu^*_{vM,HG} = \left\{\frac{n\sqrt{\pi}}{2}\int_0^{2\pi}\{f_0(\theta)r''(\theta)\}^2 d\theta\right\}^{2/5}. \quad (12)$$

The optimal $MISE^*_{vM,HG}$ *of HG estimator is obtained by remplacing the* $\nu^*_{vM,HG}$ *given by (12) in* $MISE\left(\tilde{f}_{vM,HG}(\theta)\right)$ *given in (11)*

$$MISE^*\left(\tilde{f}_{vM,HG}(\theta)\right) = \frac{1}{2\sqrt{\pi}}\left(\frac{\sqrt{\pi}}{2}\int_0^{2\pi}\{f_0(\theta)r''(\theta)\}^2 d\theta\right)^{1/5} \left(1+\frac{1}{4}\left\{\int_0^{2\pi}\{f_0(\theta)r''(\theta)\}^2 d\theta\right\}^{-1}\right) n^{-4/5}.$$

5 The JSH estimator

Based on the same idea of (Jones et al., 1999) and (Hirukawa and Sakudo, 2019), JSH-von Mises kernel density estimator will be defined as follows:

$$\tilde{f}_{vM,JSH}(\theta) = \tilde{f}_{vM,HG}(\theta)\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{1}{2\pi I_0(\nu)}\exp\{\nu\cos(\theta-\Theta_i)\}}{\tilde{f}_{vM,HG}(\Theta_i)}\right\}, \quad (13)$$

where $\tilde{f}_{vM,HG}(\theta) = f(\theta)\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{1}{2\pi I_0(\nu)}\exp\{\nu\cos(\theta-\Theta_i)\}}{f(\Theta_i)}\right\}$,
$f(\theta)$ is the parametric model, which can takes different distributions. In the case where this model takes the form of a uniform distribution, the JSH estimator is reduced into an MBC JLN estimator, which explains why the JLN estimator is a special case of the general estimator JSH.
In this work, the parametric model exploited is the von-Mises law of parameters μ and k, expressed as:

$$f(\theta;\mu,k) = \frac{1}{2\pi I_0(k)}\exp\left(k\cos(\theta-\mu)\right),\ \ 0 \le \theta < 2\pi,$$

and $\hat{\mu}$ and $\hat{k}$ are the estimated parameters of the μ and k respectively by the maximum likelihood (ML) method.

5.1 Asymptotic properties

The asymptotic properties of the JSH-vM kernel estimator requires certain conditions and assumptions.

Set μ_0 and k_0 the values which minimizes the Kullback–Leibler distance of $f(\theta;\mu,k)$ from the true $f(\theta)$. We also denote $r_0(.) = f(.)/f(.;\mu_0,k_0)$

A3. $\{\Theta_i\}_{i=1}^n$ are i.i.d random variables drown from a univariate distribution having a density f with support $[0,2\pi)$.
A4. For a given design point $\theta \in [0,2\pi)$, $f(\theta)$, $f(\theta;\mu_0,k_0) > 0$, and $r_0(\theta)$ has four continuous and bounded derivatives in the neighborhood of θ.

Note that these assumptions have been discussed in (Hirukawa and Sakudo, 2019).

Theorem 5.1. *Under the Assumptions 1-4, the bias and variance of* JSH-vM *kernel estimator defined by (5.1), for a given $\theta \in [0,2\pi)$, are given by:*

(i) The bias of JSH-vM *kernel estimator is given by:*

$$Bias\left(\tilde{f}_{vM,JSH}(\theta)\right) = -f(\theta)\,q\,(\theta,r_0)\ \frac{1}{\nu^2} + o\left(\frac{1}{\nu^2}\right), \tag{14}$$

where $q(\theta,r_0) = \left(\frac{r_0''(\theta)}{4r_0(\theta)}\right)''$.
(ii) The variance of JSH-vM *kernel estimator is given by:*

$$Var\left(\tilde{f}_{vM,JSH}(\theta)\right) = \frac{\nu^{1/2}}{2n\sqrt{\pi}}f(\theta) + o\left(\frac{\nu^{1/2}}{n}\right). \tag{15}$$

Proof. *It is easy to show that when $\nu \longrightarrow \infty$ and $\nu^{1/2}/n \longrightarrow 0$ when $n \longrightarrow \infty$,*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{1}{2\pi I_0(\nu)}\exp\{\nu\cos(\theta-\Theta_i)\}}{g(\Theta_i;\widehat{\vartheta})}\right] \simeq r(\theta) + \frac{1}{4\nu}r''(\theta), \tag{16}$$

and

$$\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{1}{2\pi I_0(\nu)}\exp\{\nu\cos(\theta-\Theta_i)\}}{g(\Theta_i;\widehat{\vartheta})}\right]\simeq\frac{\nu^{1/2}}{2n\sqrt{\pi}}\frac{r(\theta)}{g_0(\theta)}. \tag{17}$$

where $\simeq$ means asymptotically equal.
Combining the proof of Theorem 2 proposed by (Bedouhene and Zougab, 2019) and the approximations (16) and (17), we establish the results of Theorem 5.1. □

Corollary 2. *The criterion to use for the global property is the mean integrated squared error* (MISE)*, defined as:*

$$\begin{aligned} MISE\left(\tilde{f}_{vM,JSH}(\theta)\right) &= \int_0^{2\pi} bias^2\left(\tilde{f}_{vM,JSH}(\theta)\right)d\theta + \int_0^{2\pi} Var\left(\tilde{f}_{vM,JSH}(\theta)\right)d\theta \\ &= \frac{1}{\nu^4}\mathbb{E}\left[f(\Theta)\,q^2(\Theta,r_0)\right]+\frac{\nu^{1/2}}{2n\sqrt{\pi}}+o\left(\nu^{-4}+\frac{\nu^{1/2}}{n}\right). \end{aligned} \tag{18}$$

The optimal bandwidth minimizing the corresponding MISE *(18) is such that*

$$\nu^*_{JSH}=\left\{16\,n\,\sqrt{\pi}\mathbb{E}\left[f(\Theta)\,q^2(\Theta,r_0)\right]\right\}^{2/9}. \tag{19}$$

Therefore, the optimal MISE *becomes*

$$MISE^*\left(\tilde{f}_{vM,JSH}(\theta)\right)=\frac{1}{(16\sqrt{\pi})^{8/9}}\;\left\{\mathbb{E}\left[f(\Theta)\,q^2(\Theta,r_0)\right]\right\}^{1/9}n^{-8/9}. \tag{20}$$

In practice, the bandwidth parameter ν^*_{JSH} cannot be employed because it's depend on the unknown density f and its derivatives. For this, we present in the next section the unbiased cross validation UCV method for selecting the smoothing parameter.

6 Bandwidth choice by UCV method

We adopt in this section the popular unbiased cross validation (UCV) method introduced by (Rudemo, 1982) and (Bowman, 1984), and recently applied by several authors (See (Zougab and Adjabi, 2016; Harfouche et al., 2018, 2020) and (Bedouhene and Zougab, 2019) for the MBC technique, (Hagmann and Scaillet, 2007) for the semiparametric MBC technique). The optimal bandwidth ν by UCV method for a given estimator $\tilde{f}_{vM,JSH}$ is obtained by:

$$\nu_{UCV}=\arg\min_{\nu} UCV(\nu),$$

where

$$\begin{aligned} UCV(\nu) &= \frac{1}{n^2}\int_0^{2\pi}\tilde{f}^2_{vM,HG}(\theta)\left(\sum_{i=1}^{n}\frac{k_\nu(\theta-\Theta_i)}{\tilde{f}_{vM,HG}(\Theta_i)}\right)^2 d\theta \\ &- \frac{2}{n(n-1)}\sum_i\sum_{j\neq i}k_\nu(\theta-\Theta_i)\frac{\tilde{f}_{vM,HG}(\Theta_i)}{\tilde{f}_{vM,HG}(\Theta_j)}. \end{aligned} \tag{21}$$

7 Simulation study

This part is devoted to a simulation study, which consists to evaluate the performance of the $\tilde{f}_{vM}$, $\tilde{f}_{vM-JLN}$, $\tilde{f}_{vM-HG}$ and $\tilde{f}_{vM-JSH}$ estimators. This simulation study is based on 100 replications for samples sizes $n = 10$, 50, 100 and 200 for 20 circular models (distributions), that were used by (Oliveira et al., 2012), (García-Portugués, 2013) and recently (Bedouhene and Zougab, 2019, 2020). The models are classified into four sets, each with its complexity:

1. Simple models: circular uniform (M1); von Mises (M2); wrapped Normal (M3); cardioid (M4); wrapped Cauchy (M5) and wrapped skew-Normal (M6).
2. Two components models: von Mises mixtures (M7, M8 and M9); mixture of von Mises and wrapped Cauchy (M10).
3. Models with more than two components: von Mises mixtures with three components (M11, M12 and M13); von Mises mixture with four components (M14); mixture of wrapped Cauchy, wrapped Normal, von Mises and wrapped skew-Normal (M15); von Mises mixture with five components (M16).
4. Other complex models: mixture of cardioid and wrapped Cauchy (M17); mixture of von Mises (M18 and M19); mixture of two wrapped skew-Normal and two wrapped Cauchy (M20).

For the comparison, we used the classical, JLN, HG and JSH vM kernel estimators and the UCV method for bandwidth parameter selection. Note that, The parametric part $f(\theta;\mu,k)$ of the $\tilde{f}_{HG,vM}$ estimator is considered as vM distribution of μ and k parameters $vM(\mu,k)$. The parameters μ and k were estimated by maximum likelihood (ML). We examine the performances of the estimators via integrated squared error (ISE) and integrated squared bias (ISB) given respectively by:

$$ISE := \int_0^{2\pi} \Big(\tilde{f}(\theta) - f(\theta)\Big)^2 d\theta \tag{22}$$

and

$$ISB := \int_0^{2\pi} \Big(\mathbb{E}(\tilde{f}(\theta)) - f(\theta)\Big)^2 d\theta \tag{23}$$

Tables 1, 2, 3 and 4 show the average ISE of the vM kernel estimators. We can observe that,

- The means of ISE decrease as sample size n increases for the all estimators, which indicates that our estimators are consistent.
- The vM-JLN estimator performs better than the other competitors for models M5, M6, M7, M10, M11, M12, M13, M14, M16, M17, M18 and M20 and for a samples size $n = 100$ and $n = 200$ except model M10 for $n = 100$.
- The vM-JLN estimator performs better than the other estimators for models M6, M7, M8, M11, M13, M14, M16 and M17 and for a sample size $n = 50$.
- For a sample size $n = 10$ the standard vM kernel estimator is more efficient for the models M1, M4, M10, M11, M12, M16, M18, M19 and M20.
- For the rest of models M2, M3, M9 and M15 the perfomance of estimators are mixed depending on the sample size.

Tables 5 and 6 show the average ISB of the vM kernel estimators. We can observe that:

Table 1: Average integrated squared error ($\overline{ISE}$) and their standard deviation between parentheses based on 100 replications for 20 models with sample size $n = 10$

$n = 10$	$\tilde{f}_{vM}$ (Sd_{vM})	$\tilde{f}_{vM-JLN}$ (Sd_{vM-JLN})	$\tilde{f}_{vM-HG}$ (Sd_{vM-HG})	$\tilde{f}_{vM-JSH}$ (Sd_{vM-JSH})
M1	**0.028504** (0.040126)	0.044624 (0.048535)	0.085598 (0.042796)	0.065051 (0.046547)
M2	0.044730 (0.027727)	0.054532 (0.038296)	**0.041635** (0.032852)	0.056422 (0.043456)
M3	0.080912 (0.063078)	0.077513 (0.045428)	0.066713 (0.048654)	**0.061435** (0.029901)
M4	**0.031643** (0.018711)	0.035378 (0.020203)	0.032978 (0.030695)	0.036326 (0.021759)
M5	0.175951 (0.039888)	0.148754 (0.050878)	0.165277 (0.046911)	**0.136418** (0.050530)
M6	0.081913 (0.030951)	0.076863 (0.024004)	**0.071529** (0.019551)	0.075294 (0.019237)
M7	0.056374 (0.024210)	**0.054870** (0.019606)	0.067062 (0.040294)	0.057183 (0.026707)
M8	0.073435 (0.034446)	0.065588 (0.031154)	0.068817 (0.030425)	**0.062101** (0.033644)
M9	0.038088 (0.022375)	**0.037386** (0.022005)	0.058252 (0.029270)	0.043407 (0.026101)
M10	**0.101301** (0.050482)	0.107924 (0.057971)	0.104519 (0.058812)	0.109033 (0.062352)
M11	**0.051501** (0.012703)	0.057676 (0.014236)	0.069096 (0.033663)	0.067695 (0.032629)
M12	**0.066814** (0.038738)	0.074644 (0.042631)	0.101371 (0.039642)	0.104655 (0.040974)
M13	0.097510 (0.034632)	0.096302 (0.034424)	0.126583 (0.067811)	**0.095344** (0.046333)
M14	0.093109 (0.021393)	**0.081828** (0.030255)	0.116401 (0.029262)	0.091273 (0.034697)
M15	0.018991 (0.018817)	0.025672 (0.026108)	0.071526(0.068303)	0.066495 (0.051130)
M16	**0.085333** (0.014261)	0.090423 (0.014995)	0.134788 (0.055393)	0.108673 (0.028095)
M17	0.135170 (0.053391)	0.119433 (0.044483)	**0.111107** (0.019398)	0.121193 (0.047785)
M18	**0.074270** (0.037493)	0.076313 (0.037901)	0.102535 (0.105780)	0.092129 (0.076978)
M19	**0.080081** (0.027177)	0.085918 (0.036308)	0.090258 (0.024831)	0.094315 (0.035757)
M20	**0.121919** (0.029699)	0.127653 (0.039151)	0.150788 (0.047860)	0.143430 (0.040143)

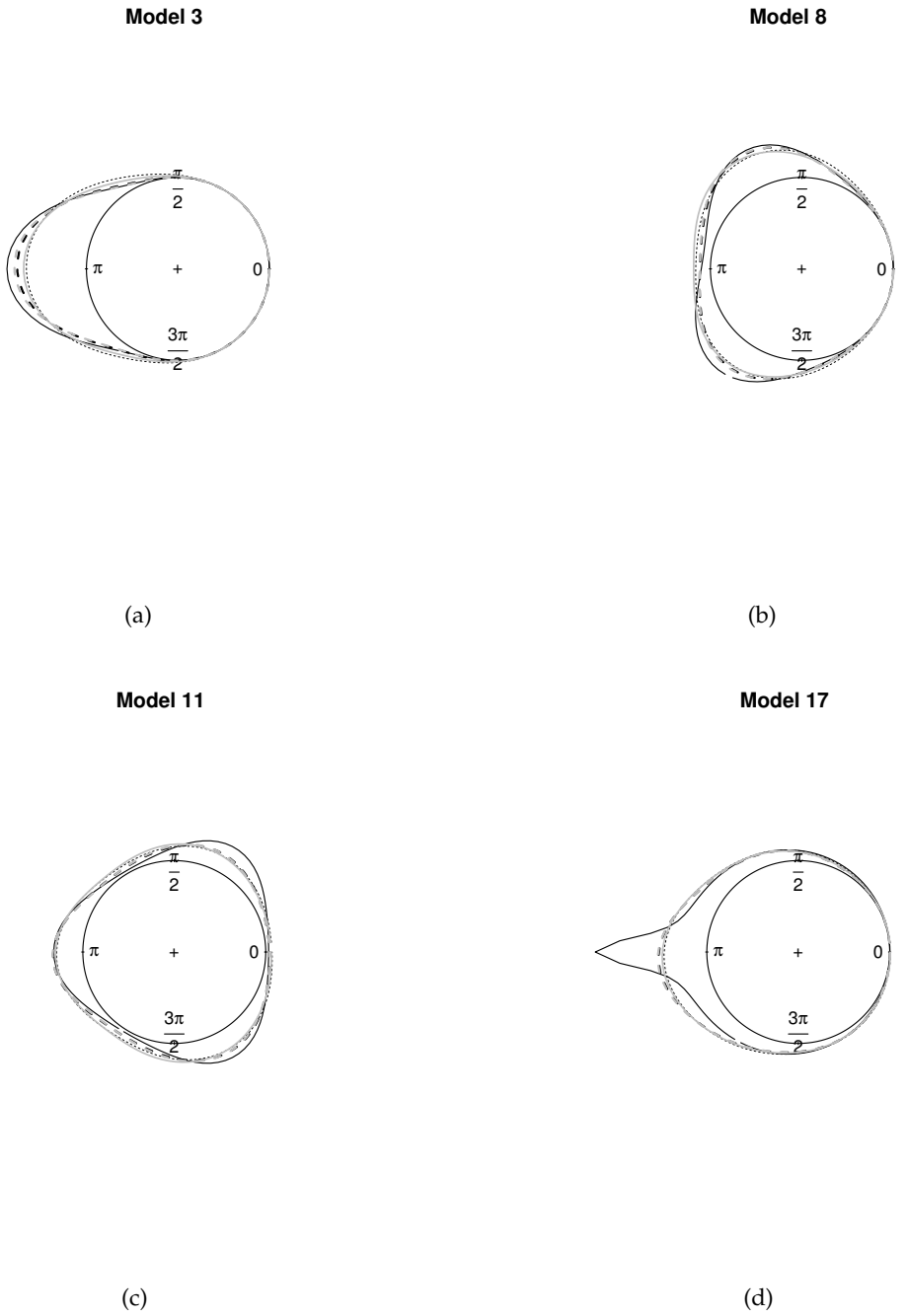
- For all estimators, the means of ISB based on 100 simulations decrease as sample size n increases.
- For all samples size the vM-JLN and vM-JSH kernel estimators outperform the classical vM and vM-HG kernel estimators except models (M1, M2, M4, M10 and M12 for $n = 10$), (M1, M2, M4 and M15 for $n = 50$) and (M1 and M2 for $n = 100$ et 200).
- The performance of vM-JLN and vM-JSH kernel estimators are mixed depending on the models, this is explained by the fact that the vM-JLN estimator is a special case of vM-JSH kernel estimators

The figure 1 presents the plot of pdf M3, M8, M11 and M17 models for each distribution family using vM, JLN-vM, HG-vM and JSH-vM estimators with UCV approach for bandwidth choice based on sample size $n = 200$ for one application. The plots show that in general the smoothing quality is satisfactory for all models. The estimators were able to reproduce the uni and several modes of the considered models, except for the model 17.

8 Illustration with real data

In this section, we illustrate the application of $\tilde{f}_{vM}$, $\tilde{f}_{vM-JLN}$, $\tilde{f}_{vM-HG}$ and $\tilde{f}_{vM-JSH}$ estimators in practice, we have analysed two real data sets.

Exemple 1. (Time series of flare azimuths): this data present a time series of measurements obtained from an experiment, to assess the relative stability of flare-projectile assemblies. A flare, attached to a projectile, is launched upward from a launch point O in a fixed direction. At some



(a) (b)

(c) (d)

Figure 1: True pdf and vM kernel estimators for models 3, 8, 11 and 17 with $n = 200$. Black solid line: true density; Black dashed line: vM estimator; Black Dotted line: vM-JLN estimator; Grey solide line: vM-HG estimator and Grey dashed line: vM-JSH estimator.

Table 2: Average integrated squared error ($\overline{ISE}$) and their standard deviation between parentheses based on 100 replications for 20 models with sample size $n = 50$

$n = 50$	$\tilde{f}_{vM}$ (Sd_{vM})	$\tilde{f}_{vM-JLN}$ (Sd_{vM-JLN})	$\tilde{f}_{vM-HG}$ (Sd_{vM-HG})	$\tilde{f}_{vM-JSH}$ (Sd_{vM-JSH})
M1	**0.001866** (0.003489)	0.006231 (0.005917)	0.013917 (0.005010)	0.011227 (0.005081)
M2	0.010313 (0.008561)	0.011044 (0.009943)	**0.008018** (0.006045)	0.011511 (0.009034)
M3	0.041835 (0.014696)	0.017860 (0.008250)	0.028309 (0.011004)	**0.014193** (0.007108)
M4	**0.008945** (0.008045)	0.014548 (0.011967)	0.014283 (0.015017)	0.018999 (0.013630)
M5	0.154229 (0.019441)	0.109994 (0.028114)	0.139815 (0.017119)	**0.103537** (0.029125)
M6	0.058575 (0.009952)	**0.047698** (0.011597)	0.048187 (0.006382)	0.051759 (0.016907)
M7	0.019584 (0.007133)	**0.014356** (0.006791)	0.022576 (0.005455)	0.014708 (0.007069)
M8	0.026306 (0.012476)	**0.018937** (0.014137)	0.034317 (0.021504)	0.020011 (0.017085)
M9	0.009666 (0.006145)	0.009945 (0.004733)	0.013562 (0.007814)	**0.009470** (0.004791)
M10	0.052599 (0.011198)	0.042107 (0.013285)	0.044380 (0.010356)	**0.042057** (0.014870)
M11	0.031501 (0.009060)	**0.024606** (0.010623)	0.036343 (0.011329)	0.025054 (0.010273)
M12	0.017454 (0.011790)	0.015778 (0.007861)	0.020647 (0.010969)	**0.014755** (0.005718)
M13	0.042008 (0.006729)	**0.031182** (0.006917)	0.048533 (0.008577)	0.032526 (0.007163)
M14	0.053977 (0.007150)	**0.040523** (0.007811)	0.064132 (0.015627)	0.043959 (0.010567)
M15	**0.009173** (0.004106)	0.012301 (0.006091)	0.019124 (0.006767)	0.017794 (0.008228)
M16	0.074266 (0.007694)	**0.062254** (0.009584)	0.074199 (0.008923)	0.063870 (0.010247)
M17	0.081349 (0.008914)	**0.071710** (0.011465)	0.078328 (0.011586)	0.071803 (0.013159)
M18	0.040710 (0.005906)	0.042476 (0.006280)	**0.040324** (0.005152)	0.043975 (0.007533)
M19	0.035764 (0.006904)	**0.035098** (0.008540)	0.041890 (0.014962)	0.035811 (0.009832)
M20	0.080598 (0.005963)	**0.072905** (0.007403)	0.086780 (0.005067)	0.075472 (0.007073)

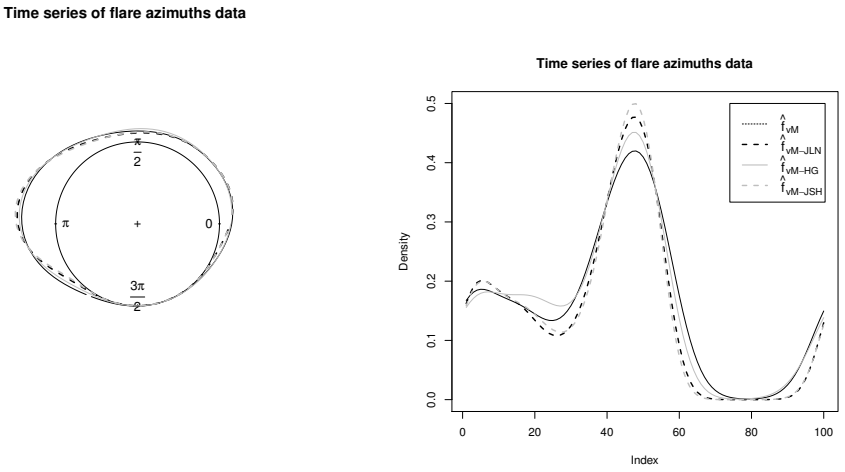
point P in space, the flare commences burning. The azimuth of P relative to O gives an indication of the variability of the assembly as more and more trials are conducted with it. The data shown are based on 60 successive launches, (see, (Fisher, 1995)).

Exemple 2. (Long-axis orientations of feldspar laths): a data of 133 measurements of feldspar laths in basalt reported by (Smith, 1988) and presented by (Fisher, 1995).

For analyzing these data sets, we applied four estimators $\tilde{f}_{vM}$, $\tilde{f}_{vM-JLN}$, $\tilde{f}_{vM-HG}$ and $\tilde{f}_{vM-JSH}$ to estimate the probability density function with vM circular kernel. The UCV technique is employed for bandwidth choice. For the $vM-HG$ kernel estimator, the parameters μ and k of parametric model are estimated by ML method. The values of $\hat{\mu}$ and $\hat{k}$ are given in table 7.

The figures 2 and 3 show the linear and circular estimators ($\tilde{f}_{vM}$, $\tilde{f}_{vM-JLN}$, $\tilde{f}_{vM-HG}$ and $\tilde{f}_{vM-JSH}$) for Time series of flare azimuths data with sample size $n = 60$ and Long-axis orientations of feldspar laths data with sample size $n = 133$, where the solide and dashed lines in black represent the $\tilde{f}_{vM}$ and $\tilde{f}_{vM-JLN}$ estimators, the solide and dashed lines in grey represent $\tilde{f}_{vM-HG}$ and $\tilde{f}_{vM-JSH}$ estimators respectively.

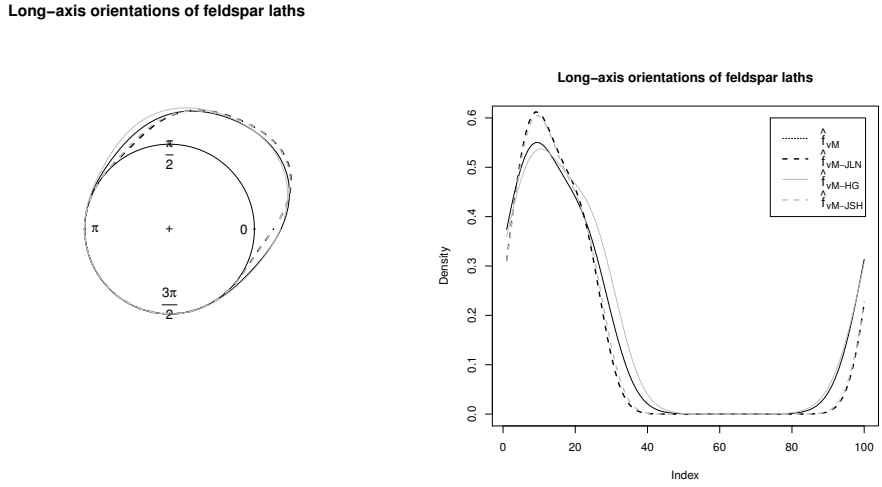
From the figures 2 and 3, we can see that all estimators are capable of reproducing the unimodality of these data sets. We also note that the smoothing quality is satisfactory and almost similar for the four estimators except in certain regions.



(a) Linear plot for Time series of flare azimuths data.

(b) Circular plot for Time series of flare azimuths data.

Figure 2: vM kernel estimator for Time series of flare azimuths data with sample size $n = 60$. Black solid line: vM estimator; Black dashed line: JLN-vM estimators; Grey solid line: HG-vM; Grey dashed line: JSH-vM.



(a) circular plot for Long-axis orientations of feldspar laths data.

(b) linear plot for Long-axis orientations of feldspar laths data.

Figure 3: vM kernel estimator for Long-axis orientations of feldspar laths data with sample size $n = 133$. Black solid line: vM estimator; Black dashed line: JLN-vM estimators; Grey solid line: HG-vM; Grey dashed line: JSH-vM.

Table 3: Average integrated squared error ($\overline{ISE}$) and their standard deviation between parentheses based on 100 replications for 20 models with sample size $n = 100$

$n = 100$	$\tilde{f}_{vM}$ (Sd_{vM})	$\tilde{f}_{vM-JLN}$ (Sd_{vM-JLN})	$\tilde{f}_{vM-HG}$ (Sd_{vM-HG})	$\tilde{f}_{vM-JSH}$ (Sd_{vM-JSH})
M1	**0.001457** (0.002025)	0.003106 (0.003671)	0.007134 (0.002430)	0.008218 (0.002656)
M2	0.007618 (0.007692)	0.006788 (0.005903)	**0.005093** (0.005007)	0.006385 (0.005543)
M3	0.037664 (0.003969)	0.013694 (0.002814)	0.026014(0.004538)	**0.010504** (0.003361)
M4	**0.006219** (0.004247)	0.006670 (0.005332)	0.006363 (0.003905)	0.008482 (0.005162)
M5	0.154077 (0.017058)	**0.106388** (0.017568)	0.138480 (0.016540)	0.119237 (0.041765)
M6	0.057178 (0.010608)	**0.047702** (0.016986)	0.047161 (0.007705)	0.052374 (0.019852)
M7	0.017531 (0.008879)	**0.010608** (0.006796)	0.021808 (0.011133)	0.011076 (0.007525)
M8	0.021781 (0.005358)	**0.011615** (0.005049)	0.027436 (0.010461)	0.011731 (0.005784)
M9	0.006189 (0.003258)	0.005196 (0.002406)	0.007646 (0.004701)	**0.004875** (0.003331)
M10	0.042288 (0.008199)	0.033879 (0.005501)	0.036614 (0.006635)	**0.033876** (0.005801)
M11	0.023124 (0.002969)	**0.012971** (0.002866)	0.028574 (0.003982)	0.013945 (0.003178)
M12	0.010648 (0.005060)	**0.009277** (0.004555)	0.013961 (0.004866)	0.009663 (0.004217)
M13	0.040233 (0.005926)	**0.024880** (0.005162)	0.047879 (0.008810)	0.026472 (0.005187)
M14	0.048432 (0.006302)	**0.033535** (0.007966)	0.055748 (0.005895)	0.035982 (0.007658)
M15	**0.007635** (0.003075)	0.008174 (0.003857)	0.008748 (0.003335)	0.008276 (0.004931)
M16	0.064286 (0.003487)	**0.053544** (0.004683)	0.068856 (0.003156)	0.054935 (0.003064)
M17	0.079762 (0.006565)	**0.066913** (0.007029)	0.076166 (0.008493)	0.067919 (0.009618)
M18	0.035264 (0.005775)	**0.034528** (0.005536)	0.036797 (0.005094)	0.036755 (0.005704)
M19	0.029451 (0.005088)	0.026673 (0.004411)	0.030842 (0.007298)	**0.026669** (0.004697)
M20	0.072511 (0.004133)	**0.062957** (0.004733)	0.075288 (0.003609)	0.063519 (0.003887)

9 Conclusion

In this work, we extended the application of the semiparametric estimators HG and JSH, for the estimation of the probability density of circular data with von Mises kernel. We have shown that the JSH estimator improves the order of magnitude of the bias, whatever the parametric model chosen. We also evaluated the performance of the proposed estimator (JSH) by a comparative study with the classical estimator $\hat{f}_{vM}$, the estimator JLN $\tilde{f}_{JLN}$ and the semiparametric estimator $\tilde{f}_{HG}$. Our study showed that the proposed JSH estimator and the JLN estimator are more efficient than the other two estimators in terms of ISE and ISB. The two estimators JLN and JSH are almost similar, this is explained by the fact that the estimator JLN is a special case of JSH.

Acknowledgments

We sincerely thank the editor and anonymous referee for their valuable comments that allowed us to improve this article.

References

Amiri, A., B. Thiam, and T. Verdebout (2017). On the estimation of the density of a directional data stream. *Scandinavian Journal of Statistics 44*(1), 249–267.

Table 4: Average integrated squared error ($\overline{ISE}$) and their standard deviation between parentheses based on 100 replications for 20 models with sample size $n = 200$

$n = 200$	$\tilde{f}_{vM}$ (Sd_{vM})	$\tilde{f}_{vM-JLN}$ (Sd_{vM-JLN})	$\tilde{f}_{vM-HG}$ (Sd_{vM-HG})	$\tilde{f}_{vM-JSH}$ (Sd_{vM-JSH})
M1	**0.000840** (0.001765)	0.002012 (0.002658)	0.005313 (0.002595)	0.004798 (0.002703)
M2	0.005359 (0.004786)	0.004924 (0.003438)	**0.002080** (0.001720)	0.004924 (0.002783)
M3	0.033638 (0.006134)	0.010285 (0.002248)	0.021982 (0.004420)	**0.006965** (0.001297)
M4	0.003911 (0.001736)	0.003399 (0.002243)	**0.003174** (0.001561)	0.003761 (0.002169)
M5	0.147711 (0.003426)	**0.098190** (0.005221)	0.133369 (0.003885)	0.104548 (0.042274)
M6	0.055673 (0.005007)	**0.039312** (0.007453)	0.045489 (0.002868)	0.042482 (0.010633)
M7	0.014555 (0.003525)	**0.006889** (0.002733)	0.017779 (0.005106)	0.007382 (0.003266)
M8	0.020626 (0.002848)	0.009474 (0.002975)	0.023749 (0.006376)	**0.008934** (0.004090)
M9	0.005869 (0.003752)	0.004805 (0.003233)	0.006548 (0.003732)	**0.004541** (0.002573)
M10	0.040668 (0.005104)	**0.031890** (0.005037)	0.035392 (0.003461)	0.031992 (0.005620)
M11	0.022043 (0.003397)	**0.011385** (0.003413)	0.025781 (0.005939)	0.011667 (0.004046)
M12	0.007827 (0.001894)	**0.004541** (0.002231)	0.012173 (0.002798)	0.004888 (0.002422)
M13	0.036175 (0.003118)	**0.020665** (0.002707)	0.040786 (0.004169)	0.021381 (0.002714)
M14	0.046397 (0.002191)	**0.029773** (0.002206)	0.050799 (0.002620)	0.031039 (0.002747)
M15	0.006984 (0.002966)	0.006893 (0.003779)	0.006948 (0.004536)	**0.006829** (0.004414)
M16	0.061072 (0.002369)	**0.049232** (0.003386)	0.064606 (0.003178)	0.050812 (0.004028)
M17	0.079144 (0.007085)	**0.066755** (0.007067)	0.075603 (0.006609)	0.067226 (0.011314)
M18	0.034933 (0.003938)	**0.032185** (0.003686)	0.033361 (0.002792)	0.032942 (0.004285)
M19	0.028672 (0.002181)	0.023275 (0.001647)	0.030065(0.003587)	**0.023245** (0.001921)
M20	0.071069 (0.001973)	**0.060446** (0.002518)	0.072983 (0.001576)	0.060781 (0.001884)

Bai, Z. D., C. R. Rao, and L. C. Zhao (1988). Kernel estimators of density function of directional data. *Journal of Multivariate Analysis 27*(1), 24–39.

Batschelet, E. (1981). Circular statistics in biology. *Academic Press, New York*.

Bedouhene, K. and N. Zougab (2019). Nonparametric multiplicative bias correction for von mises kernel circular density estimator. *Communications in Statistics-Simulation and Computation*, 1–19.

Bedouhene, K. and N. Zougab (2020). A bayesian procedure for bandwidth selection in circular kernel density estimation. *Monte Carlo Methods and Applications 26*(1), 69–82.

Bowers, J. A., I. D. Morton, and G. I. Mould (2000). Directional statistics of the wind and waves. *Applied ocean research 22*(1), 13–30.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika 71*(2), 353–360.

Brunsdon, C. and J. Corcoran (2006). Using circular statistics to analyse time patterns in crime incidence. *Computers, Environment and Urban Systems 30*(3), 300–319.

Fisher, N. I. (1995). *Statistical analysis of circular data*. cambridge university press.

Funke, B. and R. Kawka (2015). Nonparametric density estimation for multivariate bounded data using two non-negative multiplicative bias correction methods. *Computational Statistics and Data Analysis 92*, 148–162.

Table 5: Average integrated squared bias ($\overline{ISB}$) based on 100 replications for 20 models with samples size $n = 10$ and $n = 50$.

	$n = 10$				$n = 50$			
	ISB_{vM}	ISB_{vM-JLN}	ISB_{vM-HG}	ISB_{vM-JSH}	ISB_{vM}	ISB_{vM-JLN}	ISB_{vM-HG}	ISB_{vM-JSH}
M1	**0.005219**	0.008038	0.021296	0.010257	**0.000119**	0.000168	0.002499	0.000714
M2	0.014345	0.010257	**0.001855**	0.002298	0.003316	0.001879	**0.000361**	0.000877
M3	0.049497	0.037105	0.034794	**0.022167**	0.038346	0.012247	0.025256	**0.008201**
M4	0.008672	0.007433	**0.005598**	0.006518	**0.000925**	0.002404	0.003753	0.003842
M5	0.152022	0.113836	0.140459	**0.100007**	0.148395	0.100820	0.133217	**0.093216**
M6	0.056014	0.043471	0.050594	**0.038419**	0.053628	**0.040018**	0.043361	0.041668
M7	0.027229	0.021211	0.024952	**0.019256**	0.012103	**0.004418**	0.014214	0.004566
M8	0.038949	0.027765	0.030994	**0.021352**	0.018364	**0.007779**	0.025824	0.008485
M9	0.010064	**0.009071**	0.025743	0.015451	0.003066	0.001300	0.005760	**0.001122**
M10	0.052183	0.045454	**0.042911**	0.043626	0.047571	0.033729	0.038858	**0.032650**
M11	0.023463	0.020239	0.025946	**0.019101**	0.021022	**0.010154**	0.025315	0.010812
M12	**0.034089**	0.035811	0.048286	0.037611	0.008910	0.005016	0.011260	**0.004757**
M13	0.047050	0.040661	0.052122	**0.031847**	0.033815	**0.020224**	0.038620	0.020977
M14	0.065126	**0.043785**	0.066986	0.045752	0.044921	**0.029098**	0.053015	0.031722
M15	0.005597	0.008574	0.035514	**0.022371**	**0.005597**	0.008574	0.035514	0.022371
M16	0.073712	0.070974	0.080689	**0.065076**	0.065496	**0.049189**	0.064059	0.050704
M17	0.108569	0.088024	0.097095	**0.078830**	0.076755	0.064740	0.073507	**0.063844**
M18	0.050216	**0.049516**	0.056337	0.049953	0.032852	**0.029673**	0.032681	0.031185
M19	0.038135	**0.034433**	0.039294	0.035087	0.027261	**0.022225**	0.031473	0.022280
M20	0.083918	**0.080912**	0.090967	0.083591	0.072655	**0.061176**	0.078217	0.063647

García-Portugués, E. (2013). Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. *Electronic Journal of Statistics 7*, 1655–1685.

Hagmann, M. and O. Scaillet (2007). Local multiplicative bias correction for asymmetric kernel density estimators. *Journal of Econometrics 141*(1), 213–249.

Hall, P., G. P. Watson, and J. Cabrera (1987). Kernel density estimation with spherical data. *Biometrika 74*(4), 751–762.

Harfouche, L., S. Adjabi, N. Zougab, and B. Funke (2018). Multiplicative bias correction for discrete kernels. *Statistical Methods and Applications 27*(2), 253–276.

Harfouche, L., N. Zougab, and S. Adjabi (2020). Multivariate generalised gamma kernel density estimators and application to non-negative data. *International Journal of Computing Science and Mathematics 11*(2), 137–157.

Hirukawa, M. (2010). Nonparametric multiplicative bias correction for kernel-type density estimation on the unit interval. *Computational Statistics and Data Analysis 54*(2), 473–495.

Hirukawa, M. and M. Sakudo (2014). Nonnegative bias reduction methods for density estimation using asymmetric kernels. *Computational Statistics and Data Analysis 75*, 112–123.

Hirukawa, M. and M. Sakudo (2019). Another bias correction for asymmetric kernel density estimation with a parametric start. *Statistics and Probability Letters 145*, 158–165.

Table 6: Average integrated squared bias ($\overline{ISB}$) based on 100 replications for 20 models with samples size $n = 100$ and $n = 200$.

	$n = 100$				$n = 200$			
	ISB_{vM}	ISB_{vM-JLN}	ISB_{vM-HG}	ISB_{vM-JSH}	ISB_{vM}	ISB_{vM-JLN}	ISB_{vM-HG}	ISB_{vM-JSH}
M1	**3.04082e-05**	6.34841e-05	0.001504	0.000482	**0.000130**	0.000426	0.002935	0.001184
M2	0.003346	0.001899	**0.000185**	0.000800	0.002320	0.001072	**0.000240**	0.000379
M3	0.035586	0.010318	0.023821	**0.006829**	0.032850	0.009040	0.021154	**0.005623**
M4	0.002290	0.001300	0.001273	**0.001134**	0.001808	0.000601	0.000627	**0.000183**
M5	0.151102	**0.102017**	0.135469	0.111047	0.146730	**0.096512**	0.132177	0.098760
M6	0.053433	**0.041933**	0.042452	0.044853	0.054779	**0.037884**	0.044704	0.040683
M7	0.012055	**0.003418**	0.015161	0.003627	0.011890	**0.003048**	0.014729	0.003406
M8	0.019035	0.007261	0.024175	**0.007019**	0.018971	0.006902	0.021794	**0.006061**
M9	0.003343	0.002097	0.003830	**0.001869**	0.003254	0.001306	0.003497	**0.001161**
M10	0.039047	0.028795	0.033483	**0.028344**	0.039227	0.029577	0.034169	**0.029576**
M11	0.019389	**0.007878**	0.024427	0.008800	0.019270	**0.007573**	0.022607	0.007841
M12	0.006522	**0.003136**	0.009393	0.003439	0.006229	**0.002157**	0.010371	0.002493
M13	0.035241	**0.019218**	0.031964	0.020593	0.033787	**0.017780**	0.038108	0.018465
M14	0.043921	**0.027408**	0.050541	0.029725	0.043617	**0.027364**	0.048747	0.028654
M15	0.004455	0.002909	0.004828	**0.001855**	0.004336	0.002906	0.003675	**0.001368**
M16	0.059258	**0.046620**	0.063620	0.048120	0.059188	**0.046421**	0.062583	0.047996
M17	0.076536	**0.063448**	0.073174	0.063749	0.074252	0.063933	0.073700	**0.061833**
M18	0.031513	**0.029350**	0.032692	0.030959	0.031188	**0.027903**	0.030837	0.028397
M19	0.026697	0.022108	0.028112	**0.022010**	0.026548	0.020033	0.027767	**0.019929**
M20	0.068609	**0.057966**	0.071069	0.058526	0.068106	**0.057721**	0.070831	0.058020

Table 7: Estimate parameters with ML method for Time series of flare azimuths and Long-axis orientations of feldspar laths data sets

	Time series of flare azimuths	Long-axis orientations of feldspar laths
$\hat{\mu}_{ML}$	3.050395	3.136083
$\hat{k}_{ML}$	1.343009	1.015225

Hjort, N. L. and I. K. Glad (1995). Nonparametric density estimation with a parametric start. *The Annals of Statistics 23*(3), 882–904.

Jammalamadaka, S. R. and U. J. Lund (2006). The effect of wind direction on ozone levels: a case study. *Environmental and Ecological Statistics 13*(3), 287–298.

Jones, M. C., O. Linton, and J. P. Nielsen (1995). A simple bias reduction method for density estimation. *Biometrika 82*(2), 327–338.

Jones, M. C., D. F. Signorini, and N. L. Hjort (1999). On multiplicative bias correction in kernel density estimation. *Sankhyā: the indian journal of statistics 61*, 422–430.

Klemelä, J. (2000). Estimation of densities and derivatives of densities with directional data. *Journal of Multivariate Analysis 73*(1), 18–40.

Kokonendji, C. C., T. Senga Kiessé, and N. Balakrishnan (2009). Semiparametric estimation for count data through weighted distributions. *Journal of Statistical Planning and Inference 139*(10), 3625–3638.

Mann, K. A., S. Gupta, A. Race, M. A. Miller, and R. J. Cleary (2003). Application of circular statistics in the study of crack distribution around cemented femoral components. *Journal of biomechanics 36*(8), 1231–1234.

Mardia, K. V. and P. E. Jupp (2009). *Directional statistics*, Volume 494. John Wiley & Sons.

Marzio, M. Di, A. Panzera, and C. C. Taylor (2009). Local polynomial regression for circular predictors. *Statistics and Probability Letters 79*(19), 2066–2075.

Marzio, M. Di, A. Panzera, and C. C. Taylor (2011). Kernel density estimation on the torus. *Journal of Statistical Planning and Inference 141*(6), 2156–2173.

Mooney, J. A., P. J. Helms, and I. T. Jollife (2003). Fitting mixtures of von mises distributions: a case study involving sudden infant death syndrome. *Computational Statistics and Data Analysis 41*(3-4), 505–513.

Oliveira, M., R. M. Crujeiras, and A. Rodríguez-Casal (2012). A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics and Data Analysis 56*(12), 3898–3908.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics 9*, 65–78.

Smith, N. M. (1988). Reconstruction of the tertiary drainage systems in the inverell region. *Unpublished B. Sc. (Hons.) thesis, Department of Geography, University of Sidney, Australia*.

Taylor, C. C. (2008). Automatic bandwidth selection for circular density estimation. *Computational Statistics and Data Analysis 52*(7), 3493–3500.

Terrell, G. R. and D. W. Scott (1980). On improving convergence rates for nonnegative kernel density estimators. *The Annals of Statistics 8*(5), 1160–1163.

Tsuruta, Y. and M. Sagae (2017). Higher order kernel density estimation on the circle. *Statistics and Probability Letters 131*, 46–50.

Zougab, N. and S. Adjabi (2016). Multiplicative bias correction for generalized birnbaum–saunders kernel density estimators and application to nonnegative heavy tailed data. *Journal of the Korean Statistical Society 45*(1), 51–63.

Zougab, N., L. Harfouche, Y. Ziane, and S. Adjabi (2018). Multivariate generalized birnbaum—saunders kernel density estimators. *Communications in Statistics-Theory and Methods 47*(18), 4534–4555.

SPANISH JOURNAL OF STATISTICS
Vol. **7** No. 1 2025, Pages 109–129
doi:https://doi.org/10.37830/
SJS.2025.1.06

REGULAR ARTICLE

Objective Bayesian goodness-of-fit tests for the alpha-skew-normal distribution

José Rodolfo Olmos-Zepeda

Departamento de Estadística y Ciencia de Datos, Colegio de Postgraduados, olmos.rodolfo@colpos.mx

Sergio Pérez-Elizalde

Departamento de Estadística y Ciencia de Datos, Colegio de Postgraduados, sergiop@colpos.mx

Received: November 4, 2025. Returned: . .

Abstract: The family of alpha-skew-normal (ASN) distributions is a flexible class of three-parameter probability models characterized by their location, scale, and shape. The shape parameter governs both asymmetry and uni-bimodality, allowing the distribution to model unimodal or bimodal data with varying degrees of skewness. This paper proposes an objective Bayesian goodness-of-fit test to determine whether a random sample follows an ASN distribution when parameters are unknown. The test statistics are based on empirical distribution function, whose sampling distributions depend solely on the shape parameter. Their prior predictive distributions, serving as null distributions, are obtained by integrating out the shape parameter with respect to a proper approximation of Jeffreys prior, specifically a Cauchy prior, chosen for its analytical tractability. Critical values are estimated via Monte Carlo simulation. A comprehensive simulation study demonstrates that the proposed tests maintain the nominal significance level across various scenarios and exhibit strong power properties against a range of alternative distributions. Finally, the methodology is illustrated through real-data examples, showcasing its practical applicability.

Keywords: alpha-skew-normal distribution, empirical distribution function, goodness-of-fit test, Jeffreys prior, Monte Carlo simulation, prior predictive distribution.

MSC: 60E05, 62G10, 62G30, 62E17

Keywords: alpha-skew-normal distribution, empirical distribution function, goodness-of-fit test, Jeffreys prior, Monte Carlo simulation, prior predictive distribution.

1 Introduction

The alpha-skew-normal (ASN) distribution, proposed by Elal-Olivero (2010), is an extension of the normal distribution with a shape parameter that regulates skewness differently than the asymmetric

normal model of Azzalini (1985); it also captures a uni-bimodality effect, which gives this family of distributions the flexibility to model unimodal or bimodal data sets that have some degree of skewness. The normal distribution is a special case of the ASN family.

Let $\mathbb{R}$ denote the set of real numbers and ϕ the density function of a standard normal random variable. A continuous random variable Z is said to follow an alpha-skew-normal distribution with shape parameter $\alpha \in \mathbb{R}$, denoted by $Z \sim \text{ASN}(\alpha)$, if its probability density function (pdf) is given by

$$f_Z(z;\alpha) = \frac{(1-\alpha z)^2 + 1}{2+\alpha^2}\phi(z), \quad z \in \mathbb{R}, \tag{1}$$

where α is a parameter that controls both the skewness and the uni-bimodality of the density function. For positive values of α, the distributions are skewed to the right, and for negative values, the distributions are skewed to the left. The ASN distribution is unimodal if $-1.34 < \alpha < 1.34$ and bimodal otherwise. Furthermore, when $\alpha \to \pm\infty$, Z converges in distribution to a bimodal random variable with pdf $y^2\phi(y)$ for $y \in \mathbb{R}$ (Elal-Olivero, 2010).

Let $\mu \in \mathbb{R}$ and $\sigma > 0$. If $Z \sim \text{ASN}(\alpha)$, then the location-scale ASN distribution can be defined as the distribution of the continuous random variable $X = \mu + \sigma Z$, whose probability density function is given by

$$f_X(x;\mu,\sigma,\alpha) = \left[\frac{(1-\alpha\{(x-\mu)/\sigma\})^2 + 1}{\sigma\,(2+\alpha^2)}\right]\phi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}. \tag{2}$$

A random variable X following an ASN distribution with location, scale and shape parameters μ, σ, and α, respectively, is denoted as $X \sim \text{ASN}(\mu,\sigma,\alpha)$. When $\mu = 0$ and $\sigma = 1$, the pdf (2) simplifies to equation (1). The skewness of X ranges from -0.811 to 0.811, while the kurtosis varies between -1.333 and 0.749. The cumulative distribution function (cdf) of X is given by

$$F_X(x;\mu,\sigma,\alpha) = \Phi\left(\frac{x-\mu}{\sigma}\right) + \alpha\left(\frac{2\sigma-\alpha(x-\mu)}{\sigma(2+\alpha^2)}\right)\phi\left(\frac{x-\mu}{\sigma}\right). \tag{3}$$

In practice, the ASN distribution remains relatively unknown, yet it has played a significant role in the theoretical development of other asymmetric uni-bimodal distributions. Notable examples include the alpha-skew-Laplace (Harandi and Alamatsaz, 2013), alpha-skew-logistic (Chakraborty and Hazarika, 2014), Balakrishnan-alpha-skew-normal (Chakraborty et al., 2014), and alpha-skew-normal *slash* (Gui, 2014) distributions. Additionally, (Louzada et al., 2016) and (Louzada and Ara, 2019) extended the univariate ASN distribution, as defined in equation (1), to its bivariate and multivariate counterparts, respectively.

Let $X_1, \ldots, X_n$ be a random sample of size n from a population with an unknown cdf F. This paper examines the composite goodness-of-fit problem for the $\text{ASN}(\mu,\sigma,\alpha)$ distribution, testing the null hypothesis

$$H_0 : X_1, \ldots, X_n \sim \text{ASN}(\mu,\sigma,\alpha), \quad \text{where } \mu, \sigma \text{ and } \alpha \text{ are unknown}, \tag{4}$$

versus the alternative hypothesis $H_1 : X_1, \ldots, X_n \nsim \text{ASN}(\mu,\sigma,\alpha)$.

Since this problem has not been previously studied, we propose a test procedure based on empirical distribution function (EDF) statistics as a first approach. The null distributions of the test statistics, which determine the critical threshold and depend on the shape parameter (Stephens, 1986), are their prior predictive distributions. These distributions are obtained by integrating out the shape parameter with respect to its prior distribution. To ensure objectivity, Jeffreys rule is used to derive a prior

distribution for the shape parameter. However, since the resulting Jeffreys prior is improper, we propose a proper approximation that corresponds to a well-known distribution.

The remainder of this paper is organized as follows. Section 2 presents the method for approximating Jeffreys prior distribution for the shape parameter. Section 3 defines the test statistics and describes the methodology for estimating their prior predictive distributions. Additionally, a table of critical values based on the most relevant sample quantiles from these distributions is provided. Section 4 reports the results of a Monte Carlo simulation study assessing the type I error probability and power of the tests. In Section 5, the proposed tests are applied to two real datasets: one unimodal and the other bimodal. Finally, Section 6 summarizes the main conclusions.

2 Jeffreys prior distribution for the shape parameter α

For a probability model with a one-dimensional parameter θ, the Jeffreys prior distribution of θ is proportional to the square root of its Fisher information of θ.

According to (Elal-Olivero, 2010), the Fisher information of the shape parameter in the ASN distribution with density function (1) is given by

$$\mathscr{I}(\alpha) = \frac{4(\alpha^2 b_2 + 2b_2 - \alpha^2)}{(2+\alpha^2)^2},$$

where

$$b_2 = \mathbb{E}\left[W^2 \frac{(1-\alpha W)^2}{(1-\alpha W)^2 + 1}\right],$$

and W is a standard normal random variable. Consequently, the Jeffreys prior distribution for α, denoted by $\pi_J(\alpha)$, is given by

$$\pi_J(\alpha) \propto \mathscr{I}(\alpha)^{1/2}.$$

The value of b_2 can be computed numerically for each $\alpha \in \mathbb{R}$, enabling the evaluation of $\mathscr{I}(\alpha)$ and the construction of the corresponding plot of $\pi_J(\alpha)$. The resulting graph is shown in Figure 1

Figure 1 reveals that $\pi_J(\alpha)$ exhibits similarities to well-known distributions, such as the Cauchy, generalized double Pareto (Armagan et al., 2013), and Laplace (or double exponential) distributions, each centered at zero. Since $\pi_J(\alpha)$ is improper, we approximate it using one of these three proper and computationally efficient distributions. To determine the optimal scale parameter λ for the chosen approximation, we minimize the Kullback-Leibler divergence between the normalized Jeffreys prior and each candidate distribution.

Since $\pi_J(\alpha)$ is improper, we consider its normalized version

$$p_J(\alpha) = \frac{\mathscr{I}(\alpha)^{1/2}}{C(\xi)},$$

where

$$C(\xi) = \lim_{\xi\to\infty} \int_{-\xi}^{\xi} \mathscr{I}(\alpha)^{1/2} \quad d\alpha.$$

Let $g_J(\alpha)$ denote the proposed Cauchy approximation given by

$$g_J(\alpha) = \frac{1}{\pi\lambda}\left(1 + \frac{\alpha^2}{\lambda^2}\right)^{-1}.$$

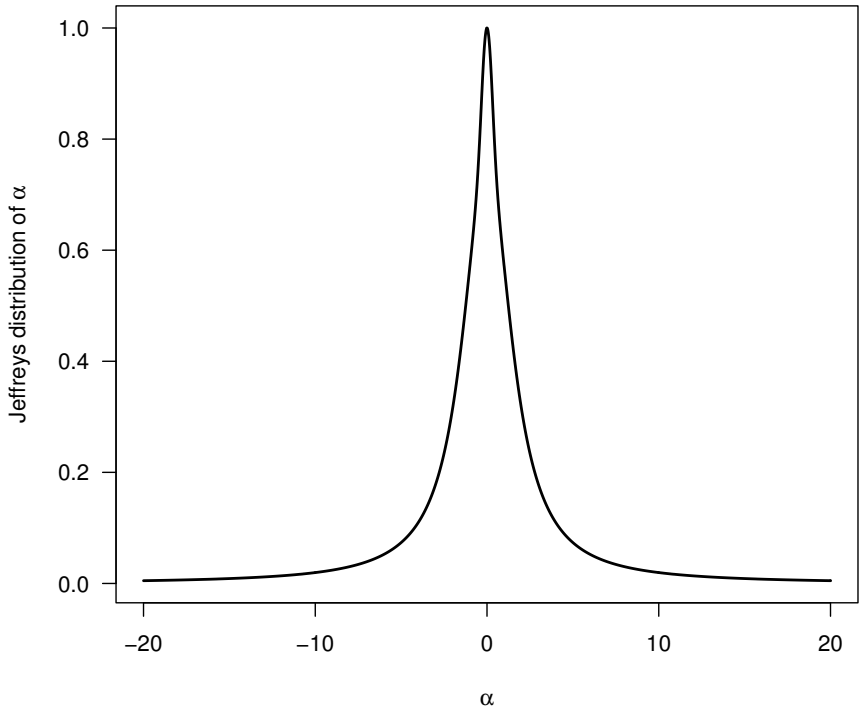


Figure 1: Jeffreys prior distribution for $\alpha \in [-20, 20]$, obtained via numerical integration.

The Kullback-Leibler divergence between $p_J(\alpha)$ and $g_J(\alpha)$ is given by

$$D_{\mathrm{KL}}(p_J, g_J) = \lim_{\xi \to \infty} \left[k(\xi) + \log \lambda + \int_{-\xi}^{\xi} \log \left(1 + \frac{\alpha^2}{\lambda^2} \right) \frac{\mathscr{I}(\alpha)^{1/2}}{C(\xi)} d\alpha \right], \tag{5}$$

where $k(\xi)$ is a term independent of λ. The first-order derivative of above equation with respect to λ is

$$\frac{\partial}{\partial \lambda} D_{\mathrm{KL}}(p_J, g_J) = \lim_{\xi \to \infty} \frac{1}{\lambda} \left[1 - \frac{1}{\lambda^2} \int_{-\xi}^{\xi} 2\alpha^2 \left(1 + \frac{\alpha^2}{\lambda^2} \right)^{-1} \frac{\mathscr{I}(\alpha)^{1/2}}{C(\xi)} d\alpha \right].$$

Setting this expression to zero yields an equation that, when solved numerically, provides the optimal scale parameter λ for the Cauchy approximation. The resulting value, obtained via numerical integration, is $\lambda_{\mathrm{CA}} = 1.48$. Similarly, for the double generalized Pareto and Laplace approximations, the optimal scale parameters are found to be $\lambda_{\mathrm{GPD}} = 1.46$ and $\lambda_{\mathrm{LAP}} = 4.94$, respectively. The resulting distributions are compared with the Jeffreys prior in Figure 2.

Figure 2 shows that the Cauchy distribution provides the best approximation to Jeffreys prior for the shape parameter of the ASN distribution. The generalized double Pareto distribution is also a viable option; however, in the tails, the Cauchy distribution aligns more closely with the normalized Jeffreys prior. Around $\alpha = 0$, both distributions exhibit minimal differences from the Jeffreys prior, suggesting that these small discrepancies should not significantly impact the desirable properties of the tests.

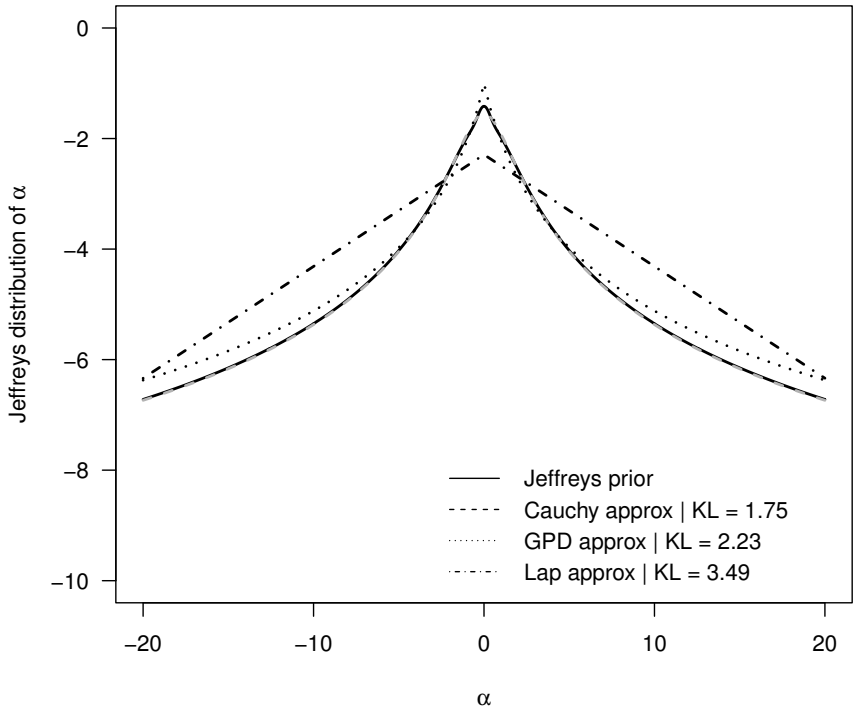


Figure 2: Approximations to the Jeffreys prior distribution for α (in logarithmic scale).

By contrast, the Laplace distribution deviates significantly over nearly the entire range of α. This observation is consistent with the Kullback-Leibler divergence values reported in Figure 2, which indicate that the Cauchy distribution results in the least information loss. Based on this, we propose using a Cauchy distribution centered at zero with a rounded scale parameter of $\lambda = 3/2$ as an approximation of $\pi_J(\alpha)$. This approximation, denoted by $\tilde{\pi}_J(\alpha)$, is given by

$$\tilde{\pi}_J(\alpha) = \frac{1}{\pi(3/2)} \left(1 + \frac{\alpha^2}{(3/2)^2}\right)^{-1}. \tag{6}$$

3 Tests for alpha-skew-normality

In this section, we introduce the test statistics used for evaluating alpha-skew-normality and describe the methodology to estimate their null distributions under the proposed Bayesian framework. We describe the procedure for estimating the quantiles of the sampling distribution of each when the hypothesized distribution is alpha-skew-normal with parameter values estimated from the data using the maximum likelihood method. Furthermore, we summarize the steps necessary to the test the null hypothesis (4) of interest.

3.1 Test statistics based on EDF

The most widely studied goodness-of-fit procedures for a specific family of distributions in the literature are those based on the EDF, a step function denoted by F_n. Given a realization $x_1, \ldots, x_n$ of size n, the EDF is computed as (Stephens, 1986)

$$F_n(x) = \begin{cases} 0 & \text{if } x < x_{(1)}, \\ i/n & \text{if } x_{(i)} \leq x < x_{(i+1)}, \quad i = 1, \ldots, n-1, \\ 1 & \text{if } x \geq x_{(n)}, \end{cases}$$

where $x_{(1)} < \cdots < x_{(n)}$ denote the ordered sample.

A goodness-of-fit test based on F_n assesses whether the sample follows a distribution with an unknown cdf F, quantifying, from a test statistic, the discrepancy between F_n and F, determining whether H_0 should be rejected. Following Stephens (1986), test statistics based on F_n can be categorized into two families: the Kolmogorov–Smirnov family that contains the Kolmogorov–Smirnov statistic D and the Kuiper statistic V, defined as

$$D = \max\left(D^+, D^-\right), \quad V = D^+ + D^-,$$

where

$$D^+ = \sup_x \left\{F_n(x) - F(x)\right\} \quad \text{and} \quad D^- = \sup_x \left\{F(x) - F_n(x)\right\}.$$

The Cramér-von Mises family that contains the Cramér–von Mises statistic W^2, Watson's statistic U^2 and the Anderson–Darling statistic A^2, defined as

$$W^2 = n \int_{\mathbb{R}} \left\{F_n(x) - F(x)\right\}^2 dF(x),$$

$$U^2 = n \int_{\mathbb{R}} \left(F_n(x) - F(x) - \int_{\mathbb{R}} \left\{F_n(x) - F(x)\right\} dF(x)\right)^2 dF(x),$$

$$A^2 = n \int_{\mathbb{R}} \frac{\left\{F_n(x) - F(x)\right\}^2}{F(x)\left\{1 - F(x)\right\}} dF(x).$$

Given an observed sample, these statistics are computed as follows (Stephens, 1974):

$$D = \max\left(D^+, D^-\right), \tag{7}$$

$$V = D^+ + D^-, \tag{8}$$

$$D^+ = \max_i \left(\frac{i}{n} - p_{(i)}\right), \tag{9}$$

$$D^- = \max_i \left(p_{(i)} - \frac{i-1}{n}\right), \tag{10}$$

$$W^2 = \sum_{i=1}^{n} \left(p_{(i)} - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n}, \tag{11}$$

$$U^2 = W^2 - n\left(\bar{p} - \frac{1}{2}\right)^2, \tag{12}$$

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log p_{(i)} + \log\left(1 - p_{(n+1-i)}\right)\right], \tag{13}$$

where $p_{(i)} = F(x_{(i)})$ represents the hypothesized cdf values, and $\bar{p}$ is the arithmetic mean of $p_{(1)}, \ldots, p_{(n)}$. For the problem considered in this paper, F corresponds to equation (3). However, since the parameters μ, σ and α are unknown, they are replaced by their maximum likelihood estimates $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\alpha}$. Hence,

$$F = F_X(x; \hat{\mu}, \hat{\sigma}, \hat{\alpha}). \tag{14}$$

When parameter estimates are used, the null distribution of the test statistics depends on the unknown shape parameter (Mateu-Figueras et al., 2009). As a result, the critical values of the tests also depend on the shape parameter. Formally, the critical constant $k_{1-\gamma}$ for a test of significance level $\gamma \in (0,1)$ satisfies

$$\max_{\alpha} \mathbb{P}\left(T_{\text{EDF}} > k_{1-\gamma} \mid H_0 \text{ is true}\right) = \gamma, \quad \alpha \in \mathbb{R},$$

where T_{EDF} represents any of the aforementioned test statistics.

There is no general approach to solving this issue. Some proposed methods in the literature include: (i) transforming the sample into one from a known distribution, reducing the problem to a setting where standard tests exist, such as normality transformations (Chen and Balakrishnan, 1995), and (ii) using the parametric bootstrap method to estimate the null distribution of the test statistic and approximate critical values via empirical quantiles (Meintanis, 2007).

In this paper, instead, we propose using the prior predictive distribution of T_{EDF}, obtained by integrating out the shape parameter using its prior distribution. In the previous section, we introduced an approximation to the Jeffreys prior for this parameter, which will be used for this purpose.

3.2 Prior predictive distribution of T_{EDF}

If $\pi(\alpha)$ is a proper prior distribution for the shape parameter, the prior predictive distribution of T_{EDF} is given by

$$\pi(t_{\text{EDF}}) = \int_{\mathbb{R}} \pi(t_{\text{EDF}}|\alpha)\pi(\alpha)\, d\alpha, \tag{15}$$

which represents the marginal distribution of the test statistic. The integral in equation (15) cannot be evaluated analytically but can be approximated by Monte Carlo simulation. To the best of our knowledge, the only reference in the literature applying this approach to goodness-of-fit tests is Cabras and Castellanos (2009), in the context of the asymmetric normal distribution introduced by Azzalini (1985), where the prior predictive p-value (Bayarri and Berger, 2000), originally proposed by Box (1980), serves as a foundation.

Bayarri and Berger (2000) suggest specifying a non-informative prior distribution for the noise parameter to ensure an objective analysis. However, a drawback of noninformative prior distributions is that they are often improper, making the prior predictive distribution (15) improper as well, rendering it invalid for practical use. Nevertheless, a proper distribution that closely approximates the Jeffreys prior for the shape parameter is available, namely, the distribution given in equation (6).

The procedure to estimate the distribution (15) of each test statistic consists of the following three steps:

1. Simulate $\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(M)} \sim \tilde{\pi}_J(\alpha)$, with $M = 50,000$.

2. Set the sample size n to n and generate a random sample from the distribution ASN(μ, σ, α) with $\mu = 0$, $\sigma = 1$ and $\alpha = \alpha^{(m)}$, for $m = 1, \ldots, M$. The location and scale parameters were set to $\mu = 0$ and $\sigma = 1$ because the distribution of test statistics based on F_n is invariant to changes in location and scale (Stephens, 1986).

3. On every m-th sample:

   (a) Obtain the maximum likelihood estimates of μ, σ and α.

   (b) Calculate the value of T_{EDF}, using the corresponding expression from equations (7), (8), (11), (12) and (13), with $\hat{p}_{(i)}$ given by

$$\hat{p}_{(i)} = \Phi\left(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}\right) + \hat{\alpha}\left(\frac{2\hat{\sigma} - \hat{\alpha}(x_{(i)} - \hat{\mu})}{\hat{\sigma}(2 + \hat{\alpha}^2)}\right)\phi\left(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}\right), \quad i = 1, \ldots, n,$$

   and $x_{(1)}, \ldots, x_{(n)}$ the sample given in step (2) sorted in non-decreasing order.

The proposed procedure generates a set of M values of T_{EDF}, which are used to estimate its null distribution for a given sample size. Similarly, the $100(1-\gamma)\%$ quantiles of its empirical distribution can be obtained to approximate the critical value $k_{1-\gamma}$ for the corresponding test. The quantiles of the statistics D and V were computed with respect to $\sqrt{n}D$ and $\sqrt{n}V$, as these statistics tend to zero as $n \to \infty$ (Lilliefors, 1967).

All computational procedures were implemented in the R software (R Core Team, 2023). Pseudo-random numbers from the ASN(μ, σ, α) distribution were generated using the acceptance-rejection method based on its stochastic representation, as described by Elal-Olivero (2010). Maximum likelihood estimates for the parameters μ, σ and α were obtained using the `optim` function. Pseudo-random numbers from the Cauchy distribution were generated using the `rcauchy` function. Table 1 presents the 90%, 95%, and 99% quantiles of the empirical distribution of the test statistic for different sample sizes.

3.3 Proposed general test procedure

The following procedure is proposed to test the null hypothesis (4) using a random sample $x_1, \ldots, x_n$ of size n, based on one of the test statistics under consideration:

(1) Compute the maximum likelihood estimates of the parameters μ, σ and α for the ASN(μ, σ, α) distribution, denoted as $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\alpha}$.

(2) Calculate the value of the chosen test statistic using the appropriate formula from equations (7), (8), (11), (12) or (13), where $\hat{p}_{(i)}$ is given by

$$\hat{p}_{(i)} = \Phi\left(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}\right) + \hat{\alpha}\left(\frac{2\hat{\sigma} - \hat{\alpha}(x_{(i)} - \hat{\mu})}{\hat{\sigma}(2 + \hat{\alpha}^2)}\right)\phi\left(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}\right), \quad i = 1, \ldots, n,$$

   where $x_{(1)}, \ldots, x_{(n)}$ are the data sorted in non-decreasing order.

(3) For a given significance level $\gamma \in (0, 1)$, obtain the $(1-\gamma)$ quantile of the test statistic corresponding to the sample size n from Table 1. If n is not listed in the table, interpolate linearly between the two closest values of n to approximate the quantile. For $n > 500$, use the quantile associated with $n = 500$.

$1-\gamma$	n	$\sqrt{n}D_n$	$\sqrt{n}V_n$	W_n^2	U_n^2	A_n^2
0.90	20	0.7856	1.2521	0.0993	0.0805	0.5842
	30	0.8084	1.2836	0.1034	0.0824	0.6131
	40	0.8225	1.3032	0.1063	0.0849	0.6251
	50	0.8234	1.3196	0.1077	0.0858	0.6257
	75	0.8236	1.3245	0.1089	0.0867	0.6293
	100	0.8241	1.3244	0.1091	0.0869	0.6298
	150	0.8261	1.3289	0.1098	0.0858	0.6303
	200	0.8293	1.3424	0.1101	0.0868	0.6326
	300	0.8323	1.3461	0.1109	0.0875	0.6318
	400	0.8349	1.3444	0.1096	0.0877	0.6394
	500	0.8371	1.3467	0.1105	0.0875	0.6396
0.95	20	0.8597	1.3456	0.1219	0.0987	0.6964
	30	0.8839	1.3791	0.1302	0.1005	0.7382
	40	0.9096	1.3966	0.1386	0.1034	0.7544
	50	0.9043	1.4006	0.1347	0.1043	0.7541
	75	0.9051	1.4212	0.1341	0.1049	0.7565
	100	0.9102	1.4227	0.1356	0.1058	0.7596
	150	0.9132	1.4377	0.1365	0.1056	0.7631
	200	0.9152	1.4427	0.1379	0.1062	0.7645
	300	0.9159	1.4411	0.1369	0.1069	0.7682
	400	0.9174	1.4435	0.1408	0.1067	0.7719
	500	0.9184	1.4446	0.1406	0.1069	0.7801
0.99	20	1.0129	1.5426	0.1752	0.1393	0.9345
	30	1.0585	1.5652	0.1878	0.1426	1.0061
	40	1.0685	1.6062	0.2011	0.1513	1.1189
	50	1.0862	1.6112	0.2097	0.1511	1.1215
	75	1.0917	1.6196	0.2122	0.1518	1.1381
	100	1.0921	1.6333	0.2158	0.1518	1.1475
	150	1.0919	1.6465	0.2162	0.1568	1.1496
	200	1.0922	1.6467	0.2167	0.1573	1.1483
	300	1.0942	1.6465	0.2166	0.1594	1.1468
	400	1.0936	1.6528	0.2193	0.1599	1.1503
	500	1.0963	1.6501	0.2177	0.1594	1.1499

Table 1: $100th(1-\gamma)\%$ quantile of the empirical distribution of each test statistic, for different sample sizes n and different significance levels γ, estimated by $M = 50,000$ simulations.

(4) Reject H_0 at significance level γ if the test statistic computed in step (2) exceeds the quantile obtained in step (3).

An R script (R Core Team, 2023) containing the necessary functions to implement this test procedure is provided in the appendix.

4 Size and power of the tests

To assess whether the actual type I error probability of each test in the proposed procedure matches or does not exceed the specified significance level, and to evaluate the power of the tests against various alternative distributions, a Monte Carlo simulation study was conducted as described below.

4.1 Simulation experiment

To estimate the type I error probability of the tests, first, B samples of size n are simulated from the ASN(μ, σ, α) distribution with parameters $\mu = 0$, $\sigma = 1$ and different values of $\alpha \in [-20, 20]$ as specified in Table 2. The proposed procedure is then applied using the five test statistics under consideration, with the significance level set at $\gamma = 0.05$. Finally, the proportion of times that H_0 is rejected over the B repetitions is computed. The power of the test was estimated in the same manner, using samples from alternative distributions. For this study, a variety of distributions are considered, including symmetric and asymmetric unimodal distributions, asymmetric bimodal distributions, and mixtures of two normal distributions.

1. $t(v)$: Student's t with v degrees of freedom. When $v = 1$ we have the standard Cauchy distribution.
2. $\mathrm{Lap}(0, 1)$: standard Laplace.
3. $\mathrm{Logit}(0, 1)$: standard logistic.
4. $\mathrm{LN}(\mu, \sigma)$: lognormal with location and scale parameters μ and σ, respectively.
5. $\mathrm{Ga}(a, b)$: gamma with shape and scale parameters a and b, respectively
6. $\mathrm{Exp}(1)$: standard exponential.
7. $\chi^2(\nu)$: chi-square with ν degrees of freedom.
8. $\mathrm{Gu}(0, 1)$: standard Gumbel.
9. $\mathrm{SN}(\lambda)$: standard asymmetric normal with shape parameter λ.
10. BASN(α): Balakrishnan alpha-skew-normal with parameter $\alpha \in \mathbb{R}$. Its probability density function is

$$f_X(x; \alpha) = \frac{1}{C(\alpha)} \left[\frac{(1 - \alpha x)^2 + 1}{2 + \alpha^2} \right]^2 \phi(x), \quad x \in \mathbb{R},$$

    where $C(\alpha) = 3 - 4(2 + \alpha^2)^{-1}$. The distribution is unimodal for $-0.96 < \alpha < 0.96$.
11. ASLap(α): alpha-skew-Laplace with parameter $\alpha \in \mathbb{R}$. Its probability density function is

$$f_X(x; \alpha) = \frac{(1 - \alpha x)^2 + 1}{4(1 + \alpha^2)} f_{\mathrm{LP}}(x), \quad x \in \mathbb{R},$$

    where $f_{\mathrm{LP}}(x)$ denotes the standard Laplace distribution. The distribution is unimodal for $-1 < \alpha < 1$.
12. ASLogit(α): alpha-skew-logistic with parameter $\alpha \in \mathbb{R}$. Its probability density function is

$$f_X(x; \alpha) = \frac{3(1 - \alpha x)^2 + 1}{6 + \pi^2 \alpha^2} f_{\mathrm{LG}}(x), \quad x \in \mathbb{R},$$

    where $f_{\mathrm{LG}}(x)$ denotes the standard logistic distribution. The distribution is unimodal for $-0.8 < \alpha < 0.8$.
13. mixN: mixture of two normal distributions N(μ_A, σ_A) and N(μ_B, σ_B) with mixing parameter $w \in (0, 1)$.

4.2 Results

According to Table 2, the estimated type I error probability of the tests is generally less than or equal to the specified significance level. This indicates that, under the considered configurations of the shape parameter and sample size, all five tests under study maintain a level of γ. Instances where the estimated probability slightly exceeds the nominal significance level may be attributed to the inherent variability of the simulation.

Several key observations from Tables 3 and 4 are as follows:

- The estimated power of the tests behaves consistently, increasing as the sample size increases.
- When the alternative distribution has heavy tails, as in the case of the standard Cauchy or Student's t distribution with 2 degrees of freedom, the tests exhibit high power even for sample sizes as small as $n = 20$. However, as the degrees of freedom of the Student's t distribution increase, the power of the tests decreases.
- For the standard logistic distribution, the tests exhibit very low power, even with sample sizes of $n = 50$ and $n = 100$, indicating that the tests may struggle to distinguish the logistic distribution from the ASN model. In contrast, the standard Laplace distribution is better identified.
- For most of the unimodal asymmetric alternative distributions studied, the tests demonstrate strong discriminatory capability. However, they perform poorly with the Gumbel distribution, where power estimates remain low even for $n = 100$. Similarly, the tests fail to effectively detect the asymmetric normal distribution for the considered values of the asymmetry parameter. However, the power improves as the absolute value of the shape parameter increases. For the standard exponential distribution, the asymmetric Student's t with 1 degree of freedom, and the chi-square distribution with 2 degrees of freedom, the tests show favorable power. However, as the degrees of freedom increase in the last two cases, the power of the tests decreases.
- Among the specified asymmetric bimodal distributions, the BASN distribution is the most effectively discriminated. The tests demonstrate high power for sample sizes of $n = 100$, with power increasing as bimodality and asymmetry intensify.
- When the alternative distribution is a mixture of two normal distributions, the estimated power of the tests increases as the value of one of its parameters increases or as the mixing proportion deviates from $1/2$.

n	α	D	V	W^2	U^2	A^2
20	-20	0.04	0.05	0.04	0.04	0.04
	-15	0.04	0.04	0.05	0.04	0.05
	-10	0.05	0.04	0.05	0.05	0.04
	-5	0.04	0.04	0.03	0.04	0.05
	-3	0.04	0.06	0.03	0.04	0.04
	-1	0.02	0.04	0.02	0.06	0.03
	0	0.02	0.04	0.02	0.04	0.02
	1	0.03	0.04	0.02	0.04	0.02
	3	0.04	0.04	0.03	0.04	0.04
	5	0.05	0.05	0.04	0.05	0.05
	10	0.05	0.06	0.04	0.04	0.05
	15	0.05	0.05	0.05	0.05	0.04
	20	0.04	0.04	0.04	0.04	0.04
50	-20	0.04	0.05	0.04	0.05	0.04
	-15	0.04	0.04	0.05	0.05	0.05
	-10	0.05	0.05	0.05	0.04	0.05
	-5	0.04	0.06	0.04	0.05	0.04
	-3	0.04	0.04	0.04	0.06	0.05
	-1	0.02	0.04	0.03	0.04	0.04
	0	0.01	0.04	0.01	0.03	0.03
	1	0.03	0.04	0.03	0.04	0.04
	3	0.04	0.05	0.04	0.04	0.04
	5	0.05	0.05	0.05	0.04	0.05
	10	0.05	0.05	0.05	0.05	0.05
	15	0.04	0.05	0.04	0.04	0.04
	20	0.04	0.04	0.04	0.05	0.04
100	-20	0.05	0.05	0.03	0.05	0.05
	-15	0.05	0.05	0.05	0.05	0.04
	-10	0.04	0.04	0.04	0.04	0.04
	-5	0.04	0.04	0.03	0.05	0.04
	-3	0.05	0.05	0.03	0.04	0.03
	-1	0.03	0.04	0.03	0.04	0.03
	0	0.01	0.04	0.01	0.04	0.02
	1	0.02	0.04	0.04	0.04	0.03
	3	0.04	0.05	0.03	0.05	0.04
	5	0.05	0.05	0.05	0.05	0.05
	10	0.04	0.05	0.05	0.04	0.05
	15	0.05	0.05	0.05	0.05	0.05
	20	0.04	0.05	0.04	0.05	0.04

Table 2: Estimated probability of type I error of the tests for different values of n and different values of the parameter α, with $\gamma = 0.05$.

	$n = 20$					$n = 50$				
Alternative distribution	D	V	W^2	U^2	A^2	D	V	W^2	U^2	A^2
Symmetric unimodal distributions										
$\mathrm{Ca}(0,1)$	0.73	0.76	0.80	0.75	0.81	0.96	0.95	0.99	0.96	0.99
$t(2)$	0.45	0.46	0.50	0.47	0.51	0.65	0.65	0.67	0.66	0.70
$t(3)$	0.29	0.30	0.35	0.29	0.35	0.40	0.39	0.45	0.40	0.48
$\mathrm{Lap}(0,1)$	0.08	0.08	0.14	0.10	0.15	0.25	0.26	0.30	0.26	0.34
$\mathrm{Logit}(0,1)$	0.04	0.05	0.07	0.05	0.07	0.06	0.08	0.10	0.06	0.09
Asymmetric unimodal distributions										
$\mathrm{LogN}(0,0.5)$	0.18	0.19	0.22	0.18	0.24	0.30	0.31	0.35	0.31	0.39
$\mathrm{Exp}(1)$	0.42	0.41	0.51	0.41	0.54	0.70	0.69	0.77	0.70	0.80
$\mathrm{Ga}(0.5,1)$	0.73	0.72	0.81	0.72	0.84	0.95	0.95	1.00	0.95	1.00
$\mathrm{Ga}(2,1)$	0.23	0.24	0.30	0.25	0.33	0.48	0.51	0.55	0.47	0.58
$\chi^2(2)$	0.40	0.41	0.45	0.42	0.47	0.59	0.61	0.65	0.63	0.66
$\chi^2(4)$	0.08	0.09	0.10	0.08	0.15	0.23	0.24	0.28	0.24	0.32
$\mathrm{Gu}(0,1)$	0.10	0.09	0.12	0.10	0.11	0.14	0.15	0.17	0.14	0.17
$\mathrm{NS}(0.5)$	0.04	0.05	0.06	0.05	0.06	0.04	0.05	0.03	0.05	0.03
$\mathrm{NS}(2.5)$	0.15	0.14	0.19	0.16	0.21	0.29	0.26	0.34	0.25	0.35
Asymmetric bimodal distributions										
$\mathrm{BASN}(2)$	0.27	0.28	0.31	0.26	0.31	0.46	0.47	0.52	0.45	0.53
$\mathrm{BASN}(6)$	0.43	0.41	0.45	0.42	0.48	0.63	0.60	0.64	0.61	0.65
$\mathrm{ASLap}(2)$	0.10	0.12	0.15	0.12	0.14	0.26	0.27	0.32	0.27	0.35
$\mathrm{ASLap}(6)$	0.33	0.33	0.35	0.33	0.36	0.48	0.50	0.53	0.49	0.56
$\mathrm{ASLogit}(2)$	0.17	0.18	0.19	0.16	0.19	0.30	0.29	0.32	0.29	0.31
$\mathrm{ASLogit}(6)$	0.33	0.32	0.36	0.33	0.38	0.46	0.47	0.50	0.46	0.55
$\mathrm{mixN}(0.5,-2,2,1,1)$	0.08	0.12	0.09	0.13	0.10	0.07	0.10	0.07	0.09	0.06
$\mathrm{mixN}(0.3,-2,2,1,1)$	0.13	0.12	0.16	0.13	0.16	0.36	0.35	0.40	0.34	0.42
$\mathrm{mixN}(0.3,-3,3,1,1)$	0.21	0.21	0.24	0.20	0.24	0.51	0.49	0.54	0.50	0.54
$\mathrm{mixN}(0.7,-3,3,2,2)$	0.14	0.14	0.15	0.13	0.16	0.31	0.30	0.36	0.31	0.37

Table 3: Estimated power of tests with different values of n and $\gamma = 0.05$.

	$n = 100$					$n = 300$				
Alternative distribution	D	V	W^2	U^2	A^2	D	V	W^2	U^2	A^2
Symmetric unimodal distributions										
Ca$(0, 1)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$t(2)$	0.90	0.89	0.95	0.88	0.96	1.00	1.00	1.00	1.00	1.00
$t(3)$	0.61	0.62	0.65	0.61	0.68	0.96	0.95	1.00	0.96	1.00
Lap$(0, 1)$	0.44	0.44	0.50	0.43	0.50	0.92	0.91	0.96	0.92	0.97
Logit$(0, 1)$	0.09	0.09	0.10	0.08	0.10	0.12	0.11	0.15	0.13	0.16
Asymmetric unimodal distributions										
LogN$(0, 0.5)$	0.50	0.52	0.56	0.51	0.61	0.82	0.82	0.85	0.81	0.86
Exp(1)	0.93	0.91	0.96	0.90	0.99	1.00	1.00	1.00	1.00	1.00
Ga$(0.5, 1)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Ga$(2, 1)$	0.71	0.72	0.75	0.70	0.80	1.00	1.00	1.00	0.99	1.00
$\chi^2(2)$	0.91	0.93	0.95	0.92	0.98	1.00	1.00	1.00	1.00	1.00
$\chi^2(4)$	0.50	0.51	0.56	0.52	0.60	0.85	0.87	0.93	0.85	1.00
Gu$(0, 1)$	0.21	0.21	0.25	0.20	0.25	0.40	0.42	0.45	0.40	0.46
SN(0.5)	0.03	0.04	0.02	0.05	0.02	0.01	0.04	0.03	0.03	0.01
SN(2.5)	0.31	0.32	0.35	0.33	0.36	0.55	0.56	0.59	0.55	0.60
Asymmetric bimodal distributions										
BASN(2)	0.76	0.75	0.80	0.74	0.82	1.00	0.98	1.00	0.96	1.00
BASN(6)	0.86	0.87	0.90	0.86	1.00	1.00	1.00	1.00	1.00	1.00
ASLap(2)	0.52	0.51	0.55	0.52	0.60	0.91	0.90	0.95	0.91	1.00
ASLap(6)	0.77	0.76	0.80	0.76	0.82	1.00	1.00	1.00	1.00	1.00
ASLogit(2)	0.43	0.44	0.47	0.42	0.51	0.70	0.72	0.75	0.70	0.80
ASLogit(6)	0.67	0.69	0.72	0.70	0.74	0.95	0.96	1.00	0.97	1.00
mixN$(0.5, -2, 2, 1, 1)$	0.06	0.05	0.04	0.05	0.05	0.06	0.05	0.05	0.05	0.03
mixN$(0.3, -2, 2, 1, 1)$	0.48	0.46	0.50	0.46	0.52	0.65	0.61	0.71	0.63	0.72
mixN$(0.3, -3, 3, 1, 1)$	0.69	0.68	0.72	0.69	0.72	0.80	0.76	0.81	0.78	0.83
mixN$(0.7, -3, 3, 2, 2)$	0.69	0.68	0.70	0.70	0.71	0.81	0.79	0.82	0.78	0.84

Table 4: Estimated power of tests with different values of n and $\gamma = 0.05$ (continued).

5 Application examples

In this section, we illustrate the application of the proposed tests using two datasets, one unimodal and one bimodal, to demonstrate their practical performance in real-world situations.

5.1 Unimodal data

The variable of interest in this dataset is the relative change in the length of corn seeds under compressive stress, referred to as `strain`. To assess the mechanical damage sustained by corn seeds when subjected to compressive forces, an experiment was conducted in which seeds with varying moisture levels and different endosperm types were compressed until rupture occurred. This dataset, originally analyzed by González-Estrada and Cosmes (2019), contains 90 observations (in millimeters) of `strain` measured in corn seeds with floury endosperm and 8% moisture. The histogram of the data is shown in Figure 3.

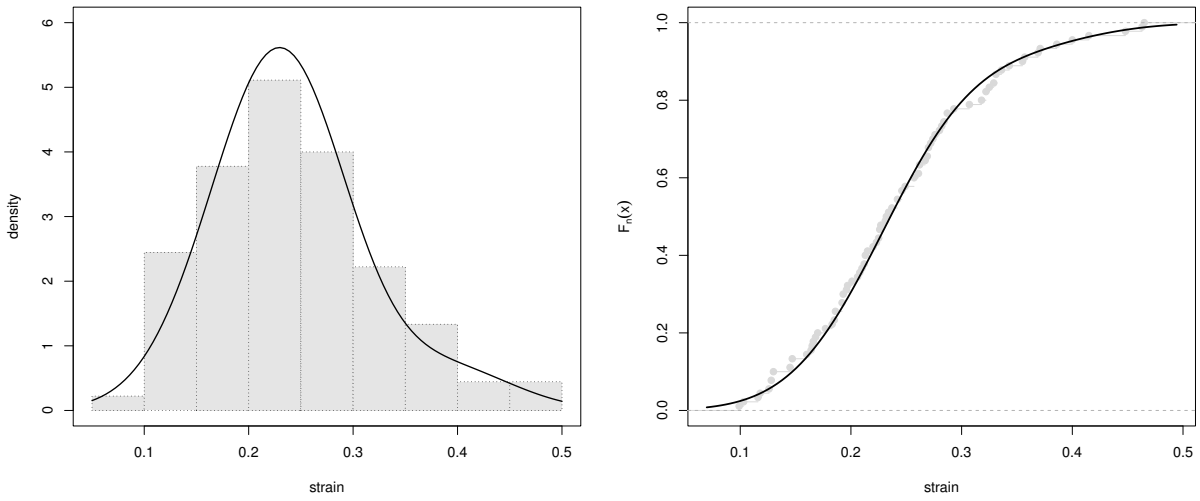


Figure 3: Histogram and empirical distribution function of the `strain` data. The solid line represents the fitted $\mathrm{ASN}(\hat{\mu} = 0.29, \hat{\sigma} = 0.07, \hat{\alpha} = 1.01)$ distribution.

Applying the proposed procedure to this dataset, the test statistic values obtained are

$$\sqrt{n}D_n = 0.5403, \;\; \sqrt{n}V_n = 0.8931, \;\; W_n^2 = 0.0266, \;\; U_n^2 = 0.0264 \;\text{ and }\; A_n^2 = 0.2272.$$

The corresponding 5% significance level quantiles for each test statistic are 0.9081, 1.4247, 0.1338, 0.1051 and 0.7568, respectively. As all computed test statistic values fall below their critical quantiles, the hypothesis of alpha-skew-normality is not rejected. This suggests that the ASN distribution provides a plausible model for describing the stochastic behavior of the data. This conclusion is further supported by Figure 3, which demonstrates a good agreement between the empirical distribution function of the sample and the fitted ASN distribution function.

5.2 Bimodal data

The following example consists of 63 observations of the breaking strength of 1.5 cm-long glass fibers, obtained by the National Physical Laboratory in the United Kingdom. These data can be found, for example, in the text by Jones and Pewsey (2009). The histogram in Figure 4 reveals a small bump, which may indicate the presence of bimodality.

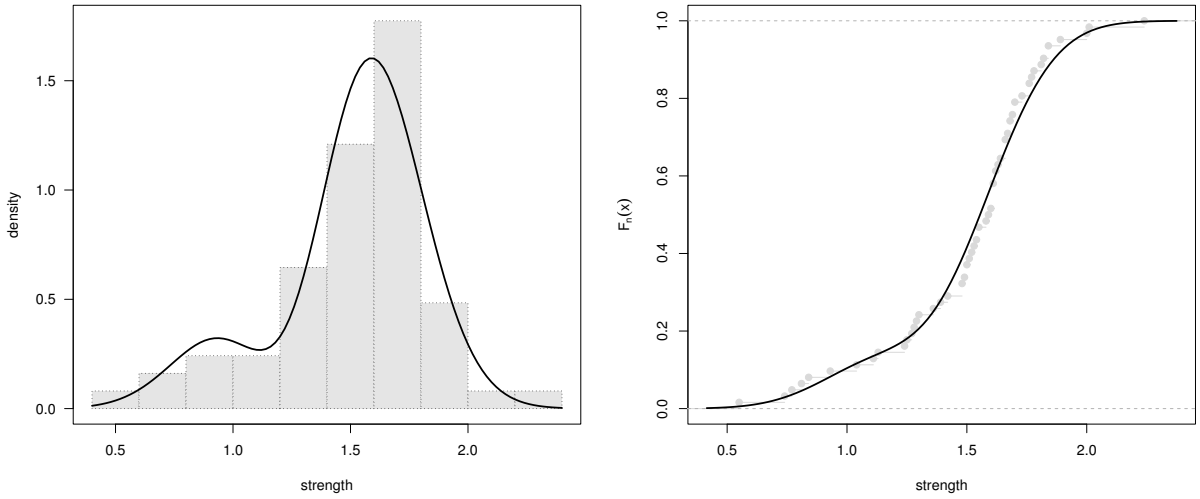


Figure 4: Histogram and empirical distribution function of the `strength` data. The solid line represents the fitted $\text{ASN}(\hat{\mu} = 1.32, \hat{\sigma} = 0.26, \hat{\alpha} = -1.58)$ distribution.

Applying the proposed procedure to the `strength` data, the observed values of the test statistics are

$$\sqrt{n}D_n = 0.7625, \quad \sqrt{n}V_n = 1.2784, \quad W_n^2 = 0.0830, \quad U_n^2 = 0.0788 \text{ and } A_n^2 = 0.4364.$$

At the 5% significance level, the critical quantiles for each test statistic are 0.9046, 1.4114, 0.1344, 0.1041 and 0.7533, respectively. Since all observed test statistics fall below their corresponding quantiles, the results support the alpha-skew-normality hypothesis, suggesting that the ASN distribution provides an appropriate approximation to the frequency distribution of the data. This conclusion is further supported by Figure 4, which shows a good agreement between the empirical distribution function of the data and the fitted ASN distribution function.

6 Conclusions

This paper presents a general testing procedure based on classical goodness-of-fit tests for assessing the validity of any member of the $\text{ASN}(\mu, \sigma, \alpha)$ family of distributions when the parameters μ, σ and α are unknown. The null distribution of each test statistic was approximated using its prior predictive distribution, effectively eliminating dependence on the shape parameter.

The Monte Carlo simulation study provides evidence that the type I error probability of the tests does not exceed the nominal significance level, indicating that the use of the five proposed tests is statistically valid in terms of maintaining the intended significance level. Furthermore, the tests

demonstrate strong power against the considered alternative distributions. Among the tests studied, those based on the W^2 and A^2 statistics demonstrate appreciably higher power.

References

Armagan, A, D B Dunson, and J Lee (2013). Generalized double pareto shrinkage. *Statistica Sinica 23*, 119–143.

Azzalini, A (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics 12*, 171–178.

Bayarri, M J and J O Berger (2000). p-values for composite null models. *Journal of the American Statistical Association 95*(452), 1127–1142.

Box, G (1980). Sampling and bayesian inference in scientific modeling and robustness. *Journal of the Royal Statistical Society 143*, 383–430.

Cabras, Samuel and María Eugenia Castellanos (2009). Default bayesian goodness-of-fit tests for the skew-normal model. *Journal of Applied Statistics 36*(2), 223–232.

Chakraborty, S and P J Hazarika (2014). Alpha-skew-logistic distribution. *Journal of Mathematics 10*(4), 36–46.

Chakraborty, S, S Shah, and P J Hazarika (2014). *Balakrishnan-alpha-skew-normal distribution: properties and applications*. Dibrugarh, Assam, India: Department of Statistics, Dibrugarh University.

Chen, G and N Balakrishnan (1995). A general purpose approximate goodness-of-fit test. *Journal of Quality Technology 27*(2), 154–161.

Elal-Olivero, David (2010). Alpha-skew-normal distribution. *Proyecciones Journal of Mathematics 29*(3), 224–240.

González-Estrada, E and W Cosmes (2019). Shapiro–wilk test for skew normal distributions based on data transformations. *Journal of Statistical Computation and Simulation 89*(17), 3258–3272.

Gui, W (2014). A generalization of the slashed distribution via alpha-skew-normal distribution. *Statistical Methods and Applications 23*(4), 547–563.

Harandi, S S and M Alamatsaz (2013). Alpha-skew-laplace distribution. *Statistics and Probability Letters 83*, 774–782.

Jones, M C and A Pewsey (2009). Sinh-arcsinh distributions. *Biometrika 96*(6), 761–780.

Lilliefors, H W (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association 62*(318), 399–402.

Louzada, F and A Ara (2019). The multivariate alpha-skew-normal distribution. *Bulletin of the Brazilian Mathematical Society 50*, 823–843.

Louzada, F, A Ara, and G Fernandes (2016). Bivariate alpha-skew-normal distribution. *Communications in Statistics - Theory and Methods 46*(12), 6098–6111.

Mateu-Figueras, G, P Puig, and A Pewsey (2009). Goodness-of-fit tests for the skew-normal distribution when the parameters are estimated from the data. *Communications in Statistics - Theory and Methods 36*(2), 1735–1755.

Meintanis, S G (2007). A kolmogorov–smirnov type test for skew normal distributions based on the empirical moment generating function. *Journal of Statistical Planning and Inference 137*, 2681–2688.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Stephens, M (1974). Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association 69*(347), 730–737.

Stephens, M (1986). Test based on edf statistics. in *Goodness-of-fit techniques*. marcel dekker. pp. 97–193.

7 Appendix

```
# Function that calculates the density of the alpha-skew-normal
# distribution of location-scale:
dasnls <- function(x, mu, sigma, alpha){
  w <- (1/sigma)*(x - mu)
  num <- (1 - alpha*w)^2 + 1
  den <- sigma*(2 + alpha^2)
  res <- (num/den)*dnorm(w)
  return(res)
}

# Function that calculates the cumulative density of the
# alpha-skew-normal distribution of location-scale:
pasnls <- function(x ,mu, sigma, alpha){
  w <- (1/sigma)*(x - mu)
  num <- 2*sigma - alpha*(x - mu)
  den <- sigma*(2 + alpha^2)
  res <- pnorm(w) + alpha*dnorm(w)*(num/den)
  return(res)
}

# Function that fits a alpha-skew-normal distribution of location-scale
# to data, using maximum likelihood estimation:
library(moments)
mle_asn <- function(data){
  ll <- function(par){(-1)*sum(log(dasnls(data, mu=par[1], sigma=par[2],
                                          alpha=par[3])))}
  if (skewness(data) > 0){
    init1 <- c(mean(data), sd(data), 1)
    init2 <- c(mean(data), sd(data), 5)
    init3 <- c(mean(data), sd(data), 10)
    opt1 <- optim(init1, ll) ; opt2 <- optim(init2, ll)
    opt3 <- optim(init3, ll)
    indmin <- which.min(c(opt1$value, opt2$value, opt3$value))
```

```
    if(indmin == 1) out <- opt1$par ; if(indmin == 2) out <- opt2$par
    if(indmin == 3) out <- opt3$par
  } else {
    init1 <- c(mean(data), sd(data), -1)
    init2 <- c(mean(data), sd(data), -5)
    init3 <- c(mean(data), sd(data), -10)
    opt1 <- optim(init1, ll) ; opt2 <- optim(init2, ll)
    opt3 <- optim(init3, ll)
    indmin <- which.min(c(opt1$value, opt2$value, opt3$value))
    if(indmin == 1) out <- opt1$par ; if(indmin == 2) out <- opt2$par
    if(indmin == 3) out <- opt3$par
  }
  return(out)
}

# Function that obtains, for a set of data, the value of empirical
# distribution function statistic with respect to the alpha-skew-normal
# distribution:
edfs_asn <- function(data, mu, sigma, alpha){
  n <- length(data)
  xsort <- sort(data)
  pos <- 1:n
  z <- pasnls(xsort, mu, sigma, alpha)
  Dp <- max((pos/n) - z)
  Dm <- max(z - ((pos - 1)/n))
  D <- sqrt(n)*max(Dp, Dm)
  V <- sqrt(n)*(Dp + Dm)
  W2 <- sum((z - ((2*pos - 1)/(2*n)))^2) + (1/(12*n))
  U2 <- W2 - (n*(mean(z) - 0.5)^2)
  aux <- sapply(1:n, function(i){1 - z[n + 1 - i]})
  A2 <- -n - (1/n)*sum((2*pos - 1)*(log(z) + log(aux)))
  return(round(c(D, V, W2, U2, A2), 4))
}

# Function that performs linear interpolation to obtain the constant
# critical:
# tab: table of critical constants.
# n: size of the sample in turn.
interp <- function(tab, n){
  ns <- c(5, 10, 20, 30, 40, 50, 75, 100, 150, 200, 300, 400, 500)
  pos_upp <- 1
  while (n > ns[pos_upp]){
    pos_upp <- pos_upp + 1
  }
  pos_low <- length(ns)
  while (n < ns[pos_low]){
    pos_low <- pos_low - 1
  }
  q_upp <- tab[pos_upp, 3:7]
  q_low <- tab[pos_low, 3:7]
  aux <- ((n - ns[pos_low])/(ns[pos_upp] - ns[pos_low]))
  q_aux <- aux*(q_upp - q_low) + q_low
  return(q_aux)
```

```
}

# Table of critical constants (Table 1):
col1 <- rep(c(0.90, 0.95, 0.99), each=13)
col2 <- rep(ns, times=3)
col3 <- c(0.6911, 0.7384, 0.7856, 0.8084, 0.8225, 0.8234, 0.8236, 0.8241,
          0.8261, 0.8293, 0.8323, 0.8349, 0.8371, 0.7541, 0.8172, 0.8597,
          0.8839, 0.9096, 0.9043, 0.9051, 0.9102, 0.9132, 0.9152, 0.9159,
          0.9174, 0.9184, 0.8833, 0.9578, 1.0129, 1.0585, 1.0685, 1.0862,
          1.0917, 1.0921, 1.0919, 1.0922, 1.0942, 1.0936, 1.0963)
col4 <- c(1.1166, 1.1856, 1.2521, 1.2836, 1.3032, 1.3196, 1.3245, 1.3244,
          1.3289, 1.3424, 1.3461, 1.3444, 1.3467, 1.1976, 1.2721, 1.3456,
          1.3791, 1.3966, 1.4006, 1.4212, 1.4227, 1.4377, 1.4427, 1.4411,
          1.4435, 1.4446, 1.3359, 1.4681, 1.5426, 1.5652, 1.6062, 1.6112,
          1.6196, 1.6333, 1.6465, 1.6467, 1.6465, 1.6528, 1.6501)
col5 <- c(0.0807, 0.0898, 0.0993, 0.1034, 0.1063, 0.1077, 0.1089, 0.1091,
          0.1098, 0.1101, 0.1109, 0.1096, 0.1105, 0.0982, 0.1108, 0.1219,
          0.1302, 0.1386, 0.1347, 0.1341, 0.1356, 0.1365, 0.1379, 0.1369,
          0.1408, 0.1406, 0.1417, 0.1674, 0.1752, 0.1878, 0.2011, 0.2097,
          0.2122, 0.2158, 0.2162, 0.2167, 0.2166, 0.2193, 0.2177)
col6 <- c(0.0718, 0.0756, 0.0805, 0.0824, 0.0849, 0.0858, 0.0867, 0.0869,
          0.0858, 0.0868, 0.0875, 0.0877, 0.0875, 0.0879, 0.0909, 0.0987,
          0.1005, 0.1034, 0.1043, 0.1049, 0.1058, 0.1056, 0.1062, 0.1069,
          0.1067, 0.1069, 0.1215, 0.1308, 0.1393, 0.1426, 0.1513, 0.1511,
          0.1518, 0.1518, 0.1568, 0.1573, 0.1594, 0.1599, 0.1594)
col7 <- c(0.4853, 0.5409, 0.5842, 0.6131, 0.6251, 0.6257, 0.6293, 0.6298,
          0.6303, 0.6326, 0.6318, 0.6394, 0.6396, 0.5602, 0.6394, 0.6964,
          0.7382, 0.7544, 0.7541, 0.7565, 0.7596, 0.7631, 0.7645, 0.7682,
          0.7719, 0.7801, 0.7116, 0.9011, 0.9345, 1.0061, 1.1189, 1.1215,
          1.1381, 1.1475, 1.1496, 1.1483, 1.1468, 1.1503, 1.1499)
tabq <- data.frame(col1, col2, col3, col4, col5, col6, col7)
colnames(tabq) <- c("P", "n", "D", "V", "W2", "U2", "A2")

# Function that obtains the critical constant corresponding to each test:
# tab: table of critical constants.
# sig: significance level of the test (0.01, 0.05, 0.10).
# naux: size of the sample in turn.
qtab <- function(tab, sig, naux){
  tab_aux <- tab[which((1 - sig) == tab$P),]
  if (naux %in% tab_aux$n){
    q_aux <- tab_aux[which(naux == tab_aux$n), 3:7]
  } else {
  if (naux > 500){
    q_aux <- tab_aux[which(500 == tab_aux$n), 3:7]
  } else {q_aux <- interp(tab_aux, naux)}
  }
  return(q_aux)
}

# Function that performs, from a set of data, the test for
# alpha-skew-normality:
# sig: significance level of the test (0.01, 0.05, 0.10).
# pv: boolean object that takes the value of TRUE if the
```

```
#     p-value of the test is desired, otherwise the value
$     of FALSE is specified.
# The routine returns the following data:
# edfsobs: observed value of each test statistic.
# quantile: value of the critical constant for each test.
# rejectH0: boolean object that takes the value of TRUE if H0 is rejected,
#           otherwise FALSE means that it is not rejected.
# mles: maximum likelihood estimates for the parameters.
# pvalue: p-value for each test.
edfstest_asn <- function(data, sig, pv){
  naux <- length(data)
  mle <- mle_asn(data)
  tobs <- edfs_asn(data, mle[1], mle[2], mle[3])
  qp <- qtab(tabq, sig, naux)
  test <- tobs > qp
  if (pv == FALSE){
    print(list(edfsobs=tobs, quantile=qp, rejectH0=test, mles=mle))
  }
  if (pv == TRUE){
    M <- 1E3
    aux <- edfsdist_asn(naux, rcauchy(M, 0, 3/2))
    pv <- apply(sapply(1:M, function(j){aux[,j] > tobs}), 1, mean)
    print(list(edfsobs=tobs, quantile=qp, rejectH0=test, mles=mle,
               pvalue=pv))
  }
}
```

Spanish Journal of Statistics
Vol. **xx** No. xx 2025, Pages 131–139
doi:https://doi.org/10.37830/
SJS.2025.1.07

Invited Article

Address at the 2024 Spanish National Award in Statistics

Concha Bielza

Universidad Politécnica de Madrid, mcbielza@fi.upm.es

Received: November 5, 2025. Returned: November 25, 2025.

Abstract: This article is based on the address given upon receiving the 2024 Spanish national award in statistics, a ceremony made especially meaningful by the attendance of His Majesty the King of Spain. It offers a personal overview of my research trajectory, shaped by the long-standing interplay between statistics and artificial intelligence, and by their applications in neuroscience and industry. After beginning my career in statistical decision theory and probabilistic graphical models, Bayesian networks soon became the central framework of my work, enabling rigorous reasoning under uncertainty across domains.

In neuroscience, my contributions span neuronal classification, spatial analysis of synapses, modeling of dendritic arborizations, biomarker discovery for neurological disorders, and the decoding of brain activity, among others. These efforts, supported by landmark programs such as the Cajal Blue Brain Project and the Human Brain Project, were driven by the need for models capable of capturing complex, high-dimensional, and often unconventional data.

Around 2018, my research turned increasingly toward Industry 4.0, where real-time data streams, dynamic systems, and predictive and prescriptive maintenance posed demanding methodological challenges. This led to advances in dynamic Bayesian networks, latent-variable models, and probabilistic evolutionary algorithms for high-dimensional optimization.

Alongside scientific discovery, I have remained committed to teaching, mentoring, and knowledge transfer through initiatives such as the Machine Learning and Advanced Statistics Summer School at the Universidad Politécnica de Madrid. I conclude with reflections on interdisciplinary work, convergence of statistics and machine learning, ethical use of data and algorithms, and the importance of inspiring new generations of statisticians.

Keywords: Bayesian networks, probabilistic machine learning, Bayesian decision theory, reasoning under uncertainty, heuristic optimization, temporal data, interpretable models, artificial intelligence, neuroscience, industry 4.0

MSC: 62-09, 62P99, 68T37, 90B50

Published by the Spanish National Statistical Institute

1 Introduction

This article offers a written version of the speech I delivered on April 7, 2025, the day I received the 2024 Spanish national award in statistics. I am grateful to the editor of the *Spanish Journal of Statistics* for the invitation to share it with the journal's readers. The ceremony took place at the Royal Spanish Academy of Exact, Physical, and Natural Sciences in Madrid. The event was made especially memorable by the presence of His Majesty King Felipe VI of Spain.

The video of the ceremony can be seen in https://www.youtube.com/watch?v=3wqZ8h6OrYg.

2 Opening remarks

Good morning. Your Majesty, Secretary of State, President of the Royal Spanish Academy of Exact, Physical, and Natural Sciences, President of the National Statistics Institute, authorities, guests, colleagues, family members, and friends.

I would like to begin by expressing my gratitude to His Majesty for his presence, which elevates this ceremony and lends even greater significance to this event. To the Secretary of State, for the institutional support he represents. And to this Royal Academy, for welcoming us into this magnificent home as host of this occasion.

My special thanks go to the National Statistics Institute for establishing this prestigious award, which highlights the value of statistics as an essential tool for informed decision-making and encourages us to continue developing methodologies that transform data into useful knowledge for society.

It is a great honor to receive this award, which recognizes years of effort and dedication and brings visibility to my work. The stature of previous laureates—three of whom are here today—confirms its relevance (see Figure 1).

Figure 1: His Majesty King Felipe VI (center), together with four of the five laureates—shown from left to right: 2022 (Enrique Castillo), 2021 (Wenzeslao González), 2024 (the author, Concha Bielza), and 2020 (Daniel Peña)—and the Presidents of the National Statistics Institute and of the Royal Academy of Sciences.

My thanks also go to the selection committee for granting me this distinction, to those who supported my candidacy with their letters, and to those who encouraged me to apply. I am likewise grateful to the other candidates, who help keep interest in this award alive. And thank you, Pedro, for your words in the *laudatio* and for being, amid the daily whirlwind, a true friend and my best companion in both hardships and achievements over many years (see Figure 2).

Figure 2: Moments from the ceremony: Pedro Larrañaga delivering the *laudatio* (left), and the applause following the award ceremony, with His Majesty King Felipe VI, the Secretary of State, and the President of the National Statistics Institute (right).

3 Foundations of my research career

My path in statistics began during my undergraduate studies in Mathematics at the Universidad Complutense de Madrid. After finishing my degree, I taught at the beautiful Royal University Center María Cristina in San Lorenzo de El Escorial (see Figure 3, left)—a private university that is part of the famous Monastery—and I also worked in a major Spanish company in the field of computing, where, without realizing it at the time, I was already moving toward artificial intelligence through what were then known as "expert systems."

In 1991, I joined the School of Computer Engineering (then still the Faculty of Computer Science) at the Universidad Politécnica de Madrid (see Figure 3, right), where I began my doctoral studies in the Department of Artificial Intelligence, which included the area of statistics and operations research. This was thanks to Sixto and David Ríos Insua, whose support was fundamental in my early career.

With David—an Academician of this Royal Academy—I carried out my PhD on decision-making under uncertainty within Bayesian decision theory. We represented the decision-making problems using influence diagrams (see Figure 4), a then-recent probabilistic graphical model, and extended them to contexts under partial information.

We applied these ideas in a decision-support system for neonatal jaundice, developed in collaboration with the Gregorio Marañón Hospital, which enabled less invasive treatments and improved clinical understanding. This work gave rise to two PhD dissertations that I supervised: one on modeling and another on explanation generation— a topic we were already addressing in the early 2000s

Figure 3: Royal University Center María Cristina in San Lorenzo de El Escorial (left) and the School of Computer Engineering at the Universidad Politécnica de Madrid (right).

and which is now reflected in European artificial intelligence regulation through the so-called right to a clear and comprehensible explanation, essential for building trust in algorithmic decisions.

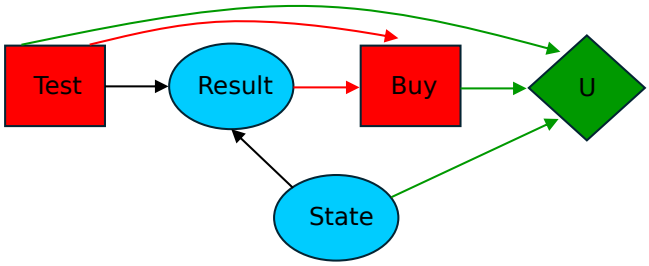
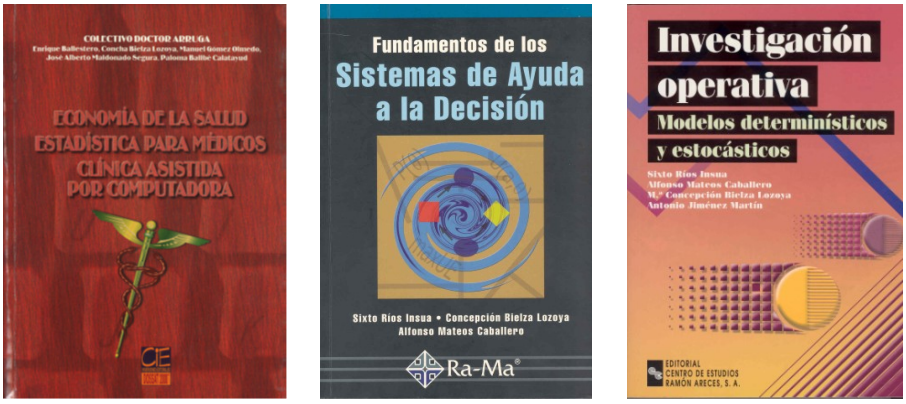


Figure 4: The structure of an influence diagram.

I also wrote several textbooks—on statistics for medical doctors, on decision-support systems, and on operations research (see Figure 5). I worked on decision-making problems across different sectors, and together with the School of Agricultural Engineering at the Universidad Politécnica de Madrid, we developed predictive models in the agri-food domain to estimate the probability of fruit damage as it moved along grading lines, which led to registered software that attracted the interest of cooperatives.

I consolidated my research independence and collaborated with international colleagues, such as Prakash Shenoy. We presented joint works at one of the early editions of AISTATS—then still a workshop—the seed of today's community that brings together statistics and artificial intelligence. I was fascinated to see how both worlds were coming to meet each other: AI was recognizing the need to incorporate uncertainty, and statistics was relying on increasingly powerful computational methods. At this intersection, probabilistic graphical models—particularly Bayesian networks—were beginning to stand out. These networks have been the cornerstone of my research. They represent uncertain knowledge through interpretable graphs and enable probabilistic reasoning of any kind.

Judea Pearl, their principal proponent, has been described as "the most original and influential thinker in statistics today." The influence diagrams that shaped my early research could be seen as their informal precursors, since Bayesian networks offered a more powerful formalism by explicitly incorporating the semantics of conditional independence, thus linking graphs and probabilities.



1999 2002 2004

Figure 5: First textbooks authored.

My first applications of Bayesian networks were in predicting surface roughness in high-speed machining, in collaboration with the Institute of Industrial Automation of the Spanish National Research Council (CSIC), a project that led to the third PhD dissertation I supervised. Indeed, some years later, industry would become central to my research.

4 Towards a data-driven neuroscience

The arrival of Pedro Larrañaga at the Universidad Politécnica de Madrid at the end of 2007 marked the beginning of my next stage, which was undoubtedly a fruitful one. Together, we created the Computational Intelligence Group from scratch, with a strong commitment to computational neuroscience, just when our university was joining ambitious international initiatives focused on the study of the brain—one of the great scientific challenges of the 21st century. I will highlight some milestones achieved over more than 15 years of work in neuroscience, mainly within two major projects: the Cajal Blue Brain Project and the Human Brain Project.

In neuroanatomy, we addressed the classification and nomenclature of neurons—a century-old debate—on the basis of their morphology, electrophysiology, and transcriptome, achieving high-impact publications in journals of the *Nature* group. It required a significant effort to establish the value of data-driven models within a community not accustomed to using them.

We also explored the spatial organization of synapses and dendritic spines through spatial statistics—barely used in this field—to investigate whether their arrangement follows structured patterns or is essentially random. Another challenge was to test the hypothesis that neuronal arborizations are designed to maximize connectivity while minimizing wiring length.

It was a real gift to work with the renowned neuroscientists Javier deFelipe (CSIC) and Rafael Yuste (Columbia University), see Figure 6.

In neurological disorders such as Alzheimer's disease, Parkinson's disease, and epilepsy, we worked on the identification of biomarkers, the definition of clinical subtypes, and the prediction of disease progression and quality of life—the latter based on multiple dimensions (see Figure 7). We used highly diverse data—genetic, clinical, neuropsychological, imaging, and questionnaire-based—and collaborated with hospitals and research centers, both national and international. We

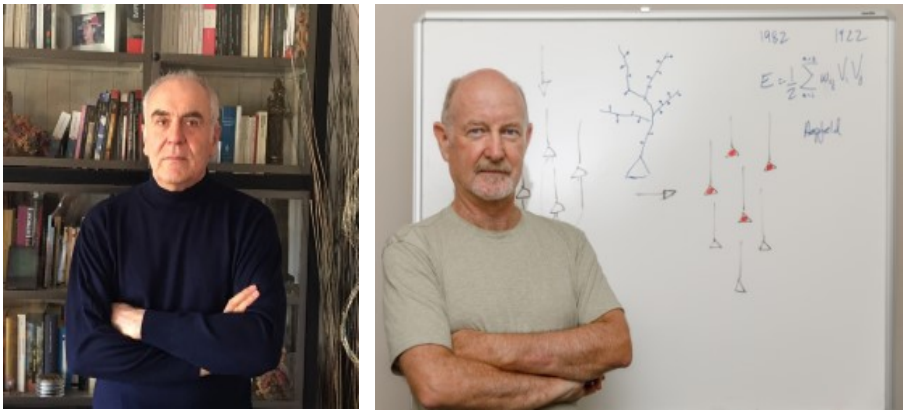


Figure 6: Javier de Felipe (CSIC) (left) and Rafael Yuste (Columbia University) (right).

also worked on classifying mental activity from magnetoencephalography recordings, contributing to the so-called brain decoding, a key line of research in brain–computer interfaces.

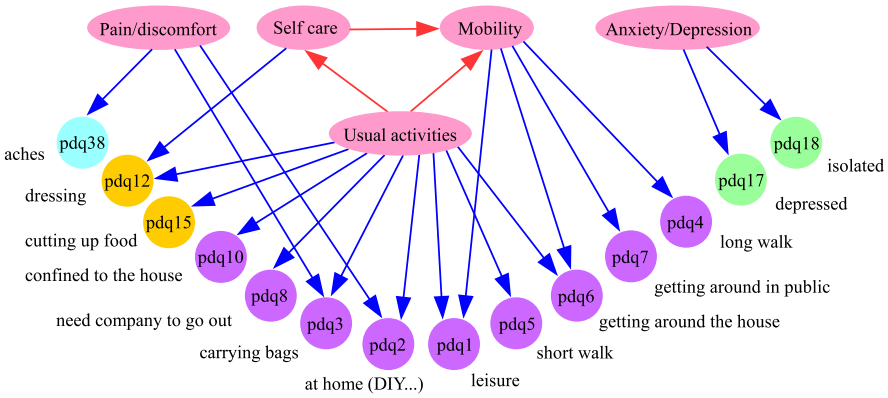


Figure 7: A Bayesian network to predict the European quality of life-5 dimensions (EQ-5D, in pink) from the 39-item Parkinson's disease questionnaire (PDQ-39).

In neuroscience, the lack of data-driven models led me to embark, with Pedro, on the writing of a book published by Cambridge University Press in 2021, which brings together methods from statistics and machine learning and their applications to real data (see Figure 8, left). After more than six years of work and nearly 700 pages, a dear collaborator—who is here today—suggested marking the occasion with steaks matching the weight of the book—almost one and a half kilos. The image has stayed with us ever since... although the meal has yet to be served.

The applications have been highly rewarding because of their real-world impact, but behind each of them there is, more often than not, a methodological gap that needed to be addressed, leading us to develop our own solutions, particularly in the area of Bayesian networks. We have often dealt with atypical types of data: directional data (such as neuronal branching angles); non-Gaussian data (common in real applications); or datasets with far more variables than observations (as in genomics).

We have also faced challenges such as noisy labels (as in the long-standing, inconsistent nomenclature of neurons); multiple interrelated classes (as in the multidimensional assessment of quality of life in Parkinson's disease); model consensus (for example, when integrating the opinions of several experts); computationally efficient learning; the construction of regularized—and even mas-

Figure 8: Recent textbooks authored.

sive—networks; the identification of the most probable or most relevant explanations for a given piece of evidence; and the algebraic characterization of decision boundaries.

5 Probabilistic modeling for Industry 4.0

Around 2018, my research began to shift toward the industrial sector, within the framework of the so-called Industry 4.0, in which sensors, machines, and people collaborate in interconnected systems that require real-time analysis. Time series and data streams arrive at high sampling rates, in the form of electrical, mechanical, or environmental signals, reflecting complex processes that evolve over time.

Working hand in hand with several companies, we have addressed multiple dynamic problems, where the aim is to monitor the operational state of assets, anticipate failures (predictive maintenance), and determine the best course of action (prescriptive maintenance).

Among these applications, I will highlight two. The first concerns industrial furnaces with tubes through which a thermal fluid circulates. The progressive accumulation of impurities on their walls hampers heat transfer and forces an increase in energy consumption to maintain the desired temperature of the fluid. If no action is taken in time, the thermal limit of the material may be reached, causing damage. Using dynamic Bayesian networks, we were able to anticipate several days in advance when a section of the furnace would require cleaning, thereby optimizing maintenance and avoiding inefficiencies or risks.

The second application focuses on estimating the remaining useful life of bearings used in machine tools (Figure 9) and in pumps at desalination plants, whose progressive degradation generates signals that change in complex ways. We developed a model with latent variables—variables that are not directly observed—that automatically captures the different phases of deterioration and adapts to the evolution of the signal.

In all these contexts, we have designed increasingly complex and adaptive dynamic Bayesian networks. It is particularly important to identify different operating modes, detect anomalies, or determine when a model ceases to be valid because the underlying process has changed.

Figure 9: A degraded bearing (image courtesy of Aingura IIoT).

To strengthen this new industrial orientation, I became involved in another book, this one on *Industrial Applications of Machine Learning*, published by CRC Press in 2019 and translated into Chinese in 2023 (see Figure 8, right).

In some of the problems we addressed, it was necessary to optimize complex, multi-objective functions, often in immense spaces where exact methods are not feasible. One example is the optimal design of a biolubricant for MotoGP racing, with millions of possible ingredients for the formulation. Here, we contributed to the development of stochastic evolutionary algorithms based on Bayesian networks—so-called estimation of distribution algorithms—which, although they provide no guarantee of optimality, make it possible to find good solutions and to better understand the optimization process.

6 Reflections and gratitude

Research is not only about discovery; it is also about teaching and sharing. I have maintained that commitment not only in the university classroom, but also in other professional settings, such as with civil servants from the Ministry of Education, the Directorate-General for the Cadastre, the Carlos III Health Institute, or the Cervantes Institute in Berlin. But if I had to highlight one initiative, it would be the Machine Learning and Advanced Statistics Summer School, which I have co-directed for many years at the Universidad Politécnica de Madrid and which is updated regularly to reflect current advances (see Figure 10).

Before closing, I would like to highlight several ideas that have been central to my career. First, the importance of interdisciplinary work in addressing today's major challenges with a richer and more complementary perspective. Second, the value of building bridges between statistics and machine learning—as Bayesian networks do—to approach the analysis of complex modern data with rigor and efficiency. Third, the need to ensure the ethical use of data and algorithms, placing people at the center and promoting statistical literacy; in this spirit, I encourage signing the manifesto promoted by the Spanish Society of Statistics and Operations Research (SEIO). And finally, the importance of a strong commitment to technology transfer, from the university to the industrial ecosystem.

I conclude these words with my deepest gratitude to those who have accompanied me on this journey, for this award belongs to them as well. To the Universidad Politécnica de Madrid, for allowing me to develop my career freely in an environment committed to research and technology transfer. To my collaborators, institutions, and companies with whom I have worked, with the hope of continuing to move forward together. To my doctoral students—past and present—for their talent,

Figure 10: Machine Learning and Advanced Statistics Summer School at the Universidad Politécnica de Madrid.

dedication, and trust, even at a time when job offers outside academia are more tempting than ever. To Laura, our program coordinator, for her constant efficiency and unfailingly positive attitude, even when facing bureaucracy. And to my friends and family, for their endless support, understanding, and love throughout the journey of life.

Allow me to mention three special people from an earlier generation, from whom one always learns. My aunt Concha—Concha Bielza, like me—who is here today and who, at almost 98 years old, remains a model of vitality and affection. And my parents, who always fostered an environment of study and instilled in my three sisters and me a culture of effort. My father would have loved to witness this moment, even if, back in the day, he advised me against studying Mathematics ("What for?"), preferring that I choose an engineering degree as he had done. And my mother, who has been able to accompany me today, making a great effort simply because she loves me.

I hope this award inspires new generations of statisticians—women and men alike—to pursue excellence, to work responsibly, and to contribute with enthusiasm to the progress of our discipline.

THANK YOU VERY MUCH.

Acknowledgments

This work was partially supported by the MINISTRY OF SCIENCE, INNOVATION AND UNIVERSITIES [Project 482 AEI/10.13039/501100011033-PID2022-139977NB-I00, Project PLEC2023-010252/MIG- 483 20232016]. Also, by the AUTONOMOUS REGION OF MADRID [Project 484 ELLIS Unit Madrid and TEC-2024/COM-89].

Acknowledgement to Reviewers SJS vol. 2025

The Editors of Spanish Journal of Statistics gratefully acknowledge the assistance of the following people, who reviewed manuscripts:

Manuela Alcañiz University of Barcelona, Spain

Catalina Bolancé, University of Barcelona, Spain

Vanesa Jordá, Department of Economics, University of Cantabria, Spain

Julio Revuelta, Department of Economics, University of Cantabria, Spain

Mercedes Tejería, Department of Economics, University of Cantabria, Spain