Statistica Applicata • Italian Journal of Applied Statistics

Volume 37 • Number 1 • 2025

Statistica Applicata

ISSN: 1125-1964

Volume 37
Number 1

ITALIAN JOURNAL OF APPLIED STATISTICS

FIRENZE
UNIVERSITY
PRESS

Statistica Applicata – ITALIAN JOURNAL OF APPLIED STATISTICS

Salvatore Ingrassia
Università di Catania, Catania, Italy
https://orcid.org/0000-0003-2052-4226
salvatore.ingrassia@unict.it

Ron S. Kenett
KPA Ltd. and Samuel Neaman Institute, Technion, Haifa, Israel
https://orcid.org/0000-0003-2315-0477
ron@kpa-group.com

Eric Marlier
LISER, Luxembourg Institute of Socio-Economic Research
Université de Luxembourg, Luxembourg
https://orcid.org/0000-0002-5559-3689
eric.marlier@liser.lu

Stefania Mignani
Università di Bologna Alma Mater, Bologna, Italy
https://orcid.org/0000-0003-4746-1130
stefania.mignani@unibo.it

Tormod Naes
NOFIMA, Oslo, Norway
https://orcid.org/0000-0001-5610-3955
tormod.naes@nofima.no

Alessandra Petrucci
Università di Firenze, Florence, Italy
https://orcid.org/0000-0001-9952-0396
alessandra.petrucci@unifi.it

Monica Pratesi
Università di Pisa, Pisa, Italy
https://orcid.org/0000-0002-9959-5483
monica.pratesi@unipi.it

Maurizio Vichi
Sapienza Università di Roma, Rome, Italy
https://orcid.org/0000-0002-3876-444X
maurizio.Vichi@uniroma1.it

Matilde Bini
Università Europea, Rome, Italy
https://orcid.org/0000-0003-1989-8685
matilde.bini@unier.it

Giovanna Boccuzzo
Università di di Padova, Padua, Italy
https://orcid.org/0000-0003-2143-7730
giovanna.boccuzzo@unipd.it

John Castura
Compusense Inc., Guelph, Canada
https://orcid.org/0000-0002-1640-833X
jcastura@compusense.com

Maurizio Carpita
Università di di Brescia, Brescia, Italy
https://orcid.org/0000-0001-7998-5102
maurizio.carpita@unibs.it

Carlo Cavicchia
Erasmus Rotterdam University, Netherlands
https://orcid.org/0000-0003-1816-3521
cavicchia@ese.eur.nl

Carolina Chaya
Universitat Politecnica de Madrid, Spain
https://orcid.org/0000-0002-5518-886X
carolina.chaya@upm.es

Alessandro Celegato
AICQ Centronord - Quality and Techonology in production
alessandro.celegato@gmail.com

Giuliana Coccia
ASVIS, Rome, Italy
giuliana.coccia1@gmail.com

Fabio Crescenzi
Independent researcher, Italy
https://orcid.org/0000-0002-2660-6209
fabio7826@gmail.com

Cristina Davino
Università di Napoli Federico II, Naples, Italy
https://orcid.org/0000-0003-1154-4209
cristina.davino@unina.it

Tonio Di Battista
Università di Chieti-Pescara “Gabriele D’Annunzio”, Pescara, Italy
https://orcid.org/0000-0003-2139-7273
tonio.dibattista@unich.it

Francesca Di Iorio
Università di Napoli Federico II, Naples, Italy
https://orcid.org/0000-0002-9586-3380
francesca.diiorio@unina.it

Simone Di Zio
Università di Chieti-Pescara “Gabriele D’Annunzio”, Pescara, Italy
https://orcid.org/0000-0002-9139-1451
simone.dizio@unich.it

Filippo Domma
Università della Calabria, Rende, Italy
https://orcid.org/0000-0002-1489-1065
f.domma@unical.it

Angela M. D’Uggento
Università di Bari, Bari, Italy
https://orcid.org/0000-0001-9768-651X
angelamaria.duggento@uniba.it

Monica Ferraroni
Università di Milano, Milan, Italy
https://orcid.org/0000-0002-4542-4996
monica.ferraroni@unimi.it

Livio Finos
Università di Padova, Padova, Italy
https://orcid.org/0000-0003-3181-8078
livio.finos@unipd.it

Giuseppe Giordano
Università di Salerno, Salerno, Italy
https://orcid.org/0000-0002-0582-109X
ggiordano@unisa.it

Michela Gnaldi
Università di Perugia, Perugia, Italy
https://orcid.org/0000-0002-2785-3279
michela.gnaldi@unipg.it

Paolo Girardi
Università Ca' Foscari, Venezia, Italy
https://orcid.org/0000-0001-8330-9414
paolo.girardi@unive.it

Domenica Fioredistella Iezzi
Università di Roma Tor Vergata, Rome, Italy
https://orcid.org/0000-0001-8041-2846
stella.iezzi@uniroma2.it

Giuseppe Lamberti
Universitat Autonoma de Barcelona, Spain
https://orcid.org/0000-0002-8666-796X
giuseppe.lamberti@uab.cat

Filomena Maggino
Sapienza Università di Roma, Italy
https://orcid.org/0000-0002-9071-9750
filomena.maggino@uniroma1.it

Angelos Markos
Democritus University of Thrace, Greece
https://orcid.org/0000-0002-4204-3573
amarkos@eled.duth.gr

Rosaria Romano
Università di Napoli Federico II, Naples, Italy
https://orcid.org/0000-0002-9708-1753
rosaroma@unina.it

Rosaria Simone
Università di Napoli Federico II, Naples, Italy
https://orcid.org/0000-0002-6844-6418
rosaria.simone@unina.it

Laura Trinchera
Neoma Business School, Paris, France
https://orcid.org/0000-0001-9679-0956
laura.trinchera@neoma-bs.fr

Maria Cristina Martini
Università di Modena e Reggio Emilia, Modena, Italy
https://orcid.org/0000-0001-5622-9187
mariacristiana.martini@unimore.it

Fulvia Mecatti
Università di Milano Bicocca, Milan, Italy
https://orcid.org/0000-0002-0403-9231
fulvia.mecatti@unimib.it.

Isabella Morlini
Università di Modena e Reggio Emilia, Modena, Italy
https://orcid.org/0000-0002-2900-2742
isabella.morlini@unimore.it

Biagio Palumbo
Università di Napoli Federico II, Naples, Italy
https://orcid.org/0000-0003-1036-8127
bpalumbo@unina.it

Antonio Punzo
Università di Catania, Catania, Italy
https://orcid.org/0000-0001-7742-1821
antonio.punzo@unict.it

Silvia Salini
Università di Milano, Milan, Italy
https://orcid.org/0000-0001-6106-9835
silvia.salini@unimi.it

A.S.A CONTACTS
Principal Contact
Francesco Palumbo (Editor in Chief)
editor@sa-ijas.org

Support Contact
Domenico Vistocco (Editorial Manager)
ijas@sa-ijas.org

JOURNAL WEBPAGE
https://oajournals.fupress.net/index.php/ijas

Statistica Applicata – Italian Journal of Applied Statistics is associated to the following Italian and international journals:

QTQM – Quality Technology & Quantitative Management (http/web.it.nctu.ew/~qtqm/)

SINERGIE – Italian Journal of Management

Pubblished: April 2025

Published by Firenze University Press
Powered by Firenze University Press

Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

Statistica Applicata – Italian Journal of Applied Statistics is a four-monthly journal published by the Associazione per la Statistica Applicata (A.S.A.), Largo Gemelli 1 – 20123 Milano, Italy (phone + 39 02 72342904). Advertising:Cleup sc, via G. Belzoni, 118/3 – 35128 Padova, Italy (phone +39 049 8753496 – Fax +39 049 9865390), email: info@cleup.it.
Rules for manuscript submission: https://www.sa-ijas.org/ojs/index.php/sa-ijas/about/submissions
Subscription: yearly €103.30; single copy €40.00; A.S.A. associates €60.00; supporting institutions: €350.00. Advertisement lower than 70%. Postal subscription Group IV, Milan. Forum licence n. 782/89. CLEUP SC on behalf of ASA, 7 March 2023.

Summary

Statistica Applicata
ITALIAN JOURNAL OF APPLIED STATISTICS

Vol. 37, Number 1

doi.org/10.26398/IJAS.210

THE TRANSITION FROM UPPER SECONDARY TO HIGHER EDUCATION: SURVEY INSIGHTS FROM ITALY

Michele Lalla

Marco Biagi Department of Economics, University of Modena and Reggio Emilia, Modena, Italy. ORCID: 0000-0002-1639-7300

Patrizio Frederic[1]

Marco Biagi Department of Economics, University of Modena and Reggio Emilia, Modena, Italy. ORCID: 0000-0001-9073-2878

Abstract. *Two data sets for 2009 were used to compare Italians and immigrants: the European Union Statistics on Income and Living Conditions (EU-SILC) and the Italian Survey on Income and Living Conditions of Families with Immigrants (IM-SILC). A sub-sample of subjects between 20 and 25 years of age was set up, containing individual, family, and contextual variables. Their effects on the choice of tertiary education (yes/no) were assessed using a Lasso method to determine the significant explanatory set of variables through a Bayesian approach also aimed at identifying interaction terms. The transition from high school to higher education showed a complex pattern involving many variables: young women continued with their education more than young men; the educational level of the parents and many components of income entered the model in a parabolic form. Significant contextual factors included the degree of urbanisation and household tenure status. New elements of this study include the sample, the Lasso method in this field, and some empirical results.*

Keywords: *Transition to university, Educational inequality, Educational territorial patterns, Lasso method, Bayesian logistic model.*

1. INTRODUCTION

Tertiary education is not compulsory in almost all educational systems and enrolment decisions constitute a difficult step for students because they are making decisions about their future without knowing much about themselves and/or the likely evolution and needs of society. These decisions may be affected

[1] Corresponding author: Patrizio Frederic, email: patrizio.frederic@unimore.it

by differences in individual characteristics and/or the socio-economic conditions of families, as well as social and contextual conditions in the area where they reside. Additionally, such decisions are likely to impact opportunities for future employment and upward mobility, while individual difficulties and critical family situations may also lead to dramatically lower grades and dropping out of school (Grove et al., 2006; Wintre et al., 2011; Armstrong and Biktimirov, 2013). All these aspects may differ among young immigrants and non-immigrants and, in the case of the former, tertiary education plays an important role not only in terms of investment in human capital, the cultural formation process, and social integration, but also as an instrument for social mobility and transformation, individual development through attuned interactions and collective healing through cooperation (Entwisle and Alexander, 1993; Ichou, 2014; Paba and Bertozzi, 2017; De Clercq et al., 2017).

The first objective of this paper is to point out the differences with respect to citizenship, a binary variable distinguishing between immigrants and non-immigrants (hereinafter also referred to as Italians), and the decision to continue with tertiary education or to discontinue their studies after finishing their upper secondary schooling, using a sufficiently large sample of immigrants in comparison to non-immigrants.

The second objective is to identify the determinants of this transition by using the Lasso method through the Bayesian approach, selecting automatically the interactions between explanatory variables, while also accounting for the marginal effects of individual characteristics, family, and social background. The data were extracted from two surveys (the reference year being 2009) carried out by the Italian National Institute of Statistics (Istat): one is the European Union Statistics on Income and Living Conditions (EU-SILC) restricted to Italy only (IT-SILC) – an annual survey conducted since 2004 coordinated by Eurostat (Istat, 2008; Eurostat, 2009) – and the other being the Italian Survey on Income and Living Conditions of families with Immigrants (IM-SILC),[2] which is a single

[2] Note that the letter S in the acronym EU-SILC is often assumed to mean “Survey”, rather than “Statistics”. The same has been done here to provide correspondence with the acronym for the Italian Survey on Income and Living Conditions of families with immigrants, where the term “immigrants” refers to individuals without Italian citizenship. The term adopted here for this group is “Immigrant Survey on Income and Living Conditions” (IM-SILC) to obtain a similar structure of the acronyms for the two surveys.

cross-sectional survey (Istat, 2009a) that involved families with at least one immigrant component resident in Italy.

Multinomial choices may be applied to the transition from upper secondary education to employment or to attend post-secondary non-tertiary or tertiary education (Nguyen and Taylor, 2003) involving several alternatives (employed, unemployed, inactive or out of the labour force, and/or distinguishing between various university degree levels). However, this paper focuses mainly on the decision to attend a degree programme rather than the other options. The binary nature of the dependent variable, hereinafter referred to as the "tertiary" (dependent) variable, implies that it is equal to 1 when an individual is attending a tertiary education level, and equal to zero otherwise, i.e., when the student has achieved an upper secondary level of education. It directly involves some specific techniques, such as ordinary logistic regression in the classical approach or a Bayesian approach, both of which were applied here. In the latter case, the set of independent variables was identified with the Lasso method, which simultaneously allows for the selection of the explanatory variables, the interaction terms, and the estimation of the model coefficients.

The paper is organised as follows. Section 2 provides an overview of the theoretical background, and Section 3 illustrates the sample, the data and some descriptive results concerning the main variables used in the subsequent analyses. Section 4 briefly explains the ordinary logistic model and the Bayesian model combined with the Lasso techniques. Section 5 describes the model obtained through the peculiar Lasso techniques for selection of the independent variables and a Bayesian approach for the estimation of parameters. Finally, Section 6 concludes with some comments and remarks.

2. BACKGROUND

Educational decisions that young people face are made at a particular stage in their lives when influences inside and outside the home are strongly felt. In this sense, such decisions strongly depend on both individual and family characteristics, as well as the environment, but also on psychological and school-related factors (Parker et al., 2004; Wilson and Gillies, 2005).

Several researchers have investigated the factors that influence the choices of young people, mainly from a socio-economic point of view, allowing for status inequalities, i.e., how individual, parental, and family characteristics affect and interact with human capital accumulation (among many others, Brunello and Checchi, 2007; Dustmann, 2008; Van de Werfhorst and Mijs, 2010). The modelling of the decision to attend a tertiary education programme is often based on human capital theory, dating back to Becker (1964), as school students are faced with two alternatives: to invest in education or to enter the labour market (Nguyen and Taylor, 2003). With respect to the explanatory variables, three sets are generally considered in these studies: personal, parental, and family/environmental characteristics.

First, individuals possessing greater ability than others may benefit from investment in further education, implying that educational achievement is an indicator of this ability. Past analyses have shown that other personal characteristics such as gender and ethnicity are significant factors (Perreira et al., 2006; Bubritzki et al., 2018). Here ethnicity was discarded because the focus was on the immigrant/Italian dichotomy. Health conditions, rarely taken into consideration proved to be associated with the choice of continuing in education and training (Lalla and Pirani, 2014; Ichou and Wallace, 2019).

Second, educational decisions reflect and originate from the context of the family, as human capital theory suggests. The effect of family background on assimilation and expectations has been thoroughly analysed and different factors have been identified as relevant in these processes: household size and family composition, educational level of the parents, socioeconomic status, language and expectations of parents, parental support and involvement, cultural background, and income (among many others, for Italy see Luciano et al., 2009; Buonomo et al., 2018). Extensive comparisons of groups of individuals at various stages of their careers have been carried out and many explanations have been given for employment and income inequalities (Glick and Hohmann-Marriott, 2007; Algan et al., 2010; Luthra and Flashman, 2017; Zwsyen and Longhi, 2018). In the absence of (reliable) income data, studies have taken the employment status of parents as proxies, while in the present study various reliable income variables and several occupational variables were included in the models.

Lastly, the social context of the community and the area of residence has also been found to be relevant (Bond Huie and Frisbie, 2000; Perreira et al., 2006; Sleutjes et al., 2018). Schooling has been analysed as a source of inequality between immigrants and natives and/or among different groups of immigrants as well. The social context includes attending kindergarten, previous experiences of success and failure, advice of teachers and peers, and the availability of schools in the area (Bertolini and Lalla, 2012; Contini, 2013). The school environment can provide strong stimuli for integration in the community as a source of potential comparison with others, and induce motivation for all to improve their knowledge and education. The context of the community of residence may refer to social characteristics of the neighbourhood (Woodraw-Lafield, 2001; Pong and Hao, 2007) and its economic characteristics. Social characteristics have often been represented considering crime level, characteristics of peers, companionship and so on, while the economic factors may refer to the employment/ unemployment rate in the area of residence, the local gross domestic product, and the value added by sector (Bertolini et al., 2015; Zwysen and Longhi, 2018). The local area may provide an important indicator summarising many effects such as segregation and favourable or unfavourable economic conditions, thus affecting decisions on whether to continue in education. Here, macro-regions and the degree of urbanisation were considered as useful indicators of immigrant concentrations, sometimes as a result of settlement preferences, as some regions and towns attract more immigrants than others. Moreover, some indicators of housing conditions, personal and family possessions were introduced into the models.

In conclusion, participation in education is a highly complex phenomenon, offering countless avenues for investigation and analysis. Consequently, it has been widely studied across the globe, particularly during and after the COVID-19 pandemic (2020-2023). In Italy too, an extensive literature has also emerged in recent years. The transition from upper secondary school to tertiary education has been investigated at the national level using time series (Minerva et al., 2022) or through macro-socioeconomic indicators at the provincial level (Bertolini et al., 2015; Paba and Bertozzi, 2017), as well as via local/ regional surveys (Vettori et al., 2020; Rondinelli et al., 2024). Other studies have focused on students' choices regarding geographic mobility (Usala et al., 2023; Vittorietti et al., 2023), which can be seen as a form of internal migration of students. Given the vast number of articles on the topic, for the sake of brevity, it is worth noting at least that several studies have utilised Program for International Student Assessment (PISA) scores as a dependent variable, analysing student performance by comparing immigrants to Italians. After controlling for the relevant variables in the PISA data set, these studies found a negative performance gap for immigrant students compared to Italians and Europeans, largely due to the immigrants' limited access to economic resources and educational materials (Murat, 2012; Murat and Frederic, 2015; Schnell and Azzolini, 2015).

3. DATA SOURCES AND PRELIMINARY EVIDENCE

The data were extracted from two surveys carried out by the Italian National Institute of Statistics (Istat) with 2009 as the reference year.

The first one was the EU-SILC, an annual survey aimed at gathering information based on nationally representative random samples of private households in each European country concerning individual socio-demographic characteristics, micro-level data on income, poverty, social exclusion and living conditions using a unique sampling design and identical definitions of the concepts currently used for these purposes (Eurostat, 2009). As a result, the target population refers to all private households and all persons aged 16 and over. The nation considered was Italy, IT-SILC, and the selected reference year, 2009, was a necessary choice because the IM-SILC (see below) was carried out by Istat only in that year. The IT-SILC target information is distributed over four different groups or data sets, each one grouping different variables: (D) Household Register, (H) Household Data, (R) Personal Register, and (P) Personal Data. The four files were matched to obtain a complete file with information at different levels. In the resulting matched file, the total number of cases was equal to the number in the personal register file (R): 51,196. However, the number of useful and manageable records remained the same as the number of personal records, each one corresponding to an interviewed individual, for a total of 43,636. Obviously, the 0–14-year age class was empty because no individuals under the age of 16 were interviewed.

The IM-SILC was funded by the Ministry of Labour and Social Policies and conducted by Istat in 2009 only, i.e., it was a one-shot survey design using a national probability sample of the population, greater than or equal to 16 years old, residing in private households in Italy. The IM-SILC design was similar to the IT-SILC project. The specificity of the reference population in the IM-SILC, compared to the IT-SILC, involved numerous expedients to improve the representativeness of the sample. (1) The sample design at the origin of the IM-SILC was based on the extraction of municipalities as primary sampling units, subdivided on the basis of the degree of urbanisation [densely-, moderately-, and thinly-populated areas (see Eurostat, 2009)] and taking into account the distribution of the main groups of foreign nationals in Italy, reducing the risk of excluding some groups of foreign nationals who could be particularly concentrated in some areas. (2) Non-respondent families were replaced with other families of the same nationality, minimizing self-selection of the most collaborative nationalities and consequent bias. (3) The questionnaires were translated into the ten most common languages among foreigners residing in Italy, to support the interviewers and facilitate the interviewees' understanding of the questions. (4) The sample was post-stratified at a geographical distribution level, taking into account, in addition to the usual constraints on the known total population, the number of families with immigrants and the foreign population classified into the 13 main nationalities residing in Italy, for better calibration with respect to the reference population (Istat, 2009b). In the resulting matched file, the total number of cases was equal to the number of cases in the Personal Register file (R): 15,036. However, the number of useful and manageable records remained the same as the number of personal records, which was 11,611, each one corresponding to an interviewed individual. Again, the 0–14-year age class was empty.

The target sample was obtained by first selecting individuals in the age range of 20 to 25, obtaining a sample of 3,166 cases. Then, in this sample, the eligible cases were only those individuals whose highest attained International Standard Classification of Education (ISCED) level (UNESCO, 2012) was equal to 3 (upper secondary education) or 4 (post-secondary non-tertiary education). The final target sample consisted of 2,874 individuals (Table 1). Further details about these two data sets can be found in Eurostat (2009), as IM-SILC is largely similar to IT-SILC. The variables introduced in the models are described in the Appendix (§7.1) and Lalla and Frederic (2020).

Table 1 shows that the sample of young immigrants aged 20 to 25 (inclusive) represents approximately 4.5% of the total. This low percentage, i.e. the small sample size, results in certain limitations. For instance, in the model estimation, many categorical variables are unable to discriminate properly between immigrants and Italians because these variables do not present observations for certain modalities among immigrants. Additionally, relationships that are statistically significant in the real world may not be significant in this sample due to the small number of immigrants in the survey. For brevity, this is sufficient to justify the use of these two data sets, even if they are not updated, because they increase the number of immigrants to 22.4%. Moreover, the data set adopted offers unique advantages not found in local surveys, such as a national perspective, detailed information on the health conditions of individuals and their parents, living conditions, and individual and family incomes collected with precision and accuracy, while distinguishing between their sources.

Table 1: Absolute frequencies and row percentages of the sample by the type of survey (TOS) and age

TOS\ Age	20	21	22	23	24	25	Total
IT-SILC Italians	388	416	382	360	373	311	2230
%	17.4	18.7	17.1	16.1	16.7	14.0	100
IT-SILC Foreigners	18	28	14	19	13	13	105
%	17.1	26.7	13.3	18.1	12.4	12.4	100
IM-SILC	62	86	93	99	88	111	539
%	11.5	16.0	17.3	18.4	16.3	20.6	100
Total	468	530	489	478	474	435	**2874**
%	16.3	18.4	17.0	16.6	16.5	15.1	100

3.1 BIVARIATE AND TRIVARIATE ANALYSIS OF THE MAIN VARIABLES

The relationship between the tertiary (binary) dependent variable and the ISCED Level Currently Attended (ILCA) showed that 55.3% of individuals, with an ISCED level equal to 3 or 4, were not enrolled in tertiary education courses (termed “not-attending”), while 44.7% were currently attending tertiary education (Table 2).

The ILCA was examined with respect to several qualitative variables and revealed many significant relationships. With respect to gender, CS(2)= 30.15 (p<0.000), where CS(g) stands for “Chi-Square with g degrees of freedom”: women tended to be attending more than men (49.5% versus 39.4%), with the exception of the post-secondary non-tertiary category, in which the percentage of men (1.5%) was unexpectedly equal to that of women (1.5%). The percentage of women not in education was lower than that of men: 49.0% versus 59.2%. The ILCA showed a significant relationship with respect to citizenship, CS(2)= 115.33 (p<0.000): fewer immigrants attended tertiary education than Italian citizens (26.6% versus 50.0%), while the percentage of immigrants not in education was higher than that of Italians (72.4% versus 48.4%), supporting well-known empirical evidence of difficulties relating to the integration process for immigrants who are also conditioned by scarce economic resources to be allocated to education.

Table 2: Absolute frequencies and row percentages of the tertiary (binary) education (EDU) dependent variable by the ISCED level currently attended (ILCA)

Tertiary\ ILCA	Not-attending	Post-Secondary EDU	Tertiary EDU	Total
Tertiary = 1			1285	1285
%			100.0	100
Tertiary = 0	1546	43		1589
%	97.3	2.7		100
Total	1546	43	1285	**2874**
%	53.8	1.5	44.7	100

A significant relationship emerged between the ILCA and self-perceived health, CS(2)= 10.87 (p<0.004), implying that individuals perceiving fair or bad or very poor health tended to discontinue their studies (66.0%) with respect to those perceiving good or very good health (53.1%), see Ichou and Wallace (2019). The ILCA was related: (*i*) to the degree of urbanisation, CS(4)= 26.26 (p<0.000) – as the density of the area increased, the ILCA increased, and (*ii*) to the Italian macro-regions, CS(8) = 24.27 (p<0.002) – as industrialisation and the possibility of finding employment increased, the percentage of individuals continuing their education decreased. This could be a possible effect. Industrialisation has a positive and indirect impact on primary education, but contributes less to increasing participation rates at higher levels of education and/or to the development of human capital through schooling (Montalbo, 2020). These effects may vary across countries due to cultural, political, and social factors (among many others, Le Brun et al., 2011; Federman and Levine, 2005) or over time (Minerva et al., 2022). In Italy, a weak regressive effect was observed in the transition to higher education in areas with abundant employment opportunities, where individuals may choose to forgo further education. The costs of tertiary education, coupled with the relatively low expected returns from a university degree, are likely to prompt some individuals to enter the labour market rather than continue their studies (Paba and Bertozzi, 2017).

The ILCA was related to the maximum ISCED level attained by the parents, CS(12)= 198.80 (p<0.000), but it was also related to the father’s and mother’s level of educational attainment. The ILCA yielded significant relationships also with several variables describing the working conditions of both parents, although the correlation was often weak.

The ILCA was analysed with respect to the main quantitative variables.

The age of fathers, with respect to the ILCA and citizenship, showed that the fathers of immigrants were younger than the fathers of Italians by about 12 years. Similarly, the mothers of immigrants were younger than the mothers of Italians by about 12 years.

Disposable Family Income (DFI) per capita (in thousands of euros) is reported in Table 3 by the ILCA and citizenship. On average, the DFI per capita for immigrants was reported to be significantly lower than that of Italians by about 4,000 euros: about 35.7%.

Table 3: Absolute frequencies (n), means, and standard deviations (SD) of the disposable family income per capita (in thousands of euros) by citizenship and by the ISCED level currently attended (ILCA) by their children (E=Education)

Citizenship\ ILCA		Not-attending	Post-Secondary E	Tertiary E	Total
Italian citizen:	*n*	1080	36	1114	2230
	Means	*11.389*	*11.508*	*12.543*	*11.967*
	SD	6.999	6.432	9.147	8.153
Foreign citizen:	*n*	466	7	171	644
	Means	*7.777*	*5.868*	*7.563*	*7.699*
	SD	5.315	3.489	5.877	5.452
Total:	***n***	1546	43	1285	**2874**
	Means	*10.300*	*10.590*	*11.880*	*11.011*
	SD	6.742	6.376	8.942	7.835

The size of immigrant families proved to be smaller than those of Italians, although non-significantly for the marginal effects of citizenship with F(1;2868)= 1.16 (p=0.282), but it was statistically significant for the ILCA (p<0.001) and for their interaction (p<0.001). Given that the total fertility rate of immigrant women is generally higher than that of Italian women, one might expect that the size of immigrant families would be larger than that of Italians. However, many immigrants come to work in Italy without their families, and this presumably accounts for the decrease in the size of immigrant families.

Citizenship was examined with respect to some other variables, even if it was not a target dependent variable. Its relationship with the maximum ISCED level attained by parents was statistically significant, CS(6)= 217.01 (p<0.000).

The two-sample Kolmogorov-Smirnov test (K-S) of the equality of distribution functions showed that they were statistically different (combined K-S= 0.265, p<0.000). Immigrant parents had attained upper secondary education levels more frequently than Italians (73.1% versus 42.4%) as expected because other empirical findings have revealed this tendency (Bertolini and Lalla, 2012; Bertolini et al., 2015). In fact, this was also the case with vocational qualifications achieved through post-secondary non-tertiary education (1.6% versus 0.6%). On the contrary, this behaviour was not evident for post-tertiary education, as immigrant parents tended to avoid this type of education (0.9% versus 2.9%), seeking employment immediately after a degree because of scarce economic resources compared to non-immigrants (Forster and van de Werfhorst, 2020).

Citizenship was significantly related to the degree of urbanisation, CS(2)= 19.18 (p<0.000), which was confirmed by the two-sample K-S test of the equality of distribution functions (combined K-S= 0.078, p<0.004). Immigrants tended to settle in densely populated areas more than Italians (36.2% versus 35.3%) or in intermediate areas (46.6% versus 39.6%). As expected, the reverse was true for thinly populated areas (17.2% versus 25.1%). Interaction of foreign nationals with other foreign nationals is facilitated in densely populated areas, but at the same time, integration measures for immigrants may be more efficient in highly populated cities than in other areas.

Citizenship showed a significant relationship with the Italian macro-regions, i.e., the geographical subdivision of Italy into five zones (the North-West, North-East, Centre, South, and the Islands): CS(4)= 50.58 (p<0.000). The immigrants tended to establish themselves in the North-East (24.4% versus 20.0% of Italians), in the North-West (19.9% versus 16.6%), in the Centre (25.6% versus 23.4%), where Rome attracts many immigrants, and the Islands (14.8% versus 10.9%), prefiguring a sort of embryonal segregation (Andersson et al., 2018). If the North-South contrast framework is applied, then the data show that immigrants tend to settle more frequently in the North than in the South. However, the Islands exhibit percentage differences similar to those in the North, particularly Sicily, as they are primary points of entry. Immigrants often need time and favourable conditions to continue the journey toward northern Italy and other European countries.

Citizenship yielded a significant relationship with the index summarising the total self-perceived health of parents, CS(3)= 134.99 (p<0.000) and the K-S test (combined K-S= 0.245, p<0.000), implying that when the number of health problems increased, the percentages of Italians decreased, but they were always higher than that of immigrants, although in slightly nonlinear way. For example, the percentage of immigrants with parents without health problems was greater than that of Italians: 84.0% versus 59.5%.

Citizenship proved to be associated with many variables describing working conditions. The relationship between citizenship and the parents' activity status was statistically significant: CS(4)= 105.20 (p<0.000). Immigrants presented lower percentages than those of Italians for the category "both parents employed" and the category "at least one parent is retired": 21.9% and 1.4% versus 34.1% and 9.8%, respectively. Immigrants presented higher percentages than those of Italians for "employment of father only" and for "employment of mother only": 41.6% and 19.3% versus 31.9% and 13.8%, respectively. Citizenship revealed a significant relationship with the maximum position of parents on the job, CS(4)= 134.03 (p<0.000) and K-S= 0.489 (p<0.000), implying that with higher positions (i.e., when one of the parents has a high position), the percentage of Italians increases, although in a slightly nonlinear way. For example, there was a lower percentage of immigrants in managerial positions with respect to Italians: 0.9% versus 4.9%. The difference concerning the position of executive director was 1.1% versus 6.7%. Citizenship yielded a significant association also with the working conditions of parents, CS(5)= 147.14 (p<0.000).

4. MODEL BY BAYESIAN LASSO SELECTION OF REGRESSORS

Let Y be the binary variable denoting for the i-th individual, the dichotomised choice with respect to attending a tertiary level of education (y=1) versus not attending (y=0). Let $\mathbf{x}_i$ be a vector of regressors. Let π_i be the probability that Y=1 given $\mathbf{x}_i$. Let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)$ be the parameters vector of the model. The logit model is

$$\pi_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}) / \left| 1 + \exp(\mathbf{x}'_i \boldsymbol{\beta}) \right| \tag{1}$$

The Lasso method (Tibshirani, 1996) was applied to carry out the estimation and model selection. In fact, it is a procedure involving an additional penalisation term, L_1, summed up to the negative log-likelihood of the model that depends on an additional parameter named λ, $\lambda \geq 0$. More precisely, let $\Phi(\cdot)$ be the objective function of the logit model, hence

$$\Phi(\boldsymbol{\beta},\lambda;\mathbf{x},\mathbf{y})=\sum_{i=1}^{n}\left[y_i\log(\pi_i)+(1-y_i)\log(1-\pi_i)\right]+\lambda\sum_{j=0}^{K}|\beta_j| \qquad (2)$$

where $\mathbf{x}=(\mathbf{x}_1,\dots,\mathbf{x}_n)$, $\mathbf{y}=(y_1,\dots,y_n)$, and $\pi_i=\pi_i(\mathbf{x}_i,\boldsymbol{\beta})$. Finally, $\Phi(\cdot)$ is minimised for different values of parameter λ. It should be noted that when λ=0, then $\Phi(\cdot)$ is the negative log-likelihood of the logit model. On the other hand, larger values of λ yield many β 's exactly equal to zero.

In many penalised methods, $\Phi(\cdot)$ can be interpreted as the negative logarithm of a posterior distribution in a purely Bayesian fashion. Let $p(y_i|\mathbf{x}_i,\boldsymbol{\beta})$ = $\pi_i^{y_i}(1-\pi_i)^{1-y_i}$ be the usual logit model in the usual Bayesian notation, and let $p(\boldsymbol{\beta}|\lambda)\propto\exp(-\lambda\sum_{j=0}^{K}|\beta_j|)$ be the Laplace prior distribution on coefficients $\boldsymbol{\beta}$; then the posterior distribution is

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{x},\mathbf{y},\lambda) &\propto p(\mathbf{y}|\mathbf{x},\boldsymbol{\beta})\ p(\boldsymbol{\beta}|\lambda) \\ &\propto \prod_{i=1}^{n}p(y_i|\mathbf{x}_i,\boldsymbol{\beta})\ p(\boldsymbol{\beta}|\lambda) \\ &= \prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i}\ \exp\left(-\lambda\sum_{j=0}^{K}|\beta_j|\right). \qquad (3)\end{aligned}$$

Note that $\Phi(\boldsymbol{\beta},\lambda;\mathbf{x},\mathbf{y})=-\log\left[p(\boldsymbol{\beta}|\mathbf{x},\mathbf{y},\lambda)\right]$. Hence the Lasso method can be interpreted as a maximum posterior Bayesian estimation method, where the prior distribution on β's is Laplace and λ plays the role of the hyper-parameter. Let $\hat{\boldsymbol{\beta}}_\lambda$ be the minimizing of $\Phi(\cdot)$, then $\hat{\boldsymbol{\beta}}_\lambda$ is the maximum posterior estimation of $\boldsymbol{\beta}$ conditioned to the data and λ:

$$\hat{\boldsymbol{\beta}}_\lambda=\underset{\boldsymbol{\beta}}{\arg\min}\ \Phi(\boldsymbol{\beta},\lambda;\mathbf{x},\mathbf{y})=\underset{\boldsymbol{\beta}}{\arg\max}\ p(\boldsymbol{\beta}|\mathbf{x},\mathbf{y},\lambda)\ . \qquad (4)$$

The choice of parameter λ plays a crucial role in the estimation procedure. Many different studies have focused on this issue; see Zou, Hastie, and Tibshirani (2007) for an extensive review. In addition to the classic AIC and BIC criteria, a *k-fold Cross Validation* (CV) procedure and the *One Standard Error Rule* (1SE) have been proposed. The CV procedure consists of randomly partitioning the original sample into *k* equal-sized subsamples (usually *k*= 5 or *k*= 10). Of the *k* subsamples, a single subsample is retained as validation data for testing the model and the remaining (k–1) subsamples are used as training data. The process is repeated k times, and each of the k subsamples is used exactly once as validation data. The CV for a given λ is the average of binomial deviance in each step. The optimal value of λ is

$$\lambda_{CV} = \underset{\lambda}{\arg\min}\, \mathrm{CV}(\lambda). \tag{5}$$

In order to achieve greater regularisation, the 1SE rule consists in choosing $\lambda_{1\mathrm{SE}} > \lambda_{1\mathrm{CV}}$ such that $\mathrm{CV}(\lambda_{1\mathrm{SE}}) = \mathrm{CV}(\lambda_{\mathrm{CV}}) + \mathrm{SE}[\mathrm{CV}(\lambda_{\mathrm{CV}})]$, where $\mathrm{SE}[\mathrm{CV}(\lambda_{\mathrm{CV}})]$ is the standard error estimated in the k steps.

It is well known (Hastie et al., 2015) that CV estimates prediction error at any fixed value of the tuning parameter, and thus by using it, it is implicitly assumed that achieving the minimal prediction error is the goal, which is not the case here. The 1SE rule is the best candidate for achieving the goal of recovering the true model. Actually, 1SE adds more regularisation than CV. As a result, the 1SE rule was used for selecting the variables.

The model was estimated using the glmnet (Friedman et al., 2010) package in R (R Core Team, 2019). The glmnet package, like many other penalised likelihood packages, provides point estimation for coefficients $\boldsymbol{\beta}$ and statistics for evaluating the CV, but it does not provide confidence intervals for the parameters or standard errors. However, it is possible to draw samples from the posterior distribution $p(\boldsymbol{\beta}|\mathbf{x}, \mathbf{y}, \lambda_{1\mathrm{SE}})$ and then to perform a full Bayesian analysis.

5. OUTCOMES OF THE LOGISTIC MODEL

The interpretation of coefficients is not easy, and the odds ratios (OR) reported in Table 4 and displayed in Figure 1 facilitate data interpretation only at first sight because calculations are necessary to quantify probabilities (Gould, 2000; Kleinbaum, 1994). In fact, OR is equal to odds (if the examined variable is incremented by 1)/ odds (if that variable is not incremented) or, more formally,

$$\mathrm{OR}=\left[P(y=1\,|\,x+1)/(1-P(y=1\,|\,x+1)\right]/\left[P(y=1\,|\,x)/(1-P(y=1\,|\,x)\right]. \tag{6}$$

Moreover, Table 4 only presents interaction terms of the first order because the analysis of interaction orders was limited to the first order to simplify interpretation. The interactions are indicated by the symbol ×, which is read as "by".

Let $\mathbf{x}_b$ be the binary variables. Let $\mathbf{x}_c = \boldsymbol{\mu}$ be the mean values of the continuous regressors, limited to the ages of individuals, which can never be zero in practice. Note that: (1) the product of two binary variables is again a binary variable, (2) the percentage of variation of the reference probability, $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}}$, is given by [100*(OR–1)] and is reported below in parentheses. The probability of having y=1 (i.e., of continuing one's education) was equal to $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}} = \mathbf{0.120}$, calculated at the mean values of the continuous regressors $(\mathbf{x}_c = \boldsymbol{\mu})$ and the binary variables equal to 0 $(\mathbf{x}_b)$. A binary variable having an OR greater than 1 implied that the group represented by the binary variable equal to 1 had a higher probability of having y=1 than the group identified by the binary variable equal to 0; for example, for women with an OR=1.777, the probability of continuing their education was +77.7% greater than that of men. In other words, $\pi_{w|\square}$= 1.777×0.120= 0.213, which was +77.7% greater than the probability of men: 0.120. Note that the dot in the index means keeping all other variables fixed, i.e., the binary and the continuous variables other than age being equal to zero. The successive binary variable having an OR>1 in Table 4 was "PES (Parents' Employment Status) is inactive" (x_1) × "Family living in a densely populated area" (x_2), denoted by x_{12}, which showed an OR=1.697, meaning that the odds of the event y=1, when x_{12}=1 (both x_1 and x_2 are equal to 1), were +69.7% greater than the odds of the event y=1, when x_{12}=0. Similarly, highly significant probabilities of continuing in education were observed for other interaction terms: "Father with permanent contract" × "Only mother employed" (+95.7%), "Father with permanent contract" × "Parents are managers or executives" (+132.1%), "Mother with permanent contract" × "Father is limited by health" (+64.7%), "Father with term contract" × "Mother is limited by health" (+266.5%, which is an unbelievable outcome), "TSH (Tenure Status of Household): Subtenant" × "Family living in a moderately populated area" (+46.6%), and "TSH: Free" × "Assets reduction for needs" (+173.3%).

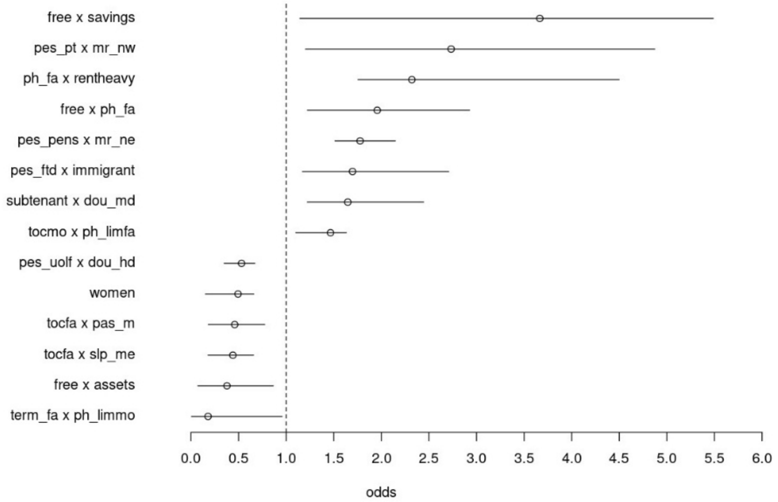

Figure 1: Odds ratios of dichotomous variables in the Bayesian model

In short, gender, favourable and stable parents' working conditions, and good actual and self-perceived health conditions strongly affected the probability of continuing from upper secondary to tertiary education, although this occurred in interaction with other factors.

The binary variables having an OR lower than 1 implied that the group represented had a lower probability of having y=1 with respect to the complementary group. In Table 4 there are six (interaction) binary variables with an OR lower than 1. For example, "Father perceives poor health" × "Rent is burdensome" had an OR=0.440 and hence its complement to one, expressed as a percentage, was equal to [100*(0.440–1)] = –56.0%. As a result, the probability of continuing in education amounted to –56.0% of the probability of the complementary group whose fathers did not perceive poor health and a burdensome rent. In other words, the group with x_{12}=1 had a probability equal to $\pi_{x_{12}=1|\square}$= 0.440×0.120= 0.053. The other five interactions were: "PES= pensioners" × "North-Est" (–50.6%), "PES: part-time" × "North-West" (–62.2%), "PES: full-time employee" × "immigrant" (–46.9%), "TSH= Free" × "Father: poor health" (–54.1%), "TSH= Free" × "Savings" (–82.1%). It is worth noting that the effect of "PES= full-time employee" × "immigrant" (–46.9%) may seem counterintuitive, as full-time parental employment would typically increase the likelihood of continuing in education. However, despite this, immigrants were still less likely than Italians to go on to tertiary education.

Table 4: Logistic regression with Lasso method and Bayesian approach: Estimated odds ratio (OR), standard errors (SE), p-values (p), and means (M)

B=Binary/ C=Continuous regressor	OR	SE	*p*	M
B- Women	1.777	0.263	0.000	0.530
C- [(Individual's age)/10]^2	0.714	0.044	0.000	5.064
C- (Father's age)/10	1.175	0.073	0.003	4.973
C- (Mother's age)/10	1.548	0.094	0.000	4.727
C- (Education Level of Father: years)^2	1.003	0.001	0.000	1.552
C- FDPI= (Father's DPI)/ 10000	1.452	0.070	0.000	2.372
C- MDPI= (Mother's DPI)/ 10000	1.285	0.062	0.000	1.248
C- FTIPC= (Family's total income per capita)/ 10000	0.314	0.046	0.000	1.101
Interactions of first order				
B- (Father: poor health) × (Burdensome rent)	0.440	0.156	0.011	0.023
B- (PES= Parents' Employment Status: pensioners) × (North-Est)	0.494	0.188	0.031	0.017
B- (PES: inactive) × (Densely populated area)	1.697	0.428	0.048	0.043
B- (PES: part-time) × (North-West)	0.378	0.237	0.042	0.008
B- (PES: full-time employee) × immigrant	0.531	0.092	0.000	0.121
B- (Father: PC= permanent contract) × (Only mother employed)	1.957	0.544	0.013	0.038
B- (Father: PC) × (Parents: manager/ executive)	2.321	0.737	0.013	0.048
B- (Mother: PC) × (Father: limited by health)	1.647	0.322	0.010	0.065
B- (Father: Term C.) × (Mother: limited by health)	3.665	1.776	0.011	0.010
B- (TSH$^+$: Subtenant) × (Moderately populated area)	1.466	0.175	0.001	0.246
B- (TSH$^+$: Free) × (Father: poor health)	0.459	0.186	0.025	0.016
B- (TSH$^+$: Free) × (Assets reduction for needs)	2.733	1.041	0.010	0.016
B- (TSH$^+$ [Tenure Status of House.]: Free) × Savings	0.179	0.220	0.023	0.003
Intercept	0.043	0.029	0.000	
Bayesian Pseudo-R square	0.227	*n* =	2874	

In short, unstable and unfavourable working conditions of parents, poor actual and self-perceived health conditions of parents, and critical and costly tenure status of the household negatively affected the probability of making the transition from upper secondary school to tertiary education, although this emerged through the interaction terms.

The continuous variables. The individual's age (range 20-25), expressed in decades, showed a parabolic and negative impact on education paths, while the ages of both parents revealed a linear positive impact on the probability of making the transition to higher education. The other continuous single variables (which may be conceptually and concretely equal to 0) entering the model showed significant effects on going on to higher education. With the increase in the parents' educational level, the probability of continuing in education increased quadratically. The father's and mother's disposable personal income (FDPI and MDPI) indicated a linear positive effect (Ochsen, 2011; Krause et al., 2015), whereas the family's total income per capita (FTIPC) yielded an unexpected negative effect, but perhaps the father's income balanced out the effect of the mother's income. In fact, FTIPC included both FDPI and MDPI. However, the algebraic sum of their impacts remained positive, implying the importance of welfare programmes to help families experiencing economic (and physical) difficulties, with the specific aim of reducing the number of students not continuing with their education. The trends of $\pi_i = P(Y=1)$ for FDPI and FTIPC are illustrated in Figure 2.

The main fault of the Lasso method in selecting significant explanatory variables concerns the possibility of selecting a theoretically unjustifiable variable, such as "Father with term contract" × "Mother is limited by health" (+266.5%) or of neglecting some important variables in the model.

The same model was estimated with the ordinary logistic procedure and the obtained odds ratios were approximately equal to those reported in Table 4, except for "Father with term contract" × "Mother is limited by health" involving an amount equal to +142.4%. Moreover, only the regressor "TSH= Free" × "Savings" was not significantly different from zero (p=0.125). The Hosmer and Lemeshow goodness-of-fit test for this classical estimation of the Bayesian model, not reported here, indicated a well-fitting model, with a p-value of 0.223, suggesting that the model's estimates fit the data at an acceptable level.

Another model was obtained using classical logistic regression with marginal effects, starting from the complete set of 65 regressors and applying backward selection. The aim was to verify the differences between the Bayesian approach, the focus of this study, with the classical method. The resulting estimates are presented in Table 5. The model also proved to be sufficiently robust to changes in the reference/base category of qualitative variables. The number of final regressors increased: 24 in the classical model versus 21 in the

Lasso Bayesian one. The components of income played a complex effect revealing eight regressors in the classical model versus three regressors in the Lasso Bayesian one.

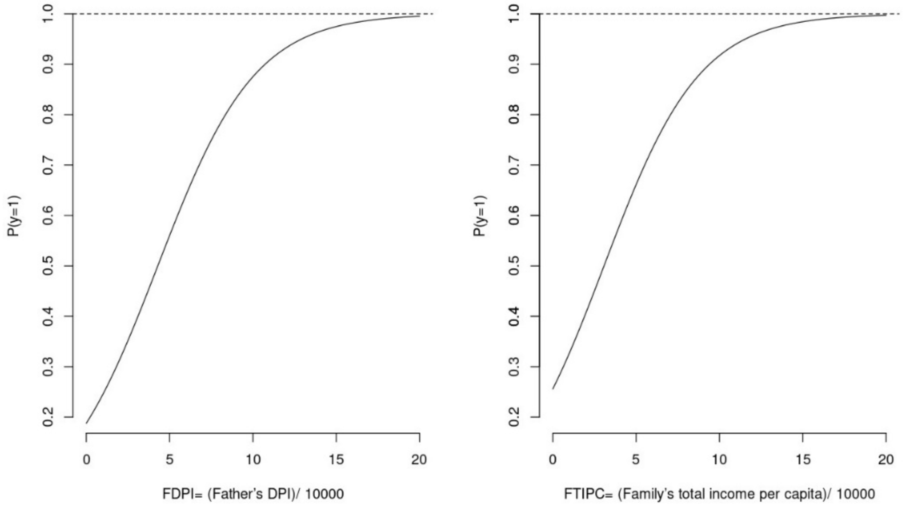

Figure 2: The probability of attending tertiary education in function of the father's disposable personal income (FDPI) and the family's total income per capita (FTIPC)

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are reported in Table 5, making it possible to compare them with those in Table 7 in the Appendix. AIC and BIC do not make much sense in comparing Bayesian and classical models, so they are not reported in Table 4. Moreover, the Hosmer and Lemeshow goodness-of-fit test indicated an unsatisfactory fitting model as its p-value was slightly lower than 0.05, implying that the model's estimates did not fit the data at an acceptable level. Since this model had only an illustrative function, it was not explored in depth. However, more information on how the regressors act on the dependent variable is provided in the Appendix (§7.2).

Table 5: Logistic regression with marginal effects: Estimated odds ratio (OR), standard errors (SE), p-values (p), and means (M)

Regressor	OR	SE	*p*	M
Woman	1.661	0.156	0.000	0.530
DPI= (Disposable Personal Income)/ 10000	0.204	0.023	0.000	0.521
DPI2 = DPI^2	1.226	0.034	0.000	0.851
SPH: Self-Perceived Health	0.408	0.093	0.000	0.055
SPH: limitation in activities	1.934	0.440	0.004	0.048
Macro-region: South	1.299	0.136	0.013	0.260
Degree of urbanisation: high density	1.286	0.156	0.038	0.355
Degree of urbanisation: average density	1.337	0.157	0.014	0.412
[(Father's age)/10]^2	1.015	0.007	0.020	26.04
(Mother's age)/10	2.785	0.656	0.000	4.727
[(Mother's age)/10]^2	0.934	0.023	0.005	23.55
ELF2 = (Education Level of Father: years)^2	1.002	0.001	0.002	133.6
ELM = Education Level of Mother: years	1.078	0.018	0.000	11.01
FDPI= (Father's DPI)/ 10,000	1.198	0.061	0.000	2.372
MDPI= (Mother's DPI)/ 10,000	1.325	0.080	0.000	1.248
MDPI2 = MDPI^2	0.985	0.005	0.003	4.205
FTI= [(Family's Total Income)/ 10,000]	0.844	0.044	0.001	3.969
FTI2= FTI^2	1.008	0.003	0.010	24.70
FTIPC2= [(Family's total income per capita)/10,000]^2	0.935	0.023	0.006	1.826
SLP (Skill Level of Parents): manager or executive	1.574	0.289	0.014	0.095
SLP: employee parent	1.577	0.252	0.004	0.104
PES: parents' unemployed or inactive	1.809	0.273	0.000	0.106
Number of optional facilities in home	1.219	0.039	0.000	4.383
Repayments of loans to banks	0.752	0.080	0.007	0.254
Intercept	0.002	0.001	0.000	1.000
Pseudo-R square	0.251	*n* =	2874	

Note: Log-Lik= -1480.1, Akaike inf criterion= 3010.1, Bayesian inf criterion= 3159.2

For the sake of brevity, no further observations on the differences will be made here, but it should be noted that the traditional fit measures are less applicable in Bayesian models. The Bayesian pseudo-R^2 is reported in Table 4 because it provides a more comprehensive assessment of model fit. Moreover, the classification table presented below indicates how well the model predicts the actual outcomes. Metrics such as false positive and false negative rates are particularly informative for understanding how the model performs in predicting whether an individual will continue or decide not to continue their education.

The performance in the correct classification seemed better in the classical model than in the Lasso Bayesian one: the ordinary post-estimation statistics are reported in Table 6, where it is possible to consider the variations.

Finally, the classification of error rates was computed for $\hat{\boldsymbol{\beta}}_{\lambda_{1\mathrm{SE}}}$ by assigning $\hat{y}_i = 0$ if $\hat{\pi}_i = \exp(x'_i\,\boldsymbol{\beta}_{\lambda_{1\mathrm{SE}}}) \big/ \left[1+\exp(x'_i\,\boldsymbol{\beta}_{\lambda_{1\mathrm{SE}}})\right] < 0.5$, and $\hat{y}_i = 1$, if $\hat{\pi}_i \geq 0.5$. The misclassified number of $\hat{y}_i$ equal to zero was 475 out of 2874, which is a false negative (minus) error rate equal to 37.0% and the misclassified number of $\hat{y}_i$ equal to one was 370 out of 2874, which is a false positive (plus) error rate equal to 23.3% (Table 6). The performance of the logistic model seemed to be slightly better than the Lasso model: the false negative rate was 28.4%, while the false positive rate was 22.8%. In the logistic model, the overall misclassification error rate was equal to 25.3% versus 29.4% in the Lasso model.

Table 6: Performance classification of the logistic and Lasso models (T=Tertiary)

	Lasso	**Model**		**Logistic**	**model**	
Classified\ T	**T=1**	**T=0**	**Total**	**T=1**	**T=0**	**Total**
Positive +	810	370	1180	920	362	1282
Negative –	475	1219	1694	365	1227	1592
Total	**1285**	**1589**	**2874**	**1285**	**1589**	**2874**
False ± rate	37.0	23.3		28.4	22.8	
for true T=0/1	P(– | T=1)	P(+ | T=0)		P(– | T=1)	P(+ | T=0)	
False ± rate	28.0	31.4		22.9	28.2	
for classified ±	P(– | $\hat{y}$ = –)	P(+ | $\hat{y}$ =+)		P(– | $\hat{y}$ = –)	P(+ | $\hat{y}$ =+)	

6. CONCLUSIONS

The key empirical evidence may be summarised as follows. In general, more women tended to continue in education than men. More women than men attended tertiary education (49.5% versus 39.4%), whereas the percentage of women not in school was lower than that of men: 49.0% versus 59.2%. Fewer young immigrants enrolled on education programmes than young Italians: 26.6% versus 50.0%. As a result, the percentage of immigrants not enrolled in education was higher than that of Italians (72.4% versus 48.4%).

In the model, self-perception of health was associated with enrolment in education: individuals with fathers perceiving poor health and a burdensome rent were 56.0% more likely not to continue their education. Immigrant status (i.e., citizenship) was not preserved as marginal effects in the explanatory variables set determined by the automatic model selection procedure of the Lasso method combined with Bayesian logistic model, but it demonstrated a combined effect in the interaction term “PES= full-time employee” × “immigrant” (–46.9%), confirming a disadvantage for immigrants compared to Italians in university enrolment. Similar findings have been observed in other countries, such as the United States (Barsha et al., 2024), France (Ichou and Wallace, 2019), Spain (Pantzer et al., 2006), among others.

The age of the parents of immigrants was significantly lower than that of the parents of Italians, showing on average a difference equal to about 12 years. The parents’ level of education had a significant impact on the probability of young people continuing in education. Analogous findings have been reported in Italy (Cantalini et al., 2020) and other countries (Wilder, 2013; Kantova, 2024). The employment status of immigrant parents was significantly lower than that of Italian parents. The same was true for disposable personal incomes and for the total income of families, some of which were well represented by a parabolic form in the model.

The empirical results are coherent with those reported in the literature and suggest that an “immigration” gradient is present in educational decisions also in Italy. Differences in educational enrolment/ attainment at the tertiary level among immigrants and Italians were explained by the socio-economic status of parents, i.e., their level of education, employment status, and occupational position. These results highlight the need for integrated policies in educational programmes, directed both at sustaining young people and helping their families, in order to stimulate and promote the enrolment of young immigrants in education programmes and to foster a complete integration process.

The outcomes obtained, when compared with those of the bibliographical references cited above, confirmed the findings of previous research while also presenting some novel insights. Two key results are highlighted here. First, the automatic selection of interactions provided interesting and interpretable outcomes, even if this strategy can lead to the selection of highly significant interaction terms, albeit difficult to interpret, and to the elimination of important variables that come into play indirectly through these interactions. This situation may be viewed in terms of the degree of urbanisation and the North-East and North-West microregions in the model in Table 4. Second, the positive impact of individual and family/ parental incomes (recorded to the highest level of precision) on university enrolment was confirmed through a national survey and with a large sample of immigrants. In contrast, previous studies had largely identified this effect through municipal or district-level surveys, or indirectly, using provincial macroeconomic data.

The affordances provided by the two cross-sectional surveys, IT-SILC and IM-SILC (reference year 2009), such as a more consistent sample size of immigrants and a national perspective, also represent limitations of this study. First, the use of IT-SILC data from 2009, which is now admittedly dated, was necessary because that year marked the last survey of its kind on immigrants across Italy. Second, individual educational performance data are absent in surveys like IT-SILC and IM-SILC.

Finally, few models with interactions exist in the literature. In fact, in the applications, the interactions should be supported by social, behavioural, psychological, and economic theories. Otherwise, they may be obtained automatically simply by using an adaptive procedure like the Lasso method and only as empirical findings. The interactions are likely to be easily found among binary or categorical variables, but this case is relatively interesting because they can be replaced with specific typologies. The same holds true for the interactions of a continuous variable with other explanatory binary variables, but the interaction between two continuous variables is difficult to grasp immediately. In general, it is useful to find a theoretical justification for the existence of the interactions, instead of blindly searching for interaction terms. However, it is highly plausible that almost all phenomena are outcomes of interactions among many variables, but the explanation of these results is likely to be complicated and challenging.

ACKNOWLEDGEMENTS

This research was carried out as part of the “Determinant and economic effects of international migration” project of the Marco Biagi Department of Economics and the Marco Biagi Foundation, with financial support from the University of Modena and Reggio Emilia, UNIMORE FAR 2017 Grant, which is gratefully acknowledged. The present paper is a major revision of “Factors affecting tertiary education decisions of immigrants and non-immigrants in Italy”, presented at the Scientific Conference on “Data-Driven Decision Making”, ASA (Associazione per la Statistica Applicata) and DISPI (Dipartimento di Scienze Politiche e

Internazionali), University of Genoa (12-14 September 2022). In addition, the authors wish to thank William Bromwich for his painstaking attention to copy-editing this paper. Last but not least, this study is dedicated to the memory of Lorenzo Bernardi for his teachings, honesty, concern, and eagerness in stimulating all who worked with him and for his generous engagement with academic and other public institutions. He has left an indelible mark in our hearts.

REFERENCES

Algan, Y., Dustmann, C., Glitz, A. and Manning, A. (2010). The economic situation of first and second-generation immigrants in France, Germany and the United Kingdom. *The Economic Journal*. 15(2): 353-385.

Andersson, E.K., Lyngstad, T.H. and Sleutjes, B. (2018). Comparing patterns of segregation in North-Western Europe: A multiscalar approach. *European Journal of Population*. 34(2): 151-168.

Armstrong, M.J. and Biktimirov, E.N. (2013). To repeat or not to repeat a course. *Journal of Education for Business*. 88(6): 339–344.

Barsha, R.A.A., Najand, B., Zare, H. and Assari, S. (2024). Immigration, educational attainment, and subjective health in the United States. *Journal of Mental Health & Clinical Psychology*. 8(1): 16-25.

Becker, G.S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Columbia University Press, New York.

Bertolini, P. and Lalla, M. (2012). Immigrant inclusion and prospects through schooling in Italy: An analysis of emerging regional patterns. In C. E. Kubrin, M. S. Zatz, and R. Martínez Jr., editors, *Punishing Immigrants: Policy, Politics and Injustice*. New York University Press, New York: 178-206.

Bertolini, P., Lalla, M. and Pagliacci, F. (2015). School enrollment of first- and second-generation immigrant students in Italy: A geographical analysis. *Papers in Regional Science*. 94(1): 141-160.

Bond Huie, S.A. and Frisbie, W.P. (2000). The component of density and the dimensions of residential segregation. *Population Research and Policy Review*. 19(6): 505-524.

Brunello, G. and Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy*. 22(52): 781-861.

Bubritzki, S., van Tubergen, F., Weesie, J. and Smith, S. (2018). Ethnic composition of the school class and interethnic attitudes: a multi-group perspective. *Journal of Ethnic and Migration Studies*. 44(3): 482-502.

Buonomo, A., Strozza, S. and Gabrielli, G. (2018). Immigrant youths: Between early leaving and continue their studies. In B. Merrill, M. T. Padilla-Carmona, and J. González-Monteagudo, editors, *Higher Education, Employability and Transitions* *to the Labour Market*. EMPLOY Project & University of Seville, Seville (ES): 131-147.

Cantalini, S., Guetto, R. and Panichella, N. (2020). Parental age at childbirth and children's educational outcomes: Evidence from upper-secondary schools in Italy. *Genus*. 76(8): 1-24.

Contini, D. (2013). Immigrant background peer effects in Italian schools. *Social Science Research*. 42(4): 1122-1142.

De Clercq, M., Galand, B. and Frenay, M. (2017). Transition from high school to university: A person-centered approach to academic achievement. *European Journal of Psychology of Education*. 32(1): 39-59.

Dustmann, C. (2008). Return migration, investment in children, and intergenerational mobility. *The Journal of Human Resources*. XLIII(2): 299-324.

Entwisle, D.R. and Alexander, K.L. (1993). Entry into school: The beginning school transition and educational stratification in the United States. *Annual Review of Sociology*. 19: 401-423.

Eurostat (2009). *Description of Target Variables: Cross–Section and Longitudinal*. EU–SILC 065 (2009 operation). Directorate F, Unit F-3. Eurostat, Luxembourg.

Federman, M. and Levine, D. I. (2005). The effects of industrialization on education and youth labor in Indonesia. *Contributions to Macroeconomics*. 5(1), 1: 1-34.

Forster, A.G. and van de Werfhorst, H.G. (2020). Navigating institutions: Parents' knowledge of the educational system and students' success in education. *European Sociological Review*. 36(1): 48-64. https://doi.org/10.1093/esr/jcz049.

Friedman, J.H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 33(1): 1-22.

Glick, J.E. and Hohmann-Marriott, B. (2007). Academic performance of young children in immigrant families: The significance of race, ethnicity and national origins. *International Migration Review*. 41(2): 371-402.

Gould, W.W. (2000). sg124: Interpreting logistic regression in all its forms. *Stata Technical Bulletin*. 53: 19–29. Reprinted in *Stata Technical Bulletin Reprints*. Vol. 9: 257–270. Stata Press, College Station, TX.

Grove, W.A., Wasserman, T. and Grodner, A. (2006). Choosing a proxy for academic aptitude. *Journal of Economic Education*. 37(2): 131-147.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, London.

Ichou, M. (2014). Who they were there: Immigrants' educational selectivity and their children's educational attainment. *European Sociological Review*. 30(6): 750-765. https://doi.org/10.1093/esr/jcu071.

Ichou, M. and Wallace, M. (2019). The Healthy Immigrant Effect: The role of educational selectivity in the good health of immigrants. *Demographic Research*. 40(4): 61-94.

Istat (2008). Ceccarelli C., Di Marco M., and Rinaldelli C., editors, *L'indagine europea sui redditi e le condizioni di vita delle famiglie* (Eu-Silc). Metodi e Norme n. 37. Istat, Rome.

Istat (2009a). Reddito e condizioni di vita delle famiglie con stranieri [electronic resource]. Rome: Istat. https://www.istat.it/it/archivio/52405. Last access: 03/01/2020.

Istat (2009b). Nota informativa sull'utilizzo dell'UDB (User Data Base) CVS 2009. Released by Istat together with data sets. Istat, Rome.

Kantova, K. (2024). Parental involvement and education outcomes of their children. *Applied Economics*. 56(48): 5683-5698.

Kleinbaum, D.G. (1994). *Logistic Regression: A Self-Learning Text*. Springer-Verlag, New York.

Krause, A., Rinne, U. and Schüller, S. (2015). Kick it like Özil? Decomposing the native-migrant education gap. *International Migration Review*. 49(3): 757-789.

Lalla, M. and Pirani, E. (2014). The secondary education choices of immigrants and non-immigrants in Italy. *Rivista Italiana di Economia Demografia e Statistica*. LXVIII(3-4): 39-46.

Lalla, M. and Frederic, P. (2020). Tertiary education decisions of immigrants and non-immigrants in Italy: An empirical approach. *DEMB Working Papers Series*. N. 168: 1-39. University of Modena and Reggio Emilia, Marco Biagi Department of Economics.

Le Brun, A., Helper, S. R. and Levine, D. I. (2011). The effect of industrialization on children's education. The experience of Mexico. *Review of Economics and Institutions*. 2(2): 1-34.

Luciano, A., Demartini, M. and Ricucci, R. (2009). L'istruzione dopo la scuola dell'obbligo. Quali percorsi per gli alunni stranieri? In G. Zincone, editor, *Immigrazione: segnali di integrazione. Sanità, scuola e casa*. il Mulino, Bologna: 113-156.

Luthra, R.R. and Flashman, J. (2017). Who benefits most from a university degree?: A cross-national comparison of selection and wage returns in the US, UK, and Germany. *Research in Higher Education*. 58(8): 843-878.

Minerva, T., De Santis, A., Bellini, C. and Sannicandro, K. (2022). A time series analysis of students enrolled in Italian universities from 2000 to 2021. *Italian Journal of Educational Research*. Anno XV(29): 09-22.

Montalbo, A. (2020). Industrial activities and primary schooling in early nineteenth-century France. *Cliometrica*. 14(2): 325-365.

Murat, M. (2012). Do immigrant students succeed? Evidence from Italy and France. *Global Economic Journal.* 12(3), Article 8: 1-22.

Murat, M. and Frederic P. (2015). Institutions, culture and background: The school performance of immigrant students. *Education Economics*. 23(5): 612-630.

Nguyen, A.N. and Taylor, J. (2003). Post-high school choices: New evidence from a multinomial logit model. *Journal of Population Economics*. 16(2): 287-306.

Ochsen, C. (2011). Recommendation, class repeating, and children's ability: German school tracking experiences. *Applied Economics*. 43(27): 4127-33.

Paba, S. and Bertozzi, R. (2017). What happens to students with a migrant background in the transition to higher education? Evidence from Italy. *Rassegna Italiana di Sociologia*. 58(2): 315-351.

Pantzer, K., Rajmil, L., Tebé, C., Codina, F., Serra-Sutton, V., Ferrer, M., Ravens-Sieberer, U., Simeoni, M.-C., Alonso, J. (2006). Health related quality of life in immigrants and native school aged adolescents in Spain. *Journal of Epidemiology and Community Health*. 60(8): 694-698.

Parker, J.D.A., Summerfeldt, L.J., Hogan, M.J. and Majeski, S.A. (2004). Emotional intelligence and academic success: Examining the transition from high school to university. *Personality and Individual Differences*. 36: 163-172.

Perreira, K.M., Harris, K.M. and Lee, D. (2006). Making it in America: High school completion by immigrant and native youth. *Demography*. 43(3): 511-536.

Pong, S. and Hao, L. (2007). Neighborhood and school factors in the school performance of immigrants' children. *International Migration Review*. 41(1): 206-241.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna (Austria). URL http://www.R-project.org/. Last access: 18/04/2020.

Rondinelli, R., Policastro, V. and Scolorato, C. (2024). How student characteristics affect mobility choices at the university level: Insights from two surveys in Campania region. *Statistica Applicata – Italian Journal of Applied Statistics*. 36(1): 7-39.

Schnell, P. and Azzolini, D. (2015). The academic achievements of immigrant youths in new destination countries: Evidence from southern Europe. *Migration Studies*. 3(2): 217- 240.

Sleutjes, B., de Valk, H.A. G. and Ooijevaar, J. (2018). The measurement of ethnic segregation in the Netherlands: Differences between administrative and individualized neighbourhoods. *European Journal of Population*. 34(2): 195-224.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*. B. 58(1): 267-288.

UNESCO Institute for Statistics (2012). *International Standard Classification of Education ISCED 2011*, Ref. UIS/2012/INS/10/REV, UNESCO–UIS, Montreal (Quebec - Canada).

Usala, C., Porcu, M. and Sulis, I. (2023). The high school effect on students' mobility choices. *Statistical Methods & Applications*. 32(4): 1259-1293.

Van de Werfhorst, H.G. and Mijs, J.J.B. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual Review of Sociology* 36: 407-428.

Vettori, G., Vezzani, C., Pinto, G. and Bigozzi, L. (2020). The predictive role of prior achievements and conceptions of learning in university success: evidence from a retrospective longitudinal study in the Italian context. *Higher Education Research & Development*. 40(7): 1564–1577.

Vittorietti, M., Giambalvo, O., Genova, V. G. and Aiello, F. (2023). A new measure for the attitude to mobility of Italian students and graduates: A topological data analysis approach. *Statistical Methods & Applications*. 32(2): 509-543.

Wilder, S. (2013). Effects of parental involvement on academic achievement: A meta-synthesis. *Educational Review*. 66(3): 377–397.

Wilson, G. and Gillies, R. (2005). Stress associated with the transition from high school to university: The effect of social support and self-Efficacy. *Australian Journal of Guidance and Counselling*. 15(1): 77-92. doi:10.1375/ajgc.15.1.77

Wintre, M.G., Dilouya, B., Pancer, S.M., Pratt, M.W., Birnie–Lefcovitch, S., Polivy, J. and Adams, G. (2011). Academic achievement in first–year university: Who maintains their high school average? *Higher Education*. 62(4): 467-481.

Woodraw-Lafield, K.A. (2001). Implication of immigration for apportionment. *Population Research and Policy Review*. 20(4): 267-289.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the Lasso. *The Annals of Statistics*. 35(5): 2173-2192.

Zwysen, W. and Longhi, S. (2018). Employment and earning differences in the early career of ethnic minority British graduates: The importance of university career, parental background and area characteristics. *Journal of Ethnic and Migration Studies*. 44(1): 154-172.

7. APPENDIX

The initial set of variables used to estimate the models and an example of models obtained by the introduction of explanatory variables blocks step by step are illustrated in this Section.

7.1 LIST OF VARIABLES

The data set contained many variables describing different aspects of each individual, as stated above. A factor analysis might have aggregated them into a reduced set. In general, there are difficulties in understanding and interpreting these factors. As a result, only the original variables indicated below were included in the models, sometimes with modifications and/or adaptations. For further details the reference provided in the introduction of Section 3 may be useful.

Qualitative variables are listed separately from quantitative variables.

Gender was dichotomised as 0 (men) and 1 (women) and termed women.

Citizenship was dichotomised as non-immigrant (0) and immigrant (1) and termed immigrants.

Self-perceived health (SPH), measured on a Likert scale (1=very good, 2=good, 3=fair, 4=bad, 5=very bad), was dichotomised assuming the value of 1 when SPH was problematic (6.4%), i.e., when the answer was fair or bad or very bad, and the value of 0 otherwise.

Suffering from any chronic illness or condition was equal to 1 for "Yes" and equal to 0 otherwise.

Limitation in activities because of health problems was dichotomised assuming the value of 1 when SPH was problematic (5.1% for severely limited or limited) and the value of 0 otherwise.

The "unmet need for medical treatment or examination" (5.0%) and the "unmet need for dental examination or treatment" (8.1%) were not included in the models, in order to reduce the number of explanatory variables, but also because this information is likely to be captured by the income of the family.

This block of variables was repeated for each young individual, their father, and their mother.

Education level of the father (ELF) and mother (ELM) were transformed into years and considered continuous variables. ELF and ELM were introduced into the models through a second-degree polynomial form: see below. Their modalities were the following: (0=ILL) illiterate, (0=NENI) no education and not illiterate, (1=PE) primary education, (2=LSE) lower secondary education, (2=VS3Y) vocational school of 2-3 years, (3=USE) upper secondary education, (4=PS-NTE) post-secondary non-tertiary education, (5=SCTE) short-cycle tertiary education, (6=BACH) bachelor's or equivalent level, (7=MAST) master's or equivalent level after bachelor, (8=PhD) research doctorate, doctor of philosophy or equivalent.

The characteristics concerning the labour market situation of the parents involved several categorical variables, which were combined between father and mother to reduce their numerosity and the results were transformed into binary variables.

The *parents' activity status* (PAS) reported the combination of the father's and mother's conditions: (1) PAS-FM equal to 1 when the father and mother were both employed and 0 otherwise, (2) PAS-F equal to 1 when only the father was employed and 0 otherwise, (3) PAS-M equal to 1 when only the mother was employed and 0 otherwise, (4) PAS-R equal to 1 when at least one of the parents was retired and 0 otherwise, (5) PAS-O equal 1 when both parents were classifiable under "other conditions" and 0 otherwise.

The *skill level of parents* (SLP) in the job was determined retaining the maximum between the positions of the father and the mother: (1) SLP-ME if at least one of the parents was a manager or executive director and the other in a lower position, (2) SLP-EMPL if at least one of the parents was an employee and the other in a lower position, (3) SLP-LAB if at least one of the parents was a labourer and the other was unemployed, (4) SLP-OTHER was the residual category containing any other situations not included above.

The *parents' employment status* (PES) was not entirely reliable, but it was constructed combining the conditions of the father and those of the mother: (1) PES-FTD if only one or both parent were full-time salaried workers, (2) PES-FTSE if only one or both parents were full-time self-employed workers, (3) PES-PT if only one or both parents were part-time salaried or self-employed workers, (4) PES-MIX if only one or both parents were full-/part-time salaried or self-employed workers but different from the previous modalities, (5) PES-PENS if at least one of the parents was retired and the other was employed part-time, unemployed or out of labour force, and (6) PES-UOLF if at least one of the parents was unemployed or out of the labour force. Note that PAS-R and PES-PENS coincide.

The *type of contract permanent* (PRM) was a binary variable equal to 1 when the father (PRM-F) or the mother (PRM-M) had a job/work contract of unlimited duration. The *type of contract temporary* (TMP) was a binary variable equal to 1 when the father (TMP-F) or the mother (TMP-M) had a job/work contract of limited duration.

The base/ reference category for these three variables is made up of those who are not in the labour market and, hence, the dichotomous variables obtained can all enter the model.

The *tenure status of the household* (TSH) presented four modalities: (1) tenant, (2) subtenant, (3) owner, (4) free accommodation.

Other five binary variables concerned household: the amount of rent was substantial, the amount of loan/mortgage was substantial, repayment of loans to banks, there was a saving in 2008, there was a reduction of disposable income for needs.

Three ordinal/counting variables summarised certain types of information: house-evaluation, optional, and needs. The *house-evaluation* (range 0-9) counted if the dwelling had a habitable kitchen, indoor flushing toilet, cellar and/or attic, terrace and/or balcony, garden, hot water, garage, roofs or ceilings or doors or floors damaged, moisture in the walls or ceilings or floors or foundations. The *optional* (range 0-7) counted if the family had a telephone (fixed landline or mobile), a dishwasher, a fridge, a VCR-DVD player, a camera, a satellite dish/antenna, and internet access. The *needs* (range 0-7) related to the lack of money in the family for necessary food, for necessary clothes, for illness to be treated, for school, for transport, for the payment of taxes, and added the request for help to purchase essential goods.

The local and geographical variables were limited to two variables.

The *macro-region* (MR) subdivision of Italy was provided by Istat. The North-West (NW) included Valle d'Aosta, Piedmont, Liguria, and Lombardy. The North-East (NE) included Trentino Alto Adige, Veneto, Friuli Venezia Giulia, and Emilia-Romagna. The Centre (C) included Tuscany, Umbria, Marche, and Latium: it was chosen as base/reference category. The South (S) included Abruzzo, Molise, Campania, Basilicata, Apulia, and Calabria. The Islands (I) included Sicily and Sardinia.

The *degree of urbanisation* (DOU) provided three modalities: densely-, moderately-, and thinly-populated area. The latter was chosen as base/reference category.

For the sake of brevity, the description of some other qualitative variable is omitted and some of the previous variables are illustrated in Figure 3.

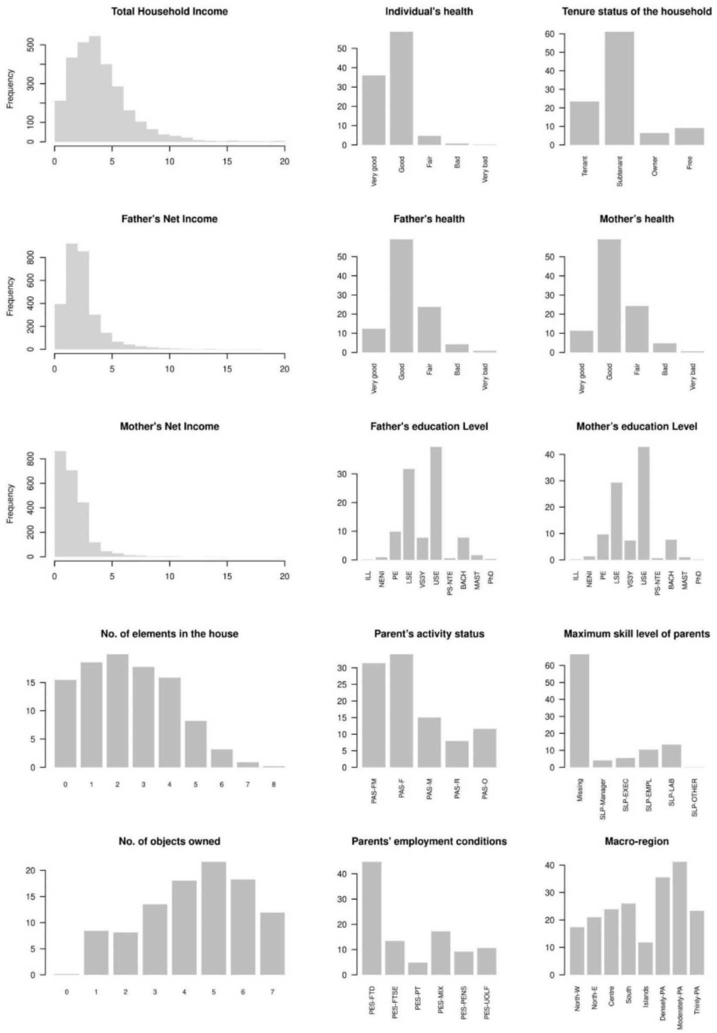

Figure 3: Descriptive plots of some qualitative and quantitative variables

Continuous variables were almost always introduced into the model through a second-degree polynomial form to capture some nonlinearities in the behaviours of individuals who took on different values in them. Only the list of these variables is reported here, for the purpose of brevity.

Age concerned the young individuals, the fathers, and the mothers: the ages were divided by 10 to have a range of values comparable with the binary variables. For the *education of parents* see above.

Income concerned many variables and components, which were all divided by 10,000. The net *disposable personal income* (DPI), as for age, concerned the young individuals, the father (FDPI), and the mother (MDPI). The net *disposable family income* (DFI) was available and *family income per capita* (FIPC) was calculated using the *number of family members*. The income variables were mutually correlated, and the correlation coefficients differed significantly from zero, but the values were surprisingly low, except for the coefficient between the total income of the family and the father's income (r=0.786, p<0.000). However, DPI should be used in the model with caution because its value was zero in the case of 1122 individuals (39.0%) and 723 of the latter (64.4%) had achieved or were currently attending tertiary education. Only 25 individuals (0.9%) reported negative income. Some continuous variables are illustrated in Figure 3.

7.2 LOGISTIC REGRESSION MODELS WITH VARIABLES BLOCKS

The original data set contained many variables describing different aspects of each individual, as stated above. A factor analysis might have aggregated them into a reduced set. However, in general, there are difficulties in understanding and interpreting these factors. As a result, some variables were omitted to reduce their number and only the original variables, most of them described above, were included in the models, sometimes with modifications and/or adaptation.

Table 7 shows the odds ratios for five different models, each one obtained adding a block of variables representing a dimension or a specific situation. Column (1) shows only the estimated odds ratios, without the standard errors for shortness, of the first block of variables referring to the young individuals (*Model 1*) constituting the sample cases. Furthermore, it also shows the estimated odds ratios of the models containing only the single next added block. The remaining fours columns concern respectively the addition of the father's data block (*Model 2*), then the addition of the mother's data block (*Model 3*), then the tenure status of the household data block (*Model 4*), and finally the addition of the block containing the Macro-Region (MR) and the degree of urbanisation.

Table 7: Logistic regressions with marginal effects and estimated odds ratio for different models: (1) single block of variables (M2) blocks 1+2, (M3) blocks 1+2+3, (M4) blocks 1+2+3+4, (M5) blocks 1+2+3+4+5

Regressor/ Block	(1)	(M2)	(M3)	(M4)	(M5)
1. YOUNG INDIV.					
Intercept	2.746^{***}	$0.246^{\#}$	0.010^{***}	0.009^{***}	0.008^{***}
Woman	1.320^{***}	1.373^{***}	1.607^{***}	1.648^{***}	1.651^{***}
$(Age/10)^2$	$0.906^{\#}$	0.925	0.939	0.950	0.938
Immigrant	0.356^{***}	0.507^{***}	0.632^{***}	0.909	0.946
DPI	0.177^{***}	0.183^{***}	0.182^{***}	0.181^{***}	0.186^{***}
DPI2 = DPI^2	1.231^{***}	1.227^{***}	1.227^{***}	1.225^{***}	1.219^{***}
SPH: Self-Perc Health	0.395^{***}	0.440^{***}	0.454^{***}	0.447^{***}	0.446^{***}
SPH: chronic illness	1.743^{*}	1.857^{**}	1.907^{**}	1.929^{**}	1.940^{**}
SPH: limitat. activ.	1.069	0.924	0.864	0.874	0.880
2. FATHER BLOCK					
Intercept	0.007^{***}				
[(Father's age)/10]	3.489^{***}	1.388	1.315	1.206	1.145
$[(Father's\ age)/10]^2$	0.919^{**}	0.997	0.982	0.985	0.990
ELF (Educ. Level)	0.973	0.945	0.925	0.910	0.901
ELF^2	1.006^{*}	1.007^{**}	$1.006^{\#}$	1.007^{*}	1.007^{*}
FDPI= (Father's DPI)	1.109^{*}	1.164^{***}	1.070	1.022	1.026
$FDPI^2$	0.998	0.996	0.999	1.000	1.000
SPHF: SPH of father	0.793^{*}	$0.794^{\#}$	0.814	0.838	0.868
SPHF: chronic illness	1.081	1.023	1.006	0.989	0.961
SPHF: limitat. activ.	1.168	1.191	1.259	$1.315^{\#}$	$1.327^{\#}$
PRM-F: permanent	1.081	0.959	0.953	0.934	0.950
TMP-F: temporary	0.830	0.878	0.932	0.990	1.042
No. of observations	2874	2874			
Log Likelihood	-1674.66^{+}	-1577.61			
Akaike Inf Criterion	3367.31^{+}	3195.22			
Bayesian Inf Criterion	3420.98^{+}	3314.49			

Notes: $^{+}$ First block only. # p<0.1; * p<0.05; ** p<0.01; *** p<0.001 (*continue*)

Table 7 (continued from previous page): Logistic regressions with marginal effects and estimated odds ratio for different models: (1) single block of variables (M2) blocks 1+2, (M3) blocks 1+2+3, (M4) blocks 1+2+3+4, (M5) blocks 1+2+3+4+5

Regressor/ Block	(1)	(M2)	(M3)	(M4)	(M5)
3. MOTHER BLOCK					
Intercept	0.002^{***}				
(Mother's age)/10	3.970^{***}		2.366^{***}	2.303^{***}	
$[(\text{Mother's age})/10]^2$	0.915^{***}		$0.954^{\#}$	$0.955^{\#}$	
ELM (Educ. Level)	1.074		1.126	1.107	
ELM^2	1.002		0.999	0.999	
MDPI= (Moth. DPI)	$1.081^{\#}$		1.160^{**}	1.145^{**}	
MDPI^2	0.998		0.993^{**}	0.994^{*}	
SPHM: SPH of mother	0.844		0.938	0.962	
SPHM: chronic illness	1.115		1.084	1.104	
SPHM: limitat. activ.	1.090		0.930	0.970	
PRM-M: permanent	1.363^{***}		1.138	1.135	
TMP-M: temporary	1.115		0.997	1.048	
4. TENURE STATUS					
Intercept	0.211^{***}				
TSH: Subtenant	1.473^{**}			1.091	
TSH: Owner	1.155			1.009	
TSH: Free	1.061			1.105	
Rent is substantial	0.957			0.963	
Loan is substantial	0.745^{*}			$0.751^{\#}$	
Repayment to bank	0.757^{**}			0.741^{**}	
House-evaluation	0.975			0.976	
No. of optional	1.325^{***}			1.183^{***}	
No. of observations	2874	2874	2874	2874	
Log Likelihood	-1674.66^{+}	−1577.61	−1520.33	−1499.51	
Akaike Inf Criterion	3367.31^{+}	3195.22	3102.65	3077.02	
Bayesian Inf Criterion	3420.98^{+}	3314.49	3287.52	3309.59	

Notes: $^{+}$ First block only. # p<0.1; * p<0.05; ** p<0.01; *** p<0.001 (*continue*)

Table 7 (continued from previous page): Logistic regressions with marginal effects and estimated odds ratio for different models: (1) single block of variables (M2) blocks 1+2, (M3) blocks 1+2+3, (M4) blocks 1+2+3+4, (M5) blocks 1+2+3+4+5

Regressor/ Block	(1)	(M2)	(M3)	(M4)	(M5)
5. MR AND DOU					
Intercept	0.629^{***}				
North-West	0.847				1.044
North-Est	$0.808^{\#}$				1.081
South	1.286^{*}				1.349^{*}
Islands	0.836				0.923
Densely-pop area	1.622^{***}				1.352^{*}
Moderately-pop area	1.293^{**}				1.354^{*}
No. of observations	2874	2874	2874	2874	2874
Log Likelihood	-1674.66^{+}	−1577.61	−1520.33	−1499.51	-1491.71
Akaike Inf Criterion	3367.31^{+}	3195.22	3102.65	3077.02	3073.42
Bayesian Inf Criterion	3420.98^{+}	3314.49	3287.52	3309.59	3341.78

Notes: $^{+}$ First block only. # p<0.1; * p<0.05; ** p<0.01; *** p<0.001

The blocks concerning the working condition of parents (13 binary variables) and the family data (five binary variables and 10 continuous variables) were not shown to reduce the length of the Table. According to the Bayesian information criterion (BIC), the *Model 3*, including the father's and mother's data, showed the best fitting model in Table 7. Overall, *Model 3* turned out to be the best model (with the lowest BIC) also adding the other two omitted blocks. According to the Akaike information criterion (AIC) the best model resulted the *Model 6*, not reported here, given by Model 5 plus the block of working conditions of parents.

doi.org/10.26398/IJAS.215

SCHISTOSOMIASIS IN UGANDA: WHAT FACTORS AFFECT MAINLY THE SPREAD OF THE INFECTION?

Francesca Bassi
Department of Statistical Sciences, University of Padova, Italy. ORCID: 0000-0002-3257-7029. francesca.bassi@unipd.it. Corresponding author.

Salvatore Ingrassia
Department of Economics and Business, University of Catania, Italy. ORCID: 0000-0003-2052-4226.

Saint Kizito Omala
Department of Statistical Methods and Actuarial Science, Makerere University, Uganda. ORCID: 0000-0003-4073-5565

Chiara Tognon
Department of Statistical Sciences, University of Padova, Italy.

Abstract. *Schistosomiasis represents a heavy burden for developing countries. In particular, the infection is quite widespread in Uganda where it reaches high prevalence in many regions. In this paper, we investigate the factors that mainly influence the probability of contracting the infection based on a dataset formed by 24,918 observations. Data was collected between April and June 2017 through a survey on households in some districts of Uganda. Due to the hierarchical structure of the data, the analysis has been carried out with multilevel regression models. Results show that hygienic conditions and the absence of water resources strongly correlate with the infection's spread.*

Keywords: NTD (neglected tropical disease), Africa, Risk factors, Prevalence

1. INTRODUCTION

Schistosomiasis is an infectious disease, also known as bilharzia, and it is caused by a species of parasitic worms of the genus *Schistosoma*. It occupies the third place among tropical infections with the most disastrous effects of mortality and morbidity in developing countries. It only precedes malaria helminthiasis infection (Ahmed, 2020). Transmission occurs through contact with water contaminated by excrements that contain parasite eggs. The larvae open once inside the human host, and the female specimens continue the reproduction by depositing new eggs (WHO, 2021a). Parasites survive, on average, between three and 10 years inside the human body, although they can survive for even 40 years. In these cases, if a patient becomes infected during childhood or adolescence, he can carry the infection up to adulthood, even after the disappearance of medical symptoms. Thus, the reported number of people infected could be underestimated. Furthermore, schistosomiasis can cause comorbidity with other infectious diseases, such as hepatitis, HIV and malaria (Ahmed, 2020). Some studies report that genital *Schistosoma* infection increases the risk of contracting HIV by three/four times (Colley et al., 2014).

From a social point of view, the disease can be disabling: in children, it can cause anaemia, growth problems, malnutrition and impaired cognitive capacity; while in adults, it can have straining economic effects owing to its impact on their ability to work. All daily activities that involve contact with unsafe water constitute a risk factor for the transmission of the infection. Schistosomiasis infections occur among the poorest and most rural communities, where the main occupations are agriculture and fishing, activities that involve assiduous contacts, without any hygienic control, with polluted water which carries the parasites. The Special Program for Research and Training in Tropical Diseases of the WHO has started the development of vaccines for the most diffused and most dangerous viruses, including schistosomiasis (Assad and Torrigiani, 1985).

Uganda is one of the countries where schistosomiasis is endemic; there are areas with a high prevalence of infection among the population. An estimated over four million people have become infected with the disease, and 55% of the Ugandan population is at risk (Loewemberg, 2014).

In recent years, the government of Uganda, with support from some other organizations, has tried to improve health conditions. In particular, the Ministry of Health devised a specific plan to control neglected tropical diseases (Borne and NTDD, 2021). The project involves specialists from various fields, including scientists and experts in communication, technicians, entomologists and analysts, who collaborate to identify the most problematic areas and plan immediate intervention.

In this framework, the present paper aims to investigate the main factors that contributed to the spread of the infection in Uganda. To this end, a dataset comprising 24,918 individuals has been analyzed, with data collected through a survey of households in some districts of Uganda. Due to the hierarchical structure of the data, the analysis has been conducted using multilevel regression models.

The rest of the paper is organized as follows: In Section 2, we provide an overview of the features and geographic distribution of schistosomiasis, also reporting measures which were put in place to limit the occurrence of the infection and outlining the goals set for the future by health authorities. We analyze, in more detail, the case of Uganda, giving an overall picture of the development of schistosomiasis in the country and a brief excursus of the demographic characteristics of the population. In Section 3, we describe our dataset, the sampling and collection method, and perform exploratory analyses. In Section 4, we introduce multilevel modelling, highlighting the importance of evaluating the inclusion of a hierarchical component in the analyses in the epidemiological field. These models and their properties are described from a statistical point of view. In particular, the multilevel logistic regression model is estimated to identify factors which mostly affect prevalence. In Section 5, we report the results of the estimation of the multilevel models. These results show which factors contribute more to increase the risk of infection, considering both within and between groups variability. Section 6 contains a discussion and concluding remarks.

2. DEMOGRAPHIC AND HYGIENIC FRAMEWORK

Uganda is located in the Central-Eastern Africa. To begin with, we illustrate the demographic and hygienic context in Uganda that influences the spread of the infection of schistosomiasis. The most common type is the *Schistosoma mansoni*, which causes infections at the intestinal level; it is estimated that more than four million people have become infected with the disease and that 55% of the Ugandan population is at risk (Loewemberg, 2014). Across the country, the prevalence in the population is 22%, according to a PMA (2020). The prevalence of the infection varies a lot, and there are communities where positivity reaches 92%, while there are areas where it is almost absent (Exum et al., 2019). The problem mainly concerns rural areas, but it can also affect urban areas, such as the district of Kampala, Uganda’s capital city, where the prevalence is 10%. The lethality of the infection is 1.8 over 100,000 inhabitants, according to WHO (2019).

The main reasons for such a high prevalence are lifestyle, habits and the lack of health and hygiene rules. Another problem that contributes to potentially infecting the water is the management of the sewerage system for the disposal of excrement. According to a report by the African Development Fund, in 2017, only 7% of the country was covered by a sewer network (African Development Fund, 2017). The most used system, especially in rural areas but also in the suburbs of large cities, is that of the latrines, which are very often without covers. Frequently, the territorial conformation and the absence of carriage roads make emptying these latrines difficult, making them unusable and, above all, creating problems when the rains carry around the faecal material in excess. For this reason, latrines are very often emptied illegally in the streams of water along the roadsides, favouring the proliferation of bacteria and parasites.

Poor health conditions are also associated with a lack of basic hygienic standard practices, such as washing hands with soap after going to the bathroom; in 2014, only 29% of the population followed this habit (Loewenberg, 2014).

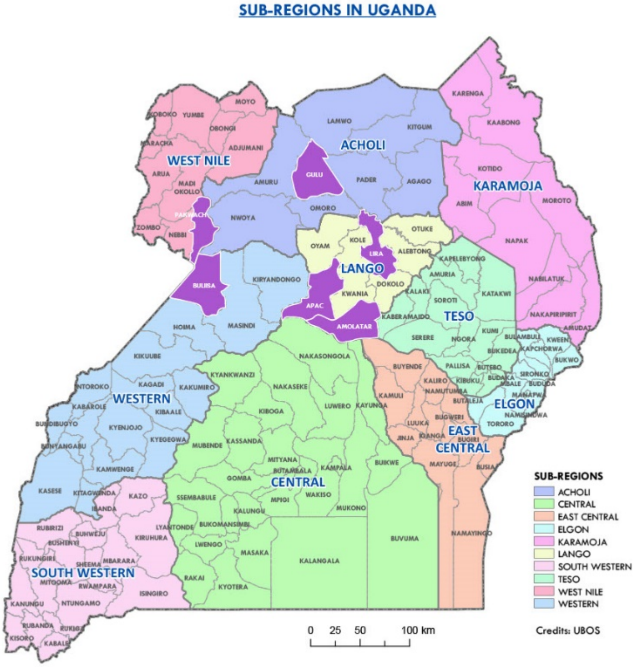

Figure 1. Districts selected for the survey (in purple)

2.1 DEMOGRAPHICAND FAMILY OVERVIEW

To better understand the problem, we present here a brief overview of Uganda’s main demographic and family characteristics. According to the data collected through the Uganda Demographic and Health Survey (Uganda Bureau of Statistics, 2021), the population amounts to 41 million, and 27% of them live in urban centres. The territory is divided into five regions and 15 sub-regions, each divided into districts and parishes (Figure 1). Buganda (Central 1-Central 2) is the most inhabited region, with nearly 10 million residents (23% of the Ugandan population). Regions also differ by household size; in Teso, a rural area, for example, the average number of household members is 5.9. In Kampala, it is lower, equal to 3.4. The population is very young (54% under 18 and 4% over 60), and life expectancy is 63 years (United Nations, Department of Economic and Social Affairs, Population Division, 2019). In some areas (South-East and

Kampala), the literacy rate is over 80%; in the North, only slightly over 30%. An important portion of the population does not have access to health facilities in case of illness, especially in rural areas, either due to the distance, the costs, or because they do not consider it necessary. As for the work situation, most of it is related to the primary sector: 68% of the population is engaged in agriculture, fishing and forestry. The data from the survey reported median monthly earnings for an employed person of UGX 200,000 (about 50 euros); 30% of households live below the poverty line, which means less than US$ 1.77 per person per month. The housing situation is an important indicator for assessing the conditions of life of the population. The fact that there are certain in-household services and features has a major impact on the health and well-being of individuals. Fundamental to guaranteeing basic sanitary conditions is the state and the typology of the toilets. The form that is commonly used is the latrine (83% of cases), while 7% of households do not have any available toilet. The toilets equipped with a drain, which guarantees greater cleanliness and hygiene, are present in only 3% of homes. In particular, in the Northeast region of Karamoja, 70% of the population lives in houses without any form of toilet. Personal cleaning services are also not guaranteed in most of the country: 82% of households do not have tools for washing hands, and only 9% of cases are there running water for washing. As for the availability of drinking water, access to clean and safe sources is guaranteed to 79% of the population. In general, the sources of water are at a distance of less than three km from the house, and only 8% have to travel a longer distance. In this context, it is evident, especially in certain areas, that there are favourable conditions for a high incidence of the infection of schistosomiasis.

As stated above, schistosomiasis prevalence in the country is greater than 22%; however, due to the many socio-economic and cultural differences among districts just described, prevalence varies greatly across the territory; moreover, we expect differences also in the factors – and their magnitude - affecting the risk of infection.

3. EXPLORATIVE DATA ANALYSIS

The data analyzed in this paper were collected between April and June 2017, with a survey of households in some districts of Uganda, some located in the North-central area of the country and the other two, Buliisa and Pakwach, as indicated in Figure 1.

The data was collected by the Ugandan Bureau of Statistics (UBOS). A multistage cluster sampling design was used to select households (Turner, 1996), which is specific for situations where resources are scarce, and there is no sampling frame of the target population (Milligan et al., 2004). This procedure solves some of the problems of the classic Expanded Program on Immunization (EPI) sampling method used by the WHO for surveys in developing countries. The EPI method is based on information available from the most recent census. Areas are selected with probabilities proportional to the number of inhabitants; this last information is, however, often inaccurate in regions where the rate of population growth, even in short periods, is very high. Households in the sampled areas are selected by the interviewer using a non-random sampling procedure (Brogan et al., 1994). In the multistage cluster sampling method, sampling areas are determined with probability proportional to the population counted in the most recent census; the difference consists in how households are selected. Using maps, sampled areas are divided into smaller segments, containing more or less the same number of houses. A sample of segments is randomly selected, and all households are included in cluster sampling.

With reference to this specific survey, 20 areas corresponding to the parishes (municipalities) within the districts were selected with probability proportional to the population. Every parish was then divided into segments with around 30 households. Interviews were administered to each head of the selected households; questions regarding schistosomiasis with reference to all household members were posed, and information on household members and characteristics of households was also collected. The achieved sample comprised 24,918 individuals from a total of 3,966 households.

Table 1. Variables collected in the survey: brief description

Name	**Description**
Second-level variables	
District	There are seven districts (6, 14, 16, 20, 21, 22, 24)
Parish	Each district contains 20 parishes
Members	Number of family components
Watersource	There are 8 categories: pipe at residence, public water tap, deep well, hand pump well, open well, river/stream, lake, spring.
Timesource	Time to reach water: less than 30 minutes, 31-60 minutes, 61-90 minutes, more than 60 minutes

Wateryear	Binary variable indicating if water is available all year or not.
Bicycle	Binary variable indicating if the family owns a bicycle.
Radio	Binary variable indicating if the family owns a radio.
Penpres	Binary variable indicating if the family owns a pen for cows.
Toilet	Type of toilet present in household: no, shared, family use only, public, other
Garbage	Binary variable indicating if there is garbage near the house.
Feces	Binary variable indicating if there are excrements near the house.
First-level variables	
Age	There are 8 categories in years: <5, 5-9, 10-14, 15-19, 20-29, 30-39, 40-49, >49.
Sex	2 categories: male, female.
Marital	5 categories: single, married, widowed, divorced, separated.
Education	5 categories: no education, primary level, secondary education, diploma, university.
Nourished	Binary variable indicating if the person is malnourished.
Extended-sprain	Binary variable indicating if the person has abdominal sprain.
Abdominal-pain	Binary variable indicating if the person has abdominal pain.
Rash	Binary variable indicating if the person has skin rash.
Canread	Binary variable indicating if the person can read.
Relation	Relation with head of household

Table 1 describes the variables used for the analyses; a few others were discarded because of too many missing data. The questionnaires administered in the district of Buliisa (number 20) presented very large percentages of missingness, therefore, we eliminated the corresponding records from the dataset. A few other records collected in other districts were eliminated either for missing data in many variables or for irrelevance, for example, data on district 14, where only five interviews were conducted. In cases of a percentage lower than 1 of missingness, we proceeded with imputation[1]. We ended up with 21,029 observations on

[1] For some variables it was possible to make imputations of missing values with the following rules for each specific variable:
"Totmembers": has been imputed with the number of the members corresponding to the same "idnumber";
"Marital": for all those under the age of 16 it was assumed that they were not married;
For the other variables, we studied the distribution of the missing values within families. If data was missing for one variable for all the components, a decision was made to eliminate the observations relating to the entire household; otherwise, if the absence of information was only associated with a family member, the household was retained in the analysis.
Other missing data was treated during estimation as missing at random.

individuals nested in 3,280 households, distributed in five districts as reported in Table 2, which also shows the prevalence of schistosomiasis in the districts and the overall population: we can observe great heterogeneity across the territories (confirmed by a Chi-squared test with p-value lower than 0,001).

Table 2. Individuals and household distribution in districts and prevalence of schistosomiasis

District	Individuals		Households		Prevalence
	n	%	*n*	%	%
6 *Pakwach*	4,050	18.39	666	20.30	33.83
16 *Gulu*	3,868	20.87	574	17.50	8.79
21 *Apac*	4,389	20.47	667	20.34	15.13
22 *Lira*	4,305	21.00	682	20.79	10.59
24 *Amolatar*	4,417	19.26	691	21.07	13.79
Total	21,029	100	3.280	100	16.36

Table 3 reports the prevalence of the disease by households' and members' characteristics; only statistically significance associations (Chi-squared test) are reported. We can see that the occurrence of the illness is positively associated with many characteristics of the individual and the household. Females, are more likely to get infected. The high heterogeneity observed for prevalence in the five districts may be due to the fact that families have different ways of life and different household's facilities in the different areas of the country. In Table 3, we also register the prevalence in the case of symptoms that are considered typical of schistosomiasis. The status of these symptoms is positively correlated to the prevalence (confirmed by a Chi-squared test; *p*-values are reported in Table A.1 together with Cramer V test).

Table 3. Prevalence by individuals and households' characteristics

Characteristic	Category	Prevalence %	Characteristic	Category	Prevalence %
Sex	Male	15.76	Radio	Radio	12.68
	Women	16.91		No radio	21.27
Canread	Able to read	12.19	Bicycle	Bicycle	13.87
	Unable to read	18.94		No bicycle	20.42
Age	Age <5	16.29	Water source	Deep well	8.67
	Age 5-9	16.57		Hand-pump well	18.05

	Age 10-14	15.43		Lake	7.78
	Age 15-19	18.04		Open well	13.17
	Age 20-29	13.68		Pipe	4.44
	Age 30-39	15.21		Public water tap	8.15
	Age 40-49	17.35		River	36.90
	Age >49	21.72		Spring	10.59
Education	No education	20.85	Water in a year	Water all year	14.27
	Primary education	12.70		Water not all year	23.18
	Secondary education	9.84	Time to water source	minutes to water <31	15.56
	Certificate/Diploma	10.51		minutes to water 31-60	15.39
	University degree and above	9.76		minutes to water >60	21.73
Waste management	Garbage near house	23.58	Rash	Rash	36.27
	No garbage near house	8.08		No rash	13.09
Faeces	Excrements near house	32.36	Abdominal-pain	Abdominal pain	28.54
	No excrement near house	6.86		No abdominal pain	11.38
Toilet	No toilet	18.87	Extended-spleen	Extended sleen	31.51
	Shared toilet	15.00		No extended spleen	12.98
	Household toilet	10.84	Nourished	Malnourished	23.07
	Public toilet	9.00		Nourished	12.92
	Other	25.32			

4. MULTILEVEL MODELLING IN EPIDEMIOLOGY

Multilevel models have always been applied, particularly in economic and social studies, where we often deal with hierarchical data (Guo and Zhao, 2000). Using the multilevel approach, it is possible to take into account the fact that observations might not be independent because they are nested in higher-level units. Traditional methods of statistical inference assume that observations are independent, but this is not necessarily true in hierarchical structures as in the study sample. Multilevel modelling accounts for eventual correlation among first-level units.

In epidemiological studies, the multilevel approach has historically been used less, as the focus of the investigations is usually on individual risk factors that influence the occurrence of a disease or on aggregated data (Weinmayr et al., 2017). In this case, the idea is that the individual's aspects alone are capable of explaining the causes of a disease (Diez Roux and Aiello, 2005), and it is not allowed to simultaneously consider individual and group effects on the risk of contracting the disease. More recently, the interest in multilevel analysis has also increased in epidemiological studies in order to include the possibility of variability both within and between groups.

An example in which the importance of the multilevel component is evident is the study on the propensity to smoke in adolescents, which may depend on how widespread smoking is in the group of peers over individual characteristics (Diez Roux and Aiello, 2005). Another example is that of HIV or other sexually transmitted diseases, where the transmission rate can be determined by social norms and behavioural habits of the group to which individuals belong, not only by individual behaviour (Diez Roux and Aiello, 2005).

In the epidemiological context, there is a considerable number of diseases that have a highly variable prevalence in different geographical areas. In these cases, it is important to consider the effects of contextual factors on individual health outcomes (Weinmayr et al., 2017). This approach also makes it possible to plan interventions both at the individual and community level while not considering the subdivision into groups, and there is an actual risk of drawing incorrect conclusions. Taking into account the hierarchical structure of the data considers that observations are not independent (Guo and Zhao, 2000); assuming independence, in this case, would result in biased estimates. Multilevel models permit estimation of the impact of covariates at the different levels; moreover, total variance can be decomposed in order to assess how much variability in the data is due to within and between groups factors.

4.1. MULTILEVEL REGRESSION MODEL

Equation (1) describes a multilevel regression model,

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij} \qquad (1)$$

y_{ij} is a dependent variable, observed on unit *i* (first level), belonging to group *j* (second level), $i=1,..,n_j$, $j=1,...,J$; x_{ij} is a first-level covariate, and ε_{ij} is a random error with distribution $N(0,\sigma^2)$. β_{0j} and β_{1j} are, respectively, the random intercept and the random slope.

In the case of the random intercept model, we can calculate the intra-class correlation coefficient (ICC), which is a measure of the proportion of total variance explained by between groups variability:

$$ICC = \frac{\tau_0^2}{\sigma^2 + \tau_0^2}.$$

A low value of the ICC indicates that between groups variability is low; therefore, it is not necessary to estimate hierarchical models. The minimum value of ICC necessary to perform multilevel modelling depends on sample size and other data characteristics (Musca et al., 2011).

In epidemiology, we frequently deal with a binary response variable to indicate the presence or absence of a disease. Indicating with p_{ij}=P(y_{ij}=1), a random intercept multilevel logistic regression model is defined by equation (2) (Merlo et al., 2016):

$$logit(p_{ij}) = \gamma_0 + \textstyle\sum_{h=1}^{r} \gamma_h x_{hij} + U_{0j} \qquad (2)$$

where x_h, with $h=1,..,r$, are first-level covariates; γ_0 is the mean of p_{ij} in the logistic scale and, in epidemiology, it can be seen as prevalence (Snijders and Bosker, 1999). In this case, the ICC has the following form:

$$ICC = \frac{\tau_0^2}{\frac{\pi^2}{3} + \tau_0^2}.$$

In the case of the multilevel logistic regression model, an alternative measure of between groups variability is the Median Odds Ratio (MOR):

$$MOR = \exp\left(\sqrt{2\tau_0^2}\phi^{-1}(0.75)\right) = \exp\left(\sqrt{2\tau_0^2}\cdot 0.6745\right) \cong \exp\left(\sqrt{\tau_0^2}\cdot 0.95\right).$$

where ϕ^{-1} is the 75th percentile of the standard Normal distribution. This indicator, such as the ICC, can be used to compare alternative models. It assumes values ≥1; being equal to 1 in the case of no between groups variability.

5. NUMERICAL RESULTS

In order to study which factors have the greatest effects in determining the probability of contracting the infection, we estimated logistic regression models. The classic logistic model (no multilevel) is used as the baseline to evaluate the improvement introduced by a multilevel component in the analysis. Different multilevel regression models were tested, selecting the best in terms of the lowest value of the ICC and MOR. For now, we focus on the results of the analysis, while the numerical details of fit are given in the Appendix.

In the ordinary logistic regression model (that is, no multilevel) for evaluating the probability of contracting the infection, we considered the following predictors: symptoms and individual and household characteristics; a few variables, like extended spleen and bicycle ownership, were eliminated due to multicollinearity issues.

The results showed that living in different districts has a significant effect. For example, in reference to district 6, the other districts exhibited a decrease in the probability of becoming ill. Household factors that have a statistically significant effect on the probability of contracting the illness are the type of water source, specifically when it is a river or stream; the walking time necessary to reach the water source: the longer the time, the higher the risk; the absence of a toilet in the house; the presence of garbage and excrements near the house; and the household size.

Among the individual level effects, we observe that age increases the probability of having the disease, as well as the fact that the individual is unable to read. Quite surprisingly, possession of symptoms has no significant effect.

We now proceed to evaluate whether the insertion of a second-level component, i.e., considering the hierarchical structure of our data, improves the model. As already stated, districts in Uganda differ for many characteristics that affect individuals and households, and the prevalence of schistosomiasis shows a great variability across the country. The ordinary logistic regression model assumes that individuals who make up the sample belong to the same population. Instead, it is reasonable to assume that there are differences between districts that generate a correlation structure in the variables collected on household located in the same area. The multilevel logistic regression model estimates the effects of covariates on prevalence, taking into account this correlation structure.

To this end, we estimated, as a starting point, a logistic random intercept model without covariates (unconditional) with the probability of contracting schistosomiasis as the dependent variable. This model takes into account the idea that individuals are nested within households and therefore, observations on members of the same family may be correlated. The values of the ICC and the MOR indicate significant association within the groups; therefore the multilevel approach is appropriate for analyzing these data (see Table A.2). Further, we insert covariates that might affect prevalence and improve model fit and having an effect on the variability explained by the model.

Table 4. Conditional logistic random intercept model for the probability of contracting schistosomiasis: estimation results

Variable	Estimate	Standard error	p-value	Odds-ratio
Constant	-4.63	0.43	0.00^{*}	
District 6 ref.				
16	-2.59	0.26	0.00^{*}	0.07
21	-1.40	0.25	0.00^{*}	0.25
22	-1.83	0.26	0.00^{*}	0.16
24	-1.22	0.24	0.00^{*}	0.29
Timesource <31 minutes ref.				
31-60 minutes	0.41	0.14	0.00^{*}	0.21
>60 minutes	1.98	0.18	0.00^{*}	0.52
Watersource hand-pump well ref.				
Deep well	-1.15	0.14	0.02^{*}	0.32
Lake	-1.45	0.18	0.00^{*}	0.23
Open well	-0.23	0.19	0.23	

Pipe at residence	-0.49	0.69	0.48	
Public water tap	0.12	0.38	0.74	
River/stream	0.31	0.27	0.26	
Spring	-0.11	0.25	0.44	
Wateryear	0.04	0.15	0.80	
Toiliet family ref.				
No	0.75	1.47	0.61	
Shared	0.58	1.40	0.68	
Public	-0.04	0.62	0.95	
Other	-0.78	0.17	0.00*	0.46
Radio	0.26	0.16	0.10	
Garbage	0.39	0.15	0.01*	1.47
Feces	2.19	0.17	0.00*	8.97
Members	0.23	0.03	0.00*	
F_age	1.03	0.01	0.00*	
F_canread	-1.87	0.30	0.00*	
Nourished	-0.03	0.13	0.81	
Rash	-0.06	0.13	0.63	
Abdominal-pain	0.08	0.07	0.31	
Canread	-0.17	0.10	0.09	
Education 1 ref.				
2	0.00	0.09	0.99	
3	-0.13	0.17	0.44	
4	0.33	0.28	0.23	
5	-014	0.41	0.72	
Marital 1 ref				
2	-0.09	0.12	0.70	
3	-0.07	0.17	0.93	
4	-0.12	0.19	0.78	
5	-0.08	0.56	0.69	
Sex male ref.	0.06	0.07	0.32	
Relation 1 ref.				
2	-0.05	0.12	0.70	
3	-0.02	0.17	0.93	
4	-0.05	0.19	0.78	
5	0.22	0.56	0.69	
Age in years	1.00	0.00	0.99	
τ^2_0	6.78	0.45	0.00*	
Model fit	ICC=0.6732	MOR=11.45	AIC=12,503	

* statistically significant at 5%.

Table 4 contains the estimates of the best fitting logistic random intercept model with covariates (conditional); covariates have been selected on the basis of the reduction of the ICC and MOC; alternative models have been compared with the AIC index.

With respect to the traditional logistic model, the multilevel specification has a better fit, showing a lower AIC value. Statistically significant estimated coefficients have the same sign in the two models, but different magnitude. Some second-level variables revealed statistically insignificance when considering the hierarchical structure of the data (all levels of the variable describing the type of toilet, except "others", and if water is available all year). The effects of the other second-level variables show larger estimates in the multilevel model. We inserted two new second-level variables, measuring the average age of household members (*F_age*) and the percentage of household members who can read (*F_canread*); these variables are correlated with the occurrence of schistosomiasis. In the multilevel approach, in general, first-level variables result appears less important than in the traditional logistic model.

Table 4 shows also the estimated odds-ratio for the statistically significant levels of categorical variables. These odds-ratios allow to understand better the effect of each covariate and its levels on the risk of schistosomiasis. The probability of contracting schistosomiasis increases by 47% if there is the presence of garbage near the house, and it is almost nine times higher in presence of feces. Risk of disease increases also with the distance to the water source, being 50% greater when the required time is between 30 and 60 minutes, and almost three times bigger for distances that require more than one-hour journey. Getting water from a deep well or a lake decreases the risk of infection with respect to the use of a hand pump. People leaving in districts different from Pakwach (number 6) have a lower risk of contracting the illness. From the statistically significant continuous variables, we see that one additional family member increases the risk by 25%; one year change in household members' average age increases the risk by 3%; a higher proportion of household members who can read decreases the risk.

Finally, although the multilevel logistic regression model improves the fit to the data explaining part of within-groups correlation, still the ICC and the MOR have high values, indicating that it is necessary to specify an even better model to adequately identify the factors affecting schistosomiasis. In particular, the results obtained with the logistic multilevel regression model indicate that this is a sort of household disease since many characteristics of the household and the its location appear as important. Moreover, it is very plausible that occurrence of the disease tends to happen within members of the same family. In our dataset, on average, in families where at least one member suffers from schistosomiasis, 50% of household members are ill. For these reasons, we decided to explore which conditions may cause the fact that at least one family member is infected. The following analyses, then, focus on the probability of observing at least one infection in the household. With this scope, we considered observations on 3,280 families, who become our first-level units. Information at individual level is aggregated with reference to the corresponding household. Second-level units are the 96 parishes or municipalities in which the families live.

Table 5 lists estimation results of a logistics multilevel regression model for the probability of observing at least one member of the family affected by schistosomiasis; only significant estimates are reported. We created some new variables summarizing characteristics of parishes: average number of family components (*P_members*), proportion of families who need to move for at least one hour to reach water (*P_timesorce60*), proportion of families with garbage and excrements near the house (*P_garbage*, *P_feces*), average age of family members (*P_age*). With reference to these second-level variables, we inserted in the model also the contextual effects; i.e., the deviations from the group mean for each family (*P_timesorce60_c*, *P_members_c*, *P_garbace_c*, *P_feces_c*) as suggested by Feaster et al. (2011).

Table 5. Conditional logistic random intercept model for the probability of observing at least one infected member in the family: estimation results

Variable	Estimate	Standard error	p-value	Odds-ratio
Constant	-4.91	1.26	0.00^{*}	
P_members	0.23	0.02	0.00^{*}	1.77
P_timesource_60	2.39	0.84	0.00^{*}	10-94
P_garbage	1.65	0.59	0.00^{*}	5.20
District 6 ref.				
16	-2.10	0.43	0.00^{*}	0.12
21	-1.08	0.39	0.01^{*}	0.34
22	-1.14	0.42	0.00^{*}	0.32
24	-0.66	0.40	0.10	
Watersource hand-pump well ref.				
Deep well	-0.78	0.36	0.03^{*}	0.46

Lake	-0.92	0.42	0.03*	0.40
Open well	-0.46	0.18	0.01*	0.63
Pipe at residence	0.34	0.53	0.51	
Public water tap	0.09	0.35	0.80	
River/stream	-0.10	0.31	0.74	
Spring	-0.21	0.23	0.37	
Toiliet family ref.				
No	0.50	1.11	0.65	
Shared	0.52	0.99	0.60	
Public	0.05	0.45	0.91	
Other	-0.65	0.14	0.00*	0.52
Feces	0.59	0.13	0.00*	1.81
F_age	0.02	0.01	0.03*	1.02
P_timesource_60_c	0.45	0.13	0.00*	1.56
P_members_C	0.23	0.02	0.00*	1.25
τ^2_0	1.00	0.20	0.00*	
Model fit	ICC=0.2333	MOR=2.60	AIC=3,433	

* statistically significant at 5%.

The risk of observing at least one infected person in the family decreases by 88%, 66% and 68% respectively, living in districts 16, 21, 22, with respect to district 6. Families who take water from lakes or wells show lower risks with respect to families using hand pumps; also the type of toilet access, specifically respondents that has a toilet used only a household, diminishes the risk. The risk of at least one infection in the family, on the other hand, increases with the average age of family members and excrements near the house. With reference to the parish where the family lives, factors that positively affect the risk are the average number of members in each household, the proportion of families who have a travel time greater than one hour to reach the source of water and the proportion of families in the parish who have garbage around the house. Estimation results, in this case, show that factors affecting schistosomiasis in families are related both to the characteristics of the household and of the parish where the family lives. Statistically significant second-level variables indicate that there is non-negligible heterogeneity between parishes.

6. CONCLUSIONS

In this paper, we study factors affecting the diffusion of schistosomiasis in some districts of Uganda, taking into account the hierarchical nature of the phenomenon: infected individuals belong to households, which are clustered in parishes. As recognized by the reviewed literature (Diez Roux and Aiello, 2005), in epidemiological studies, it is important to consider the multilevel structure of the data in order to obtain reliable estimates. Observations on members of the same household are correlated, as well, as observations on households living in the same parish; not considering this correlation in model specification, gives rise to biased estimates.

Estimating logistic random intercept models with our data allows us to correctly evaluate the effects on contracting schistosomiasis of all variable-levels: characteristics of individuals, families and parishes; moreover, total variance can be appropriately decomposed across within and between-groups parts.

Statistical analyses of our data identified significant associations between the prevalence of schistosomiasis and individual characteristics; however, the effects of individual traits disappeared in multilevel estimation in favour of family characteristics, especially relating to the hygienic conditions in and around the house. This result is consistent with other studies from the reference literature that identify poor sanitation as one of the main risk factors for infection (see, for example, Colley et al., 2014). Another family-related important risk factor emerging from our analyses is the contact with polluted water, as also indicated by WHO (2021), especially those families who have a long journey to reach the water source are more exposed to schistosomiasis. These results highlight the need to strengthen public information campaigns as both WHO and the Ugandan Ministry of Health are preoccupied with the prevention of the occurrence of schistosomiasis (Borne and NTDD, 2021; WHO, 2021b). These campaigns should be especially targeted for families with the highest illiteracy rates, who are also those most at risk of infection. Another significant and positive effect on the risk of infection is due to the mean age of family members, which can be explained by referring to the results in the literature on the longevity of the infection in the human body (Colley et al., 2014). Usually, schistosomiasis is contracted at a young age, and then it remains in the human body for an important part of adult life. In endemic areas, such as in the case of districts of Uganda that we analyzed, the most widespread form of schistosomiasis has a chronic characteristic due to repeated contact with infected larvae. In larger families, the risk of getting sick increases.

When estimating the logistics random intercept model for the probability of at least one infection in the family, other important aspects emerged regarding the area where the family lives: there is evidence of high heterogeneity between the different municipalities or parishes. Again, hygienic conditions in the area are strictly linked to the risk of infection. In particular, the risk is higher in areas where there is a greater proportion of households with garbage near the house or more distant from water sources. Areas with larger families also show a higher risk of schistosomiasis in the household. An important result emerging from all models is the difference between the district of Pakwach (number 6) and the others. Pakwach district has a different geographical location, being in the region of the West Nile, near the White Nile, while the others are in the central belt of the country. In Pakwach, fragility and backwardness conditions regarding developing countries emerge more sharply (UBOS, 2021). In fact, in this area, we find the highest percentages of houses with faeces or garbage around, illiteracy rates, and malnourished individuals.

For what concerns future developments of this research, it might be of some interest to estimate random slope models. Our multilevel logistic regression models assumes that the models describing prevalence in the districts have all the same slope; we allow for randomness only for the intercept. Incorporating random variation also in the slope might increase model fit.

Referring to the preventive treatments that are carried out in the areas where the disease is endemic, in a future study, it could be interesting to use this information, if available, to analyze the impact of the treatment on the probability of contracting schistosomiasis on the individual and the whole family unit.

Declarations

Ethical Approval
Not applicable

Competing interests
Not applicable

Authors’ contributions
The authors equally contributed to the analyses and the manuscript.

Funding
Not applicable

Availability of data and materials
Not applicable

REFERENCES

African Development Bank Group (2017), *Program: Kampala Sanitation Program-Phase 1 Country: Uganda* https://projectsportal.afdb.org/dataportal/VProject/show/P-UG-E00-008, accessed 14/01/22.

Ahmed, S.H. (2020), Schistosomiasis (Bilharzia), *Medscape*, https://emedicine.medscape.com/article/228392-overview, accessed 14/01/22.

Assaad, F., Torrigiani, G. (1985), Who's vaccine development programme. *European Journal of Epidemiology*, Vol.1, pp. 1–4. https://doi.org/10.1007/BF00162305

Borne V. and Neglected Tropical Diseases Division, Ministry of Health, Uganda (2021), *Sustainability Plan for Neglected Tropical Diseases Control Program 2020–2025*, https://www.health.go.ug/cause/sustainability-plan-for-neglected-tropical-diseases-control-program-2020-2025/, accessed 14/01/22.

Brogan D, Flagg EW, Deming M, Waldman R. (1994), Increasing the accuracy of the Expanded Programme on Immunization's cluster survey design. *Annals of Epidemiology*, Vol. 4, pp. 302–311.

Casulli A (2021), New global targets for NTDs in the WHO roadmap 2021–2030, *PLOS Neglected Tropical Diseases*, 15(5): e0009373, https://doi.org/10.1371/journal.pntd.000937.

Cohen J. (2016), Unfilled Vials, *Science*, Vol. 351(6268), pp. 16-19

Colley D.G., Bustinduy A.L., Secor W.E. and King C.H. (2014), Human schistosomiasis, *Lancet*, Vol. 383 (9936), pp. 2253–2264.

Diez Roux A.V. and Aiello A.E. (2005), Multilevel Analysis of Infectious Diseases, *Journal of Infectious Diseases*, Vol. 191 Suppl 1, pp. S25–S33.

Exum N.G., Kibira S.P.S., Ssenyonga R., Nobili J., Shannon A.K., et al. (2019), The prevalence of schistosomiasis in Uganda: A nationally representative population estimate to inform control programs and water and sanitation interventions, *PLoS Neglected Tropical Diseases*, Vol. 13(8), e0007617, https://doi.org/10.1371/journal.pntd.0007617.

Feaster D., Brincks A., Robbins M. and Szapocznik J. (2011), Multilevel models to identify contextual effects on individual group member outcomes: a family example, *Family Process*, Vol. 50(2), pp. 167–183.

Guo G. and Zhao H. (2000), Multilevel Modeling for Binary Data, *Annual Review of Sociology*, Vol. 26, pp. 441-462.

Lee V. E. and Bryk A. S. (1989), A multilevel model of the social distribution of high school achievement, *Sociology of Education*, Vol. 62(3), pp.172–192.

Loewenberg S. (2014), Uganda's struggle with schistosomiasis, *Lancet*, Vol. 383(9930), pp. 1707–1708.

Merlo J., Wagner P., Ghith N., and Leckie G. (2016), An Original Stepwise Multilevel Logistic Regression Analysis of Discriminatory Accuracy: The Case of Neighbourhoods and Health, *PLoS One*, Vol. 11(4), e0153778, https://doi.org/10.1371/journal.pone.0153778-

Milligan P., Njie Al. and Bennett S. (2004), Comparison of two cluster sampling methods for health surveys in developing countries, *International Journal of Epidemiology*, Vol. 33, pp. 469–476.

Musca S., Kamiejski R., Nugier A., Méot A., Er-rafiy A., Brauer M. (),Data with Hierarchical Structure: Impact of Intraclass Correlation and Sample Size on Type-I Error, *Frontiers in Psychology*, Vol. 2, Doi: 10.3389/fpsyg.2011.00074.

PMA (2020) https://www.pmadata.org/sites/default/files/data_product_results/PMA2020-Uganda-R1-Sch-brief.pd (accessed 4 August 2022).

Ross R. (1916), An application of the theory of probabilities to the study of a priori pathometry: part I, *Proceedings of the Royal Society Series A*, Vol. 92, pp. 204–30

Snijders T. and Bosker R. (1999), *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling (1 ed.)*, SAGE Publications, New York.

Uganda Bureau of Statistics (UBOS) (2021), *Uganda National Household Survey 2019/2020*, Kampala, Uganda.

United Nations, Department of Economic and Social Affairs, Population Division (2019), *World Population Prospects 2019*, Online Edition, Rev. 1, https://population.un.org/wpp/, accessed 14/01/22.

Van der Werf M.J., de Vlas S.J., Brooker S., Looman C.W., Nagelkerke N.J., Habbema J.D. et al. (2003), Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa., *Acta tropica*, Vol. 86(2-3), pp. 125–139

Weinmayr G.,Dreyhaupt J., Jaensch A., Forastiere F. and Strachan D.P. (2017), Multilevel regression modelling to investigate variation in disease prevalence across locations, *International Journal of Epidemiology*, Vol. 46(1), pp. 336–347.

World Health Organization (2019), *Global Health Estimates 2019: Deaths by Cause, Age, Sex, by Country and by Region, 2000−2019*, Geneva, https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death, accessed 14/01/22.

World Health Organization, The Global Health Observatory (2020), *Schistosomiasis: Status of Endemic Countries*, www.who.int/data/gho/data/themes/topics/schistosomiasis, accessed 14/01/22.

World Health Organization (2021a), *Schistosomiasis*, www.who.int/news-room/fact-sheets/detail/schistosomiasis, accessed 14/01/22.

World Health Organization (2021b), *Ending the Neglect to Attain the Sustainable Development Goals: a Road Map for Neglected Tropical Diseases 2021–2030. Overview*, Geneva, https://www.who.int/publications/i/item/9789240010352, accessed 14/01/22.

APPENDIX

Table A.1. p-values of the Chi-square and Cramer V test for association between each variable and prevalence

	Chi-square	Cramer V
Variable	p-value	
District	<0.001	<0.001
Parish	<0.001	<0.001
Members	<0.001	<0.001
Watersource	<0.001	<0.001
Timesource	<0.001	<0.001
Wateryear	<0.001	<0.001
Bicycle	<0.001	<0.001
Radio	<0.001	<0.001
Penpres	<0.001	0.000
Toilet	<0.001	<0.001
Garbage	<0.001	<0.001
Feces	<0.001	0.000
Age	0.011	0.012
Sex	0.025	0.087
Marital	0.115	0.880
Education	<0.001	<0.001
Nourished	<0.001	<0.001
Extended-sprain	<0.001	<0.001
Abdominal-pain	<0.001	<0.001
Rash	<0.001	<0.001
Canread	<0.001	<0.001
Relation	<0.001	<0.001

Table A.2. Unconditional logistic random intercept model: estimation results

	Estimate	Standard error	p-value
Constant γ_0	-4.01	0.12	0.00^*
τ^2_0	13.01	0.85	0.00^*
Model fit	ICC=0.7989	MOR=11.98	AIC=13,092

doi.org/10.26398/IJAS.239

MEASURING GENDER EQUALITY IN THE EUROPEAN UNION: SCRUTINIZING THE GENDER EQUALITY INDEX

Milica Maričić

Department of Operational Research and Statistics, University of Belgrade – Faculty of Organizational Sciences, Serbia. ORCID: 0000-0003-0441-9899. milica.maricic@fon.bg.ac.rs. Corresponding author.

Veljko Jeremić

Department of Operational Research and Statistics, University of Belgrade – Faculty of Organizational Sciences, Serbia. ORCID: 0000-0002-0266-5247.

***Abstract**. Over the last few decades, gender (in)equality has become a topic of high interest because it has possible implications for the global economy and the overall level of sustainability. With this in mind, the policymakers needed an accurate, reliable, and multidimensional measure of the level of gender (in)equality. Indicator-based measurement approaches deemed as an obvious solution. Among several composite indicators in the field of gender equality, the Gender Equality Index (GEI), devised by the European Institute for Gender Equality, attracts attention for its methodology, structure, and the number of yearly publications. Considering the possible consequences of GEI results on the EU level, our study aims to analyze the methodological choices of this composite measure, precisely to review its current indicator list and weights assigned to them. To conduct the GEI analysis, we applied a statistical I-distance method, i.e. a methodology that can overcome the issue of subjective weight assignment. The performed twofold I-distance approach gave us an insight into domains and total score dynamics, while the applied Composite I-distance Indicator (CIDI) methodology proposed corrections of domain weights. Finally, through the iterative exclusion of indicators by the level of their relevance, using the post hoc I-distance, we provide an in-depth analysis of the countries' rank consistency depending on the number of remaining framework indicators. The obtained results indicate that the expert-driven weights assigned to domains are supported by the data and are unbiased, but that there is place for reducing the number of framework indicators.*

***Keywords**: Gender Equality Index, European Union, ranking of countries, I-distance method, CIDI methodology*

1. INTRODUCTION

The interest and attention given to the concept of gender equality have increased over the last few decades because of its relationship with economic growth at the macro-level and micro-level (Kabeer and Natali, 2013; Girón and Kazemikhasragh, 2021). One of the first academics to point out this relation and its possible contribution is Klasen (1999). He stated that gender equality in human resource management could impact overall economic growth through *optimal use of human resources* and *family relations*. The perspective of *optimal use of human resources* suggests that gender equality will raise the productivity of human capital. On the other side, *the family relations* perspective anticipates that the next generation's productivity will increase as the positive work experience will be transmitted from mother to children. The precondition for both perspectives to have an effect is education. Growth-related impacts are the results of investments in women's education. In addition, evidence from earlier research indicates that investment in women's social capital has a higher ROI factor than in men regarding non-market return (Leeves and Herbert, 2014). Some investment effects, such as a rise in employment rates and earnings, are visible in the short term. Nevertheless, cumulative and large effects are to be felt nationwide only in the long term, i.e. in 25 years or more (Appiah and McMahon, 2002).

Another relevant aspect of gender equality, which is not visible at first sight, is its relationship with sustainability. The recognition that these two concepts are intertwined has increased in recent decades (UN, 2014a), leading to a better understanding of each. There are several reasons for this relation to appear. First, a sustainable future should be built on moral and ethical standards, which implies gender equality. Secondly, the contribution of women's knowledge can have the crucial potential to help society create a more sustainable environment and economy (Cela et al., 2013; Birindelli et al., 2019).

The acknowledgement of the gender equality concept led to a steady rise in the importance of gender (in)equality measurements (Permanyer, 2010; Dilli, Carmichael and Rijpma, 2019). The history of gender (in)equality metrics began with Kersti Yllo. In 1984, she published the article "The Status of Women, Marital Equality, and Violence against Wives" in which she presented her composite index for measuring gender inequality (Bericat, 2012). However, the importance of her ideas on gender equality measurements was recognized more than ten years later at the 1995 United Nations World Conference on Women

(Amici and Stefani, 2013). Since then, gender equality indices have proliferated on a worldwide level. Today, measuring gender (in)equality is a topic of high importance that has relevant policy applications and implications. However, such measures have not received the attention from the academic community and literature they deserve (Permanyer, 2010; Amin and Sabermahani, 2017).

Gender equality rankings can be used to stimulate countries to focus their attention on gender inequality problems and to introduce policies aimed at its reduction (Permanyer, 2010). For this reason, it is crucial to provide a transparent and unbiased ranking. Accordingly, this study will try to address several related topics. First, we will discuss several gender (in)equality indices and their frameworks to observe some of their imperfections. Secondly, we will review the methodological choices of the *Gender Equality Index* (GEI) (EIGE, 2022), a multidimensional composite index of gender equality created to provide a measure across 27 EU member states.

The United Nations Development Program (UNDP) was the first to develop a composite index on gender equality. In 1995, the UNDP published two indices with the idea of capturing gender disparities at the world level: the *Gender Development Index* (GDI) and the *Gender Empowerment Measure* (GEM) (Amici and Stefani, 2013). The GDI is often called Gender-related HDI as it is computed by calculating the score of each HDI dimension of the index for the two genders separately. Countries are later ranked based on the absolute deviation from the gender parity in the HDI (UNDP, 2023). On the other hand, GEM is a complement to the GDI. It encompasses what GDI does not - women's participation in political and economic life (Amici and Stefani, 2013). In 2000, for the 20th anniversary of Human Development Report, the UNDP presented a new measure: *Gender Inequality Index* (GII). The GII is another composite index with a goal to measure women's disadvantage in three dimensions - empowerment, economic activity and reproductive health (Permanyer, 2013). The World Economic Forum created the most recent global composite index in 2006: *the Global Gender Gap Index.* This index aims to capture the scale of gender-based disparities by tracking the country's progress in equal economic participation, educational attainment, political empowerment, and health and survival (WEF, 2022).

All these indices still have a highly complicated measuring system with conceptual and methodological flaws such as subjective weighting process and/or questionable framework indicators (Klasen, 2006; Permanyer, 2013;

Elias, 2013). Learning on their limitations, Plantenga and associates (2009) created the *European Union Gender Equality Index* (EUGEI). They wanted to emphasize the importance of measuring and determining gender equality across the EU countries. The substantial improvement this index brought to further gender equality measurement was the inclusion of unpaid time. Taking time into account was an important advancement as equal distribution of unpaid work is a precondition for an equal distribution of paid work (Plantenga et al., 2009).

Many of the indicators mentioned above were created to measure various aspects of gender (in)equality at the global level. Accordingly, they have fallen short of providing the kind of measures that would start a debate and contribute to decisions made at the EU and member-state levels. In addition, the EUGEI also did not have the expected impact on EU policymakers. To answer the lack of effective quantitative measurement of gender equality, the European Institute for Gender Equality (EIGE) created the Gender Equality Index (GEI). GEI is a composite indicator with a three-level structure (indicators – sub-domains – domains) aimed at ranking EU member states based on the achieved level of gender equality.

Considering the importance of the GEI, its results, and the policy implications it might have, it is important to evaluate this composite indicator's structure and methodological choices. The process of composite index creation and development encompasses ten steps as defined in the OECD's handbook (OECD, 2005): Theoretical framework, Data selection, Imputation of missing data, Multivariate analysis, Normalisation, Weighting and aggregation, Uncertainty and sensitivity analysis, Back to the data, Links to other indices, and Visualization of the results. In the presented study, we focused on exploring the GEI methodological choices for the 6th step: weighting and aggregation. We question if the chosen weighting scheme is appropriate and whether the index structure could be simplified by assigning some indicators the zero-weight. Among many statistical methods and methods of operational research available, we decided to use a statistical, multivariate, data-driven, distance-based analysis, the I-distance method (Ivanovic, 1977; Maričić et al., 2019). This method has been extensively used in composite indicator creation and evaluation (for example Jeremic et al., 2011; Maricic and Kostić-Stanković, 2014). I-distance stands out among many methods because it allows for the ranking of entities, without the need of the decision-maker to provide inputs on weights. Namely, the method can, based on the data, suggest data-driven weights. What is also convenient is the fact that the method can be used as post hoc analysis for exploring the composite indicator structure.

In this study, we applied the I-distance method to the GEI to evaluate its weighting scheme and structure. The analysis was threefold: first, we performed the two-fold I-distance method to aggregate the sub-domain values to domains and domains to overall I-distance values; second, we proposed new domain weights; and finally, we performed the post hoc I-distance to explore the GEI structure.

The paper is conceptualized as follows. The next section features the GEI and its structure. Next, we give an overview of the statistical multivariate method used to perform the analysis – the I-distance method and related analyses. In the section that comes after, we present and highlight the results obtained after scrutinizing the official data set for 27 member states for the year 2022. In the last two sections, we provide discussion and concluding remarks.

2. GENDER EQUALITY INDEX (GEI)

Equality between women and men is one of the EU's fundamental values incorporated in its Treaties, including the Charter of Fundamental Rights of the European Union (EUR-Lex, 2012). The European Commission concluded that the EU needed a composite index that would measure the level of gender equality in its member states with high precision. Accordingly, the first task put in front of the newly formed and long-awaited European Institute for Gender Equality (EIGE) was the creation of a gender equality index. After three years of devoted work, the Gender Equality Index (GEI) was created in 2013.

Through multiple dimensions, the GEI aims to picture how close the EU and member states have come towards achieving gender equality. Besides its complex assignment, the GEI is easy to understand and to communicate its idea as a mean of gender equality promotion. The GEI is also able to measure the progress of each member state over time (EIGE, 2022).

The index comprises six domains, which make the core index and an additional satellite domain. The satellite domain *Violence* is not included in the core framework as it focuses statistically on violence against women (EIGE, 2022). The EIGE stated that this domain is expected to be a part of the GEI from the edition 2024 after a comprehensive EU Gender-based violence survey (EU-GBV) is completed. Table 1 shows 31 indicators that make the GEI framework, divided into six domains and 14 sub-domains. The raw indicator data has been collected from the Eurostat, Gender Statistics Database, and EIGE.

Tab. 1: Domains, sub-domains, indicators and their codes used for determining countries' level of gender equality

Domains	Sub-domains	Indicators
Work (A)	Participation in work (A1)	FTE employment (A1.1)
		Duration of working life (A1.2)
	Segregation and quality of work (A2)	Sectoral segregation (A2.1)
		Flexibility of working time (A2.2)
		Career prospects index (A2.3)
Money (B)	Financial resources (B1)	Earnings (B1.1)
		Income (B1.2)
	Economic situation (B2)	Not at-risk-of-poverty (B2.1)
		Income distribution (B2.2)
Knowledge (C)	Attainment and segregation (C1)	Tertiary (C1.1) education
		People employed in education, human health and social work activities (C1.2)
	Segregation (C2)	Tertiary students in the fields of education, health and welfare, humanities and arts (C2)
Time (D)	Care activities (D1)	Childcare activities (D1.1)
		Domestic activities (D1.2)
	Social activities (D2)	Sport, culture and leisure activities (D2.1)
		Volunteering and charitable activities (D2.2)
Power (E)	Political (E1)	Ministerial representation (E1.1)
		Parliamentary representation (E1.2)
		Regional assemblies representation (E1.3)
	Economic (E2)	Members of boards (E2.1)
		Members of Central Bank (E2.2)
	Social (E3)	Share of board members of research funding organizations (E3.1)
		Share of board members in publicly owned broadcasting organizations (E3.2)
		Share of members of highest decision-making body of the national Olympic sport organizations (E3.3)
Health (F)	Status (F1)	Self-perceived health (F1.1)
		Life expectancy (F1.2)
		Healthy life years (F1.3)
	Behaviour (F2)	Non-smoking and non-drinking (F2.1)

		Doing physical activities (F2.2)
	Access (F3)	Unmet medical needs (F3.1)
		Unmet dental needs (F3.2)

Source: EIGE (2022)

EIGE states that women's and men’s participation in paid work is key to paving the way for further progress in gender equality (EIGE, 2022). Therefore, the first domain *Work* aims to measure women’s *Participation* (A1) in the labour market and *Segregation and quality of work* (A2) that women encounter. The aspect of participation is measured through FTE employment and duration of working life, while the segregation and quality of work have been measured with three indicators. One of them is sectoral segregation, i.e. the proneness of men and women to work in different occupational fields. Sectoral segregation has been widely marked as a source of the gender pay gap (Bergmann et al., 2019).

A society striving to achieve gender equality should be based on the principle of both men and women being paid equally for their work (Plantenga et al., 2013). At the same time, personal earnings allow women to be financially independent, automatically granting them equal rights (Hendriks, 2019). For this reason, financial indicators were included in the GEI framework. *Money* domain covers the difference in earnings (*Financial resources,* B1*)* and earning allocation (*Economic situation,* B2*).*

The third domain, *Knowledge,* examines the gaps between women and men regarding educational attainment and training. Today, women tend to reach or even exceed men’s educational attainment. Such changes oppose the traditional gender ideology seen by older generations (EIGE, 2022). This is why, even today, access to education for women is difficult in some more traditional societies. Without proper education, a woman’s probability of getting into the labour market decreases, leading her to economic dependency on men, especially during COVID and post-COVID pandemic (Reichelt et al., 2021). In a way, this puts *Knowledge* at the core of gender inequality.

Including *Time* in a framework for measuring gender equality is a major step forward. This domain tries to break the typical division of activities into productive and reproductive ones, which is the core of gender inequality (Crompton, 2006). Although women are more present in the labour market than in the past decades, the responsibility and burden of managing the household is still on them (Aassve et al., 2014). The idea of equal sharing of time refers to the time men and women spend on household and family activities (*Care activities,* D1*)* and leisure and community work (*Social Activities,* D2).

Domain *Power* focuses on the gap between women's and men's level of representation in the decision-making positions in the political, economic, and social spheres (EIGE, 2022). This domain complies with previous gender inequality indices such as GEM. In addition, it measures the EU's development goal of achieving a balanced participation of men and women in decision-making processes (Plantenga et al., 2009). Despite the increase in female representatives in decision-making bodies, the struggle for equity in this sphere exists (Celis and Lovenduski, 2018). The *Power* domain is seen through *Political* (E1)*, Economic* (E2)*,* and *Social* (E3) representation in public and private institutions.

The last domain, *Health,* measures the impact of gender on one's health. All individuals should have the same access to public and private goods and services (EIGE, 2022), but access to women can be difficult. Weaker labour market attachment, lower socio-economic position, and lesser participation in the public sphere are reasons for such occurrences (Heise et al., 2019). Besides the health *Status* (F1)*,* which measures self-perceived health and life expectancy, this domain measures *Access* (F3) to basic medical and dental needs, as well as healthy life *Behaviour* (F2).

Two important aspects of the GEI should be elaborated upon: the weights assigned and the aggregation method. Regarding weights assigned to framework indicators, sub-domains, and domains, they depend on the framework level. Namely, weights are equal (indicator and sub-domain level) or based on experts' opinions (domain level). On the other hand, the aggregation method is arithmetic (indicators to sub-domains) or geometric mean (sub-domains to domains and domains to overall GEI). Namely, indicators are aggregated into sub-domains by arithmetic mean and equal weights, while sub-domains are aggregated into domains by geometric mean and equal weights, and finally, the overall result of GEI is computed as the geometric mean and experts' weights obtained by Analytic Hierarchy Process (AHP) (Papadimitriou et al., 2020). Table 2 shows the GEI aggregation method and weights assigned to the indicators, sub-domains, and domains.

Tab. 2: Aggregation method and the weights assigned to the indicators, sub-domains, and domains

	Indicator level	Sub-domain level	Domain level
Weighting	Equal	Equal	AHP
Aggregation	Arithmetic	Geometric	Geometric

Source: Papadimitriou et al. (2020)

There are no specific information in any official EIGE report on the choice of the aggregation methods used and weighting schemes chosen. The 2017 report states that 3,636 formulas were considered and therefore, 3,636 indices were computed (EIGE, 2017). They mention considering four methods for assigning weights (equal weights, a modified version of equal weights, weights retrieved from statistical analysis, and finally, weights derived from expert opinions) and two aggregation methods (arithmetic and geometric). However, for the purposes of index scrutinization, it would have been better if more information was provided by the index creators.

The performance ranking of member states based on their level of gender equality was first announced biannually but is now announced annually. Our research is based on the data retrieved from the official GEI 2022 data set, depicting the situation in the year 2020. Within the research, we have applied a twofold I-distance approach, defined and calculated the *Composite I-distance Indicator* (CIDI) methodology and performed post hoc analysis. Our analysis is believed to provide additional confirmation that the GEI is a stable, coherent, and reliable metric or will point out future directions of GEI alterations.

3. METHODOLOGY

In this section, we will outline the fundamentals of the I-distance method (Section 3.1), the related Composite I-distance indicator (CIDI) methodology (Section 3.2), as well as the post hoc I-distance approach (Section 3.3).

3.1 I-DISTANCE METHOD

Composite indices have widely been criticized for their subjectivity in the indicators' selection process, weighting system and later aggregation method (Booysen, 2002; Greco et al., 2019). The weighting of indicators plays a major role in the development of a composite index, and as such, it raises uncertainty and debate along the process (Tarantola and Saltelli, 2007; Becker et al., 2017). For that, additional attention should be given to this process when creating a composite index. Some weights can be based on statistical methods, while others might depend on expert opinion to denote the policy priorities better and/or theoretical factors (OECD, 2005). Such as weighting methods, the aggregation methods also vary. Namely, the linear method is preferable when indicators have a measurement unit, while the geometric is more appropriate when no compensability between indicators should be allowed (Munda, 2008). Whichever weighting or aggregation method is used, an adequate and unbiased one cannot be easily resolved (Saisana et al., 2005).

In the 1960's a need emerged for a composite index that will rank countries based on their socio-economic development. A new index should have been created which would be able to use various indicators, to maximize the amount of information gathered, and most importantly, it had to be unbiased. Ivanovic (1977) devised a statistical method, the I-distance method, that was capable of answering all the requirements. Namely, the I-distance method is based on calculating the mutual distances between the entities being processed, whereupon they are compared to one another to create a rank (Jeremic et al., 2011). In order to calculate the distance and rank countries, it is necessary to fix one entity as a reference in the observed set using the I-distance method. The ranking of entities in the set is based on the calculated distance from the referent entity (Maricic et al., 2019).

The referent entity can be an actual or a fictive entity and the choice on that is made by the analyst. The referent entity can be a particular entity (for example Italy) or an entity which has the minimal measured values of each indicator or an entity which has the maximum measured values or even an entity which has values predefined by the analyst. Based on the choice of the referent entity, the obtained values are interpreted. If the referent entity is a particular benchmark country, the values obtained provide information on whether other entities perform better or worse than it. In practice, so far, most commonly, the referent entity was the entity which had the minimal value (Jeremic et al., 2011). The values of the obtained I-distance then provide information on how far away an entity is from the worst-case scenario. In the performed analysis, the authors used as the referent a fictive entity with minimal measured values of each indicator.

For a selected number of variables (indicators), denoted with *k*, $X^T = \left(X_1, X_2, \ldots X_k\right)$, chosen to characterize the entities, the I-distance between the entity $e_r = \left(x_{1r}, x_{2r}, \ldots x_{kr}\right)$ and the fictive entity $e_s = \left(x_{1s}, x_{2s}, \ldots x_{ks}\right)$ is defined as:

$$D(r,s) = \sum_{i=1}^{k} \frac{\left|d_i\left(r,s\right)\right|}{\sigma_i} \prod_{j=1}^{i-1} \left(1 - r_{ji.12\ldots j-1}\right) \tag{1}$$

where $e_r = \left(x_{1r}, x_{2r}, \ldots x_{kr}\right)$ and $e_f = \left(x_{1f}, x_{2f}, \ldots, x_{kf}\right)$ are values of indicators *I*, $I = \left\{1, \ldots, k\right\}, i \in I$ of the observed entity e_r and fictive entity e_f ;

$d_i(r,s)$ is the distance between the values of the indicator X_i for entities e_r and e_s e.g. the discriminate effect:

$$d_i\left(r,s\right) = x_{ir} - x_{is} \quad i \in \left\{1, \ldots k\right\} \tag{2}$$

σ_i is the standard deviation of indicator *i,* $i \in I$ and

$r_{ji.12\ldots j-1}$ is a partial coefficient of the correlation between indicators *i* and *j* where *j<i,* $i \in I$, $j \in I$, while the effects of all other indicators 1,2,…, j-1 are eliminated (Jeremic et al., 2011; Maricic and Jeremic, 2023). Partial coefficient of correlation describes the strength of a linear relationship between two variables, holding constant a number of other variables. The partial coefficient of correlation eliminated the effect of the other confounding variable(s) that is numerically related to both variables of interest (Baba et al., 2004).

The calculation of the I-distance is an iterative process, consisting of several steps. First, the value of the discriminate effect of the first variable (the most valuable variable, which provides the largest amount of information on the phenomena upon which the entities will be ranked) is calculated. Then, the value of the discriminate effect of the second variable that is not covered by the first one is calculated. This procedure is repeated for the all observed variables in the data set (Radojicic et al., 2019).

To overcome the problem of negative coefficient of partial correlation, which can occur when it is not possible to achieve the same direction of variables, it is suitable to use the square I-distance (Maricic and Kostic-Stankovic, 2016; Maricic and Jeremic, 2023). It is given as:

$$D^2(r,s)=\sum_{i=1}^{k}\frac{d_i^2(r,s)}{\sigma_i^2}\prod_{j=1}^{i-1}\left(1-r_{ji.12\ldots j-1}^2\right). \tag{3}$$

Instead of using the standard deviation of indicator i (σ_i) and partial coefficient of the correlation ($r_{ji.12\ldots j-1}$) between indicators i and j, the square I-distance uses the variance of indicator i (σ_i^2) and coefficient of partial determination ($r_{ji.12\ldots j-1}^2$) between indicators i and j, where $j<i$.

Square I-distance can be used even if the sign of the coefficients of correlation is positive. Also, when there is a large number of variables used the application of the square I-distance is recommended. The order of variables by which they are entered in the I-distance is of high importance. The first entered variable is the variable which is the most correlated with the rest (Jeremic et al., 2011). It is expected that this variable has the largest explanatory effect. The other variables are entered in the algorithm following the same procedure. When there is a large number of indicators, it can happen that the information carried by indicators which enter last in the algorithm is consumed by the indicators which entered the algorithm prior. Therefore, applying the square I-distance method is advisable as it can minimize the amount of lost information by using the partial coefficient of determination. As the presented framework has 14 sub-domains, the square method was used.

3.2 COMPOSITE I-DISTANCE INDICATOR METHODOLOGY

Besides providing rankings of entities, the I-distance method can create a more stable ranking methodology by modifying its official weights. The process of assigning adequate weights is referred to as the *Composite I-distance Indicator* (CIDI) methodology. In order to obtain weights which are not subjectively assigned, first, the correlation coefficients of each entity or domain with the I-distance value are calculated. Correlations are used as I-distance provides information on how valuable each domain is (Jeremic et al., 2011). The next step in the proposed methodology is calculating the new weights for each compounding domain based on the appropriate correlations. Weights are formed by dividing the values of correlations by the sum of correlations. The final sum of weights equals 1, thus forming a novel appropriate weighting system. The equation for determining weights is:

$$w_i = r_i / \sum_{j=1}^{k} r_j \tag{4}$$

where $r_i (i = 1,...,k)$ is the Pearson correlation coefficient of the i-th input indicator with the I-distance values (Dobrota et al., 2015).

3.3 POST HOC I-DISTANCE APPROACH

Besides providing a ranking list of entities, I-distance can be used for an in-depth analysis of the rank consistency. Namely, it can act as a post hoc approach. The post hoc approach is conducted in the following way. Using all *k* initially chosen indicators, the I-distance is calculated and the importance of the indicators for the ranking process is observed by calculating the Pearson's correlation coefficient between the indicator values and the obtained I-distance value. After each iteration, an indicator whose correlation coefficient with the I-distance value is the lowest is excluded from further analysis in the second iteration (Markovic et al., 2015; Savic et al., 2016). So, after each iteration, the number of indicators used to rank the entities is reduced. In the next iteration, the new I-distance rank is formed, and the procedure is repeated.

The I-distance post hoc approach is an iterative process, so the question that arises is when to stop excluding indicators from the framework. In the study by Markovic et al. (2015), further iterations were stopped when the sum of correlation coefficients started to plummet. However, another case can also appear when the sum of correlation coefficients increases throughout the iterative process. In that specific situation, the procedure stops when two indicators are left.

Using the post hoc I-distance method, it is not only possible to obtain information on the importance of indicators for the ranking process but also to get an insight on how the ranking of entities is sensitive to indicator exclusion. Ranking sensitivity of entities provides additional information on the contribution and consequences of including or excluding an individual indicator. Therefore, the post hoc I-distance method can be used to assess the composite indicator structure and suggest its simplification.

3.4 COMPARISON OF THE I-DISTANCE APPROACH WITH OTHER APPROACHES

This research focuses on the data-driven methodologies and one particular method, the I-distance method. Nevertheless, one can ask why the I-distance was chosen in this paper among many other available methods? Therefore, in the following paragraphs, we compare the I-distance method to several other data-driven and non-participatory methods used in the composite indicator literature.

Pea's Method

The Pena's method was devised in the same period as the I-distance method in the 1970s by Peña (1977). Interestingly, the first usage was as well in the field of quality of life composite indicators. The P2 distance or DP2 method functions similarly to the I-distance method: it calculates the distance an entity in relation to an object. This method is said to solve several issues, such as: aggregation of variables expressed in different measures, arbitrary weights and duplicity of information (Somarriba and Peña 2009). The formula for the Pena distance for a chosen entity *s* is:

$$DP2_s = \sum_{i=1}^{k} \left(\frac{\left| x_{is} - x_{ir} \right|}{\sigma_i^2} \left(1 - r_{i,i-1,\dots,1}^2 \right) \right). \tag{5}$$

The initial part of the Pena distance is the same as the initial part of the regular I-distance method – calculating the discriminant effect and taking the variability of the indicator into account. The main difference occurs in the weighting part. Pena distance considers partial coefficients of determination, while the I-distance method considers partial coefficients of correlation and coefficients of correlation. Both methods depend on the order of variables in the algorithm (Montero et al., 2010; Maricic et al., 2016).

Data Envelopment Analysis (DEA) and DEA-Like Approaches

DEA is an optimization method devised by Charnes et al. (1978), used to calculate the relative efficiency of decision-making units (DMUs) based on the measured values of inputs and outputs. The DEA method looks for data-driven weights which will maximize the overall score of the DMUs. Because the importance of the inputs and outputs does not depend on the analyst's or the experts' opinion, the application of the DEA method quickly increased, especially in policy-related settings (Cherchye et al., 2008). In the field of composite indicator creation, a special type of DEA method has been widely employed: the Benefit-of-the-Doubt (BoD) model. The BoD model is, in fact, an input-oriented DEA model (Melyn and Moesen, 1991). The goal function of the model is to maximize the value of the composite indicator by changing the weights assigned to individual indicators. The main issue with both DEA and BoD models is full freedom (Rogge and Van Nijverseel, 2019). Namely, if no weight constraints are imposed, all entities will achieve the maximum value of the composite indicator. Therefore, different approaches to weight restriction have been proposed: The upper and lower bounds of weight were generated via participatory methods, Intervals around the government-defined weights, symmetric interval ± 25% around CIDI weights, and others (Maricic and Jeremic, 2023). Although this also is a data-driven weighting approach, it significantly differs from the I-distance method. DEA and BoD models are optimization models, while the I-distance is a distance-based mathod.

Displaced Ideal Method (DI)
Displaced Ideal Method (DI) is a method based on the Euclidean distance proposed by (Zelany, 1974). The idea of the DI is to show the smallest distance of an entity from its ideal scenario. The Euclidean distance, as a distance metric, is not robust over a range of scales, which means that the computed results can be skewed if the units of the variables used have very different variabilities (Saranya and Manikandan, 2013). Similar as the I-distance method, the DI method posits that there is an inherent connection among the representative variables of a phenomenon being studied. On the other hand, the difference occurs when we observe from which the distance is observed. In the I-distance, the distance is observed between the worst case scenario, while in the ID method, iIt suggests that the optimal system should strive to minimize the gap between its current state and the ideal scenario (Magalhães-Timotio et al., 2022).

Partial Least Squares-Path Modeling
Structural equation modelling (SEM) is a statistical multivariate analysis which lies on the principles of factor analysis and regression analysis (Kline, 2005). Therefore, the analysis allows for grouping of individual indicators and exploration the relationship between the newly formed latent variables. Both

these features are quite valuable in the process of composite indicator creation as they allow for considering the role (formative and reflective) of the manifest variables (MVs) (Lauro et al., 2018). Two main approaches within the SEM literature are the covariance-based (CB-SEM) and partial least squares (PLS-SEM). The first is seen as the parametic SEM, while the second is observed as the non-parametric SEM. Within the composite indicator literature, the PSL-SEM approach is more common due to the fact that the goal of a composite indicator is to estimate the latent variables, and PLS-SEM does just that (Trinchera et al., 2008). Using SEM algorithms to create composite indicators considers taking a model based approach which creates a multidimensional latent variable measurable directly and related to its single indicators or MVs by a reflective or formative relationship or by both (Lauro et al., 2018). Although both SEM and I-distance approaches are data-driven, the first one is model based, while the other is distance-based.

The presented lietarure review indicates that within the field of composite indicator creation and evaluation there is a plethora of statistical analysis and methods of operational research which can be employed. Neither approach is flawless (Greco et al. 2018). Therefore, it is suggested that the composite indicator creator considers several approaches in quest of determing the final composite indicator methodology. The I-distance method applied in this paper is just one of the possible solutions.

4. RESULTS

This section presents the results of scrutinizing the GEI using the I-distance method and the related CIDI and Post hoc approaches. The results are organized into three subsections for better paper flow and presentation.

4.1 APPLICATION OF THE TWO-FOLD I-DISTANCE APPROACH TO THE GEI

The first direction in our research implied calculating the Total I-distance values for the GEI. Within the aim of the study to scrutinize the GEI framework, we applied the twofold I-distance approach to all EU member states and compared their rankings to the official GEI rankings. The first step in the analysis is applying the I-distance method on the sub-domain values to create new I-distance domain values. The second step sees the application of the I-distance method on the previously obtained six I-distance domain values.

As presented, the analysis used sub-domain data for one reason: data availability. The data scores available were scores of the sub-domains, domains and overall GEI. The data for indicators is available, but separately for males and females. Although the data on the indicator level is available, as there is no clear and straightforward information on how the final scores are calculated, we decided to focus on the sub-domain data. Therefore, we assumed that on the level of aggregation from indicators to sub-domains, there should be no change in weights and that equal weighting is appropriate.

Tab. 3: The rankings of countries by *Work, Money, Knowledge, Time, Power* and *Health* domains after applying the I-distance method

Member state	Work rank	Money rank	Knowledge rank	Time rank	Power rank	Health rank
Sweden	**1**	11	**1**	**1**	2	2
The Netherlands	3	10	5	3	5	3
Denmark	2	6	4	2	7	13
Belgium	14	2	2	13	6	12
Luxembourg	10	**1**	3	8	12	5
Ireland	12	7	7	7	10	**1**
Finland	5	4	11	6	4	9
Spain	18	21	6	10	3	7
France	16	12	9	11	**1**	14
Austria	8	8	12	16	14	4
Malta	9	14	8	14	19	6
Slovenia	13	3	22	9	16	16
Slovakia	21	5	10	25	24	20
Germany	15	17	25	12	8	8
Estonia	7	23	19	5	22	22
Latvia	6	26	27	4	15	27
Italy	27	18	13	17	11	10
Czech Republic	20	9	14	21	23	18
Cyprus	17	13	17	19	26	11
Lithuania	4	25	18	22	18	21
Bulgaria	19	27	15	26	9	23
Hungary	23	16	20	18	27	17
Portugal	11	19	21	23	13	24
Poland	24	15	16	20	20	25
Greece	26	22	23	27	25	15
Croatia	22	20	26	24	17	19
Romania	25	24	24	15	21	26

Source: Authors’ own work

The rankings within each domain after applying the I-distance method are presented in Table 3.

The I-distance domain results provide interesting results. Namely, two Scandinavian member states (Sweden and Denmark) are in the top 5 for three domains, which leads to the conclusion that these countries are committed to a multidimensional approach to reducing gender inequality. The domains in which the results of the “Scandinavian duo” are not the leading ones are *Money, Power* and *Health* domains. In the case of the *Money* domain, Luxembourg tops the list, followed by Belgium. Looking at the results of *Power,* France and Sweden lead the way, closely followed by Spain and Finland. On the other hand, the results of the *Health* domain pointed out Ireland as a country where both men and women have the same treatment and access to basic medical care.

Finally, the results of Romania and Croatia should be pointed out. The results of these countries are in the bottom 5 for three domains. These findings do not mean these countries are not trying to create a gender-equal society, just that there are discrepancies between them and other member states.

The final step of the two-fold I-distance approach saw the appliance of the I-distance on previously obtained domains. Table 4 shows the results of the two-fold approach: the Total I-distance value, Total I-distance ranks, and official GEI ranks.

Tab. 4: The results of the I-distance method, Total I-distance value, Total I-distance ranks and official GEI ranks of EU member states

EU Member state	Total I-distance value	Total I-distance rank	Official rank
Sweden	45.718	1	1
The Netherlands	22.157	2	3
Denmark	21.722	3	2
Belgium	19.420	4	8
Luxembourg	19.393	5	9
Ireland	19.375	6	7
Finland	17.396	7	4
Spain	14.755	8	6
France	14.300	9	5
Austria	11.298	10	10
Malta	10.300	11	13
Slovenia	8.570	12	12

Slovakia	7.387	13	24
Denmark	6.872	14	11
Spain	6.266	15	17
Latvia	6.236	16	16
Italy	6.000	17	14
Czech Republic	5.249	18	23
Cyprus	5.122	19	22
Lithuania	2.962	20	20
Bulgaria	2.697	21	18
Hungary	2.684	22	25
Portugal	2.619	23	15
Poland	2.294	24	21
Greece	1.519	25	27
Croatia	1.452	26	19
Romania	0.540	27	26

Source: Authors' own work

Consequently, Sweden and the Netherlands top the list. The obtained value of Sweden, 45.718, indicates that Sweden is furthest away in the multidimensional space from the fictive entity with minimal values of all six domains. On the other side of the ranking, Romania is quite close to the fictive entity, with a Total I-distance value of just 0.540.

After applying the two-fold I-distance method, out of 27 countries, 11 have improved their rank; five did not change their rank, and 11 dropped their rank. The I-distance results show countries that lead the way in terms of creating a more gender-equal society, the ones that have visibly improved their attitude towards gender equality, the ones that were expected to be more gender equal, and the ones that still have a way to go when implementing this concept.

Scandinavian countries have a rich history of gender equality that originates from the last decades of the 19th century. More than a hundred years later, these countries are still leading in terms of women's rights in education, voting, and political representation. One of the reasons for this lies in the famous Nordic welfare system that stands out for its universalism and its devotion to creating a society that gives women equal rights in all spheres of life (Borchorst and Siim, 2008). In several central European countries, the gender equality legislation drastically changed after they joined the EU. The EU's legal system and the potent guidance of the EU were just what these countries needed to trace their path towards raising gender equality. This can be especially noted in the case of Hungary and the Czech Republic (Velluti, 2014).

After having applied the two-fold approach, the list of countries in the top 10 did not change. Countries which improved their rank are Luxembourg and Belgium, while Denmark, France, Spain, and Finland dropped ranks. This means some aspects of gender equality need more attention and that there is place for further legislation improvements.

Younger EU member states have quite a distinctive work and family policy history compared to the EU's. In these countries, there is a strong gender – tradition-based ideology that denotes men as breadwinners and women as housewives (Hofacker et al., 2013). These discrepancies mean there are essential political, economic, and social changes related to gender equality ahead of them (Witkowska, 2013).

4.2 ASSIGNING WEIGHTS BY APPLYING THE I-DISTANCE METHOD

The second direction of our research was to scrutinize the GEI weighting scheme on the domain level. To get an in-depth analysis of the GEI, besides applying the two-fold I-distance approach, we used the CIDI methodology. Newly obtained domain weights by CIDI are calculated by dividing the correlations of domains to the Total I-distance value and the sum of domains' correlations to the Total I-distance (Dmitrovic et al., 2016). To perform the analysis, we calculated the correlations between Total I-distance values and each of the I-distance domain values

The comparison of the GEI weights and the weights proposed by CIDI is presented in Table 5. The largest differences are with the domains *Power* and *Time*. *Power* is weighted at 19% according to the official GEI ranking and 15.9% according to the CIDI methodology. As most member states have established minimum quotas for female representation in *Political* and *Economic* sphere, this domain does not need such high importance (Mateos de Cabo et al., 2011). Similarly, *Knowledge* dropped importance by our method to 20.3% from 22%. On the other hand, *Time* is assigned 15% weight according to the AHP experts' opinions, while our method gives it a higher significance, 17.3%. This domain deserves more attention, as women are facing constraints on their leisure time both within and outside home (Aitchison, 2013).

Tab. 5: Differences in CIDI weights and original GEI domain weights.

Domain	Correlation Coefficient	CIDI weights	Official GEI weights	Change from official GEI weight
Knowledge	0.897	20.3%	22%	-1.7%
Work	0.788	17.8%	19%	-1.2%
Time	0.768	17.3%	15%	2.3%
Money	0.725	16.4%	15%	1.4%
Power	0.705	15.9%	19%	-3.1%
Health	0.546	12.3%	10%	2.3%

Source: Authors' own work

The results of the CIDI methodology indicate that changes to the domain could be implemented. The smallest suggested change is for the domain *Work* and its weight should be decreased for 1.2% points, followed by the domain *Money,* whose weight should be increased by 1.4% points. The greatest change suggested is for the domain *Power,* whose importance could be reduced from 19% to 15.9%. The detected differences indicate that even though the GEI weights at the domain level are expert-driven, they are close to data-driven weights and that for only one domain, substantial changes are advised.

4.3 APPLICATION OF THE POST HOC I-DISTANCE APPROACH

Following the idea of Markovic and associates (2015), we applied the post hoc I-distance approach. Accordingly, the third phase of our study refers to 13 I-distance iterations for 14 GEI sub-domains. Namely, the iterative process was stopped when the highest average coefficient of correlation was obtained, in this case, 13 iterations. In the last iteration, we stopped excluding indicators because there were only two sub-domains left – Financial resources (B1) and Attainment and segregation (C1). Further iteration would have shown Financial resources (B1) as the most important sub-domain. Table 6 shows the ranks after the first two iterations, the ranks in the last iteration, the change in rank after the first and the last iteration, the median and the interquartile range (IQR) for all EU member states after the application of the post hoc I-distance approach.

Tab. 6: I-distance iterations, change in rank between the first and the last iteration, median, and interquartile range

Country	1st iteration	2nd iteration	…	13th iteration	Change in rank	Median	IQR
Sweden	1	1	…	6	-5	1	1
The Netherlands	2	2	…	2	0	2	0
Denmark	3	3	…	3	0	3	2
Ireland	4	5	…	5	-1	5	0
Finland	5	4	…	4	1	4	2
Luxembourg	6	6	…	1	5	4	3
Belgium	7	7	…	7	0	7	2
Spain	8	8	…	11	-3	8	2
Austria	9	10	…	8	1	9	2
France	10	9	…	10	0	9	2
Malta	11	12	…	12	-1	12	2
Slovenia	12	11	…	15	-3	13	1
Germany	13	13	…	9	4	11	1
Italy	14	15	…	14	0	15	2
Cyprus	15	18	…	13	2	17	3
Czech Republic	16	17	…	18	-2	20	3
Estonia	17	14	…	16	1	14	2
Slovakia	18	16	…	26	-8	26	5
Portugal	19	20	…	21	-2	18	3
Hungary	20	22	…	22	-2	22	4
Lithuania	21	21	…	17	4	17	5
Bulgaria	22	24	…	27	-5	25	4
Latvia	23	19	…	24	-1	23	7
Croatia	24	25	…	23	1	22	2
Poland	25	23	…	19	6	24	2
Greece	26	26	…	20	6	21	6
Romania	27	27	…	25	2	27	0

Source: Authors' own work

Indicator *Change in rank* shows an interesting result: three countries held the same rank throughout the iterations, and three countries moved up for just one place. The highest oscillations, as a result of sub-domains reduction, occurred at the bottom of the ranks, to Poland and Greece. Namely, these countries improved their ranks by six places respectively. On the other hand, the largest decrease in ranks occurred in the case of Slovakia, which dropped eight ranking places. These three countries proved to be the most sensitive to excluding indicators. A detailed frequency analysis of the change in rank is presented in Table 7. As can be seen, the majority of Member states, as much as five of them, have not changed their rank after significant indicator removal. This can be an indication that the subdomain list can be reduced.

Tab. 7: Frequency analysis of the change in rank of Member states

Change in rank	Frequency	Member states
-8	1	Slovakia
-5	2	Sweden, Bulgaria
-3	2	Slovenia, Spain
-2	3	Hungary, Portugal, Czech Republic,
-1	3	Latvia, Malta, Ireland
0	5	The Netherlands, Denmark, Belgium, France, Italy
1	4	Croatia, Estonia, Austria, Finland
2	2	Romania, Cyprus
4	2	Lithuania, Germany
5	1	Luxembourg
6	2	Greece, Poland

Source: Authors' own work

It is valuable to observe the rank median and compare it with the rank in the last iteration. For example, Sweden has a median 1, while its 13th rank is six. This indicates that Sweden dropped rank almost at the end of the analysis. On the other hand, Luxembourg has a median 4, while its 13th rank is 1, which indicates that Luxembourg improved rank almost at the end of the analysis. For the majority of member states, the median and the 13th rank are the same or close, indicating mostly stable ranks throughout the analysis.

IQR can also be used to detect the member states whose ranks oscillated. For example, the IQRs of Latvia and Greece are seven and six, respectively, indicating sharp rank changes. On the other hand, the IQR of Romania is 0, showing that the ranks of Romania have been stable.

As the final part of the descriptive analysis of the measured changes in rank, we provide a scatterplot between the rank in iteration 1 and iteration 13. The countries below the regression line are the countries whose rank improved, while the countries above are those whose rank decreased.

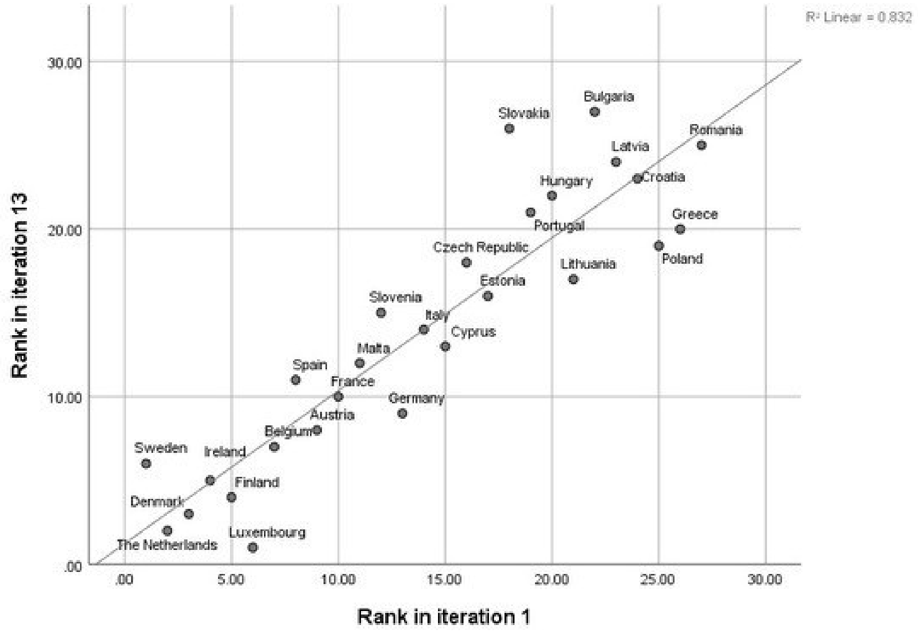

Fig. 1: Scatterplot between rank in iteration 1 and 13

Besides solely analyzing country's ranks throughout the iterations, we observed the ranks of indicators by their correlation coefficient with the Total I-distance value and the order by which they were ruled out from further observation. Table 8 presents the rank of indicators, their coefficient of correlation and the average coefficient of correlation after each iteration. As one can see, after the first iteration, the change of the average coefficient of correlation was for +0.024. With every next iteration, the quality of our model improved, as the average correlation grew from 0.640 in the first iteration to 0.891 in the last. Another interesting fact is that the importance of the sub-domains changed. In the first iterations, the sub-domain *Social activities* (D2) was top-ranked, while at the end of the analysis, the sub-domain *Financial resources* (B1) came to be the most important. The order of domains by which they were ruled out of the framework is *Power-Health-Time-Work-Knowledge-Money.*

Tab. 8: Review of excluding indicators, their coefficient of correlation and the average coefficient of correlation after each iteration

1st iteration	r	2nd iteration	r	…	12th iteration	r	13th iteration	r
D2	0.877	D2	0.884	…	B1	0.940	B1	0.941
A2	0.826	C1	0.822	…	C1	0.851	C1	0.842
C1	0.815	A2	0.803	…	A2	0.833		
B1	0.780	F2	0.760	…				
F2	0.761	B1	0.747	…				
E3	0.736	E3	0.742	…				
E1	0.716	E1	0.733	…				
F1	0.632	D1	0.674	…				
D1	0.618	E2	0.622	…				
E2	0.614	F1	0.577					
C2	0.545	C2	0.510					
B2	0.362	A1	0.399					
A1	0.348	B2	0.358					
F3	0.335							
Average r	0.640		0.664	…		0.875		0.891

In order to get a glimpse of the oscillations in I-distance ranks, Figure 2 gives a graphic overview of how the rank of the top 5 countries (according to the Total I-distance) changed during the analysis. Firstly, Sweden and the Netherlands swapped places after nine iterations. The rank of Sweden continued to decline in the next iterations, while the Netherlands remained in the top ranks. The rank of Denmark was stable until iteration 7, when it started to decrease, up to rank 6 (iteration 10). After that, it improved its rank and finished third best. The remaining two countries swapped after the first two iterations, after which Finland sharply decreased to rank 8. During the next iterations, the rank of Finland visibly changed, leading to a final improvement to rank 4. Similar movement could be detected for Ireland.

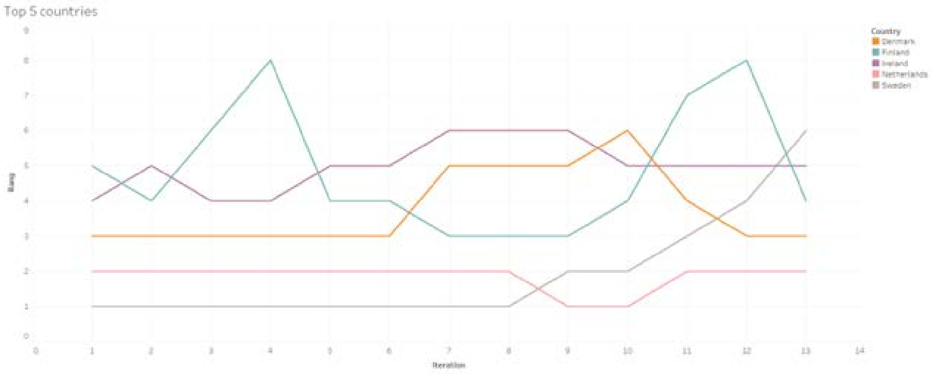

Fig. 2: Overview of the oscillations in the ranks of the top five countries according to the total I-distance value

Knowing that the bottom-ranked countries are more prone to rank oscillations, Figure 3 gives a graphic overview of how the rank of the bottom 5 countries (according to the Total I-distance) changed during the analysis. Romania had a stable rank up until the 10th iteration, when it improved its rank to 25th place on which it remained. The volatility of the remaining four observed countries is greater. Greece and Latvia had visible rank increase and decrease depending on the domains in the composite index framework. Poland and Croatia also have oscillations, but they are not that drastic.

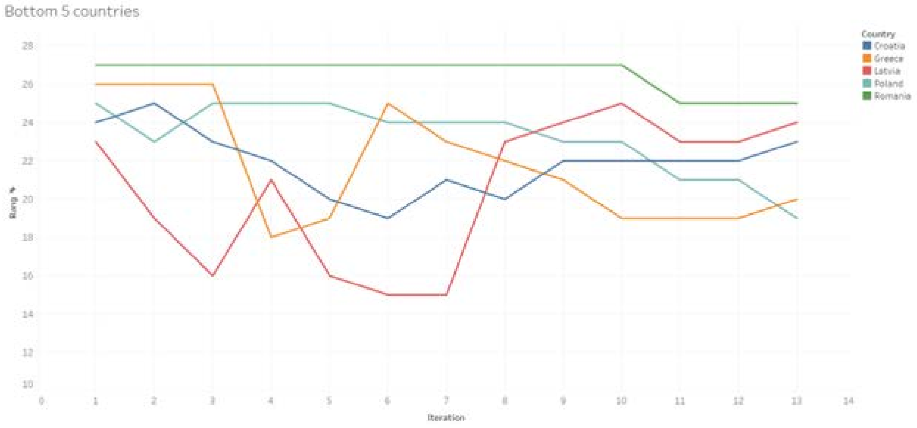

Fig. 3: Overview of the oscillations in the ranks of the bottom five countries according to the total I-distance value

4.4 FUTURE DIRECTIONS OF THE STUDY AND STUDY LIMITATIONS

The three performed analyses based on the I-distance are just one direction of the GEI revision. Namely, other statistical methods could also be applied to inspect whether the GEI is statistically sound and how to enhance it.

One of the possible directions is applying the CIDI methodology to sub-domains. Namely, CIDI can be performed in order to get an evaluation of weights assigned to sub-domains and their importance for the ranking process. Further, sensitivity and uncertainty analysis can be carried out in order to get a complete evaluation of the indicators that make the GEI framework. Also, the results of the post hoc analysis can be applied to reduce the number of framework indicators, which might result in lower calculation costs and even in annual index publications.

Besides further statistical analysis, the number of countries covered is to be expanded. In addition, other countries which are not member states may build a gender equality index modelled on the GEI. Their results can be later compared with the EU average or on a regional level. Serbia, an EU candidate country, is the first to use this opportunity as it has the know-how and expertise to do so (EIGE, 2014). Other candidate countries, like Montenegro, Bosnia and Herzegovina, and others, should also try to follow Serbia on the road for measuring gender equality and trying to create a gender-equal society.

Besides exploring the list of sub-domains and indicators of the GEI as well as the weighting schemes assigned, it would be of interest to inspect the initall structure of the GEI. For that, advanced multivariate analysis such as Hierarchical Disjoint Non-Negative Factor Analysis (Cavicchia et al., 2021), Second-Order Disjoint Exploratory Factor Analysis (Cavicchia and Sarnacchiaro, 2021), or variants of the Benefit-of-the-Doubt method (for example Verbunt and Rogge, 2018).

A limitation of the presented two-fold approach pulls along another direction of future studies. Namely, the I-distance can provide information on the relative performance of the observed countries. This statistical method calculates the distance to a referent entity rather than to a previously defined standard. This comparative analysis draws attention to the fact that there are countries that are doing better than others in terms of achieving gender equality, without elaborating the detailed reasons for such differences. The identification of the determinants of the discrepancies among the observed countries requires the assistance of sociologists, lawyers, and other specialists in the topic. Such multidisciplinary analysis could give answers to policy makers and point them out which legislative segment should be tackled in the future.

Finally, a limitation due to the data availability should be observed. As mentioned in Section 4.1, in this study, the focus was on the sub-domains and domains without deeper analysis of indicator scores and weights. The assumption made in this study was that the equal weighting from indicator scores to sub-domain scores was reasonable. Nevertheless, such a claim can be challenged, as no proof exists that all indicators within one sub-domain are equally important. Therefore, this study focused on the two levels of the GEI, and not on all three. If the indicator scores were publicly available, three-fold I-distance could have been applied, shedding light on the importance of individual indicators.

5. CONCLUDING REMARKS

Out of eight United Nations Millennium Development Goals (MDGs), which were established at the United Nations Millennium Summit in 2000, one is related to gender equality, precisely to "Promote gender equality and empower women" (UN, 2000). Despite the steady progress in the field of gender equality in access to education, labour market and politically influential positions, there is still place for further improvements (Eurostat, 2022). To continuously monitor the progress of achieving gender equality, relevant, accurate and timely gender data is of high importance. The collected data is used to calculate explicit measures of gender (in)equality. The academic community and the international institutions have acknowledged the importance of such metrics, and consequently, several gender equality indices have been created. The idea behind these indices is to draw attention of public and, more importantly, policymakers to the issue of gender-related policies and research (UN, 2012). The index which stands out is the one devised by the European Institute for Gender Equality, the Gender Equality Index (GEI).

In the previous years, the I-distance method was used with success for assessing composite indices of different purposes (Maricic and Kostic-Stankovic, 2016; Markovic et al., 2015; Maricic et al., 2019). What differentiates the I-distance method and related methods from other statistical methods is its objectiveness. Namely, the method applied in this study does not place any weighting factor on its indicators (Jeremic et al., 2014), meaning subjectively assigned weights cannot influence the final ranking. Within the aim of this research to attempt to improve the measuring system of GEI we proposed the I-distance method to be applied, together with Composite I-distance Indicator (CIDI) methodology, and post hoc analysis.

First, using the two-fold I-distance approach for assessing GEI allowed us to identify the leaders, who follows them, but who is very inefficient in that. The results revealed that Scandinavian member states top the list, that some of the founding members (France and Germany) scored below expectation, whereas South European countries are showing low efficiency in applying the gender equality concept. The obtained ranks comply with the research conducted by Earles (2014). The country ranks by the two-fold approach and the official GEI differ in the middle of the rankings, while the ranks of the top and the bottom countries have not changed. These results led to a significant level of correlation between the two ranking methodologies measured by Spearman's correlation coefficient ($r_s=0.880$, $p<0.01$). Both GEI and I-distance ranks point out one: there are considerable differences inside the EU regarding gender equality that are connected with history, tradition, culture, and the welfare of member states (Witkowska, 2013).

In the official framework, the ranks were obtained as weighted geometric mean of domain values, whilst weights were obtained through the analytical hierarchy process (AHP). Although AHP is widely preferred for solving multi-criteria decision-making problems, one of its drawbacks, which might easily be challenged, are the subjectively assigned evaluation measures (Kilincci and Asli Onal, 2011). By proposing the CIDI methodology for measuring member states' performance and gender equality, we attempted to improve GEI's ranking methodology by assigning objective weights to domains. The aim of this analysis was to test the domain weights obtained by AHP. The obtained result differs from the official ones. The biggest difference can be seen in weights assigned to domains *Power* and *Time*. Besides these, the CIDI methodology confirms the latter AHP weights.

The third analysis performed was the post hoc analysis, whose results imply that five member states had the same GEI and 13th iteration rank and that two of the top five countries did not change throughout the analysis. The correlation coefficient between official GEI ranks and the last iteration is significant ($r_s=0.833$, $p<0.01$), meaning that the country's ranks are stable. One should notice that the correlation coefficient between the first ($r_s=0.894$, $p<0.01$) and the last iteration ($r_s=0.833$, $p<0.01$) with the GEI differs for just 0.063. This leads to the conclusion that the number of indicators which comprise the GEI could be refined without significant changes to the member states ranks.

All of the above-mentioned findings provide an in-depth analysis of the GEI domains and weights assigned to them. We hope our study could act as a confirmation of the GEI's methodology and its results, but also as a guidance for possible future slight enhancements of this composite indicator. This research can also be an impetus for innovative research approaches in the field of gender equality evaluation on the EU level, which might eventually impact EU policy and legislation.

Declarations

Ethical Approval

Not applicable

Competing interests

Not applicable

Authors'contributions

The authors equally contributed to the analyses and the manuscript.

Funding

Not applicable

Availability of data and materials

Available on request from the corresponding author

REFERENCES

Amin, E. and Sabermahani, A. (2017). Gender inequality index appropriateness for measuring inequality. *Journal of Evidence-Informed Social Work. 14*: 8-18.

Aassve, A., Fuochi G. and Mencarini, L. (2014). Desperate housework: Relative resources, time availability, economic dependency, and gender ideology across Europe, *Journal of Family Issues. 35*: 1000-1022.

Aitchison, C. C. (2013) *Gender and Leisure: Social and Cultural Perspectives*. London: Routledge. ISBN: 978-0-415-26155-5.

Amici, M. and Stefani, M. L. (2013) *A Gender Equality Index for the Italian Regions* (No. 190). Bank of Italy, Economic Research and International Relations Area. ISSN: 1972-6643.

Appiah, E. N. and McMahon, W. W. (2002). The social outcomes of education and feedbacks on growth in Africa. *Journal of Development Studies.* 38: 27–68.

Baba, K., Shibata, R. and Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics, 46*(4), 657-664.

Becker, W., Saisana, M., Paruolo, P. and Vandecasteele, I. (2017). Weights and importance in composite indicators: Closing the gap. *Ecological Indicators. 80*, 12-22.

Bergmann, N., Scheele, A. and Sorger, C. (2019). Variations of the same? A sectoral analysis of the gender pay gap in Germany and Austria. Gender. *Work & Organization. 26:* 668-687.

Bericat, E. (2012). The European Gender Equality Index: Conceptual and Analytical Issues. *Social Indicators Research.* 108: 1-28.

Birindelli, G., Iannuzzi, A. P. and Savioli, M. (2019). The impact of women leaders on environmental performance: Evidence on gender diversity in banks. *Corporate Social Responsibility and Environmental Management. 26:* 1485-1499

Booysen, F. (2002). An overview and evaluation of composite indices of development. *Social Indicators Research.* 59: 115-151.

Borchorst, A. and Siim, B. (2008). Woman-friendly policies and state feminism: Theorizing Scandinavian gender equality. *Feminist Theory. 9*(2), 207-224.

Cavicchia, C., Sarnacchiaro, P. and Vichi, M. (2021). A composite indicator for the waste management in the EU via hierarchical disjoint non-negative factor analysis. *Socio-Economic Planning Sciences. 73:* 100832.

Cavicchia, C. and Sarnacchiaro, P. (2021). The effects of a new e-tivity on students'performance and satisfaction in an online course. *Statistica Applicata, 33*: 163-175.

Cela, E., Dankelman, I. E. M. and Stern, J. (2014). *Bookshelf: Powerful Synergies: Gender Equality, Economic Development and Environmental Sustainability*. New York: UNDP.

Celis, K. and Lovenduski, J. (2018). Power struggles: Gender equality in political representation. *European Journal of Politics and Gender. 1* (1-2): 149-166.

Cherchye, L., Moesen, W., Rogge, N., Van Puyenbroeck, T., Saisana, M., Saltelli, A., ... and Tarantola, S. (2008). Creating composite indicators with DEA and robustness analysis: The case of the Technology Achievement Index. *Journal of the Operational Research Society. 59*(2): 239-251.

Crompton, R. (2006). *Employment and the Family: TheRreconfiguration of Work and Family Life in Contemporary Societies*. Cambridge: Cambridge University Press.

Dilli, S., Carmichael, S. G. and Rijpma, A. (2019). Introducing the historical gender equality index. *Feminist Economics. 25*(1): 31-57.

Dmitrovic, V., Dobrota, M. and Knezevic, S. (2017). A statistical approach to evaluating bank productivity. *Management: Journal of Sustainable Business and Management Solutions in Emerging Economies. 20* (75)*:* 47-56.

Dobrota, M., Bulajic, M., Bormnann, L. and Jeremic, V. (2015). A new approach to QS university ranking using composite I-distance indicator: Uncertainty and sensitivity analyses. *Journal of the American Society for Information Science and Technology. 67* (1): 200-211.

Earles, K. (2014). Gender equality identity in Europe: The role of the EU', in D. B. MacDonald and M. M. DeCoste (ed.) *Europe in Its Own Eyes, Europe in the Eyes of the Other,* pp. 103-24. Wilfrid Laurier University Press. ISBN: 978-1554588404.

EIGE. (2022). *Gender Equality Index 2022: The COVID-19 pandemic and care. European Institute for Gender Equality*. Italy: EIGE. Available at: https://eige.europa.eu/publications/gender-equality-index-2022-covid-19-pandemic-and-care. Last access: 15/09/2022

EIGE. (2017). Gender equality index 2017: Methodological Report. Available at: https://eige.europa.eu/publications-resources/publications/gender-equality-index-2017-methodological-report. Last access: 26/10/2023

EIGE. (2014). EIGE's Gender Equality Index invigorates Serbia's path to Gender Equality. Available at: http://eige.europa.eu/content/news-article/eiges-gender-equality-index-invigorates-serbias-path-to-gender-equality. Last access: 23/11/2022

Elias, J. (2013). Davos woman to the rescue of global capitalism: Postfeminist politics and competitiveness promotion at the World Economic Forum. *International Political Sociology. 7* (2): 152-169.

EUR-Lex. (2012). *Charter of Fundamental Rights of the European Union*. 2012/C 326/02

Eurostat. (2022). SDG 5 - Gender equality - Achieve gender equality and empower all women and girls. Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=SDG_5_-_Gender_equality. Last access: 10/02/2022

Girón, A., and Kazemikhasragh, A. (2021). Gender equality and economic growth in Asia and Africa: empirical analysis of developing and least developed countries. *Journal of the knowledge economy*. 13: 1433-1443.

Greco, S., Ishizaka, A., Matarazzo, B. and Torrisi, G. (2018). Stochastic multi-attribute acceptability analysis (SMAA): An application to the ranking of Italian regions. *Regional Studies, 52*(4), 585–600.

Greco, S., Ishizaka, A., Tasiou, M. and Torrisi, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research. 141*, 61-94.

Heise, L., Greene, M. E., Opper, N., Stavropoulou, M., Harper, C., Nascimento, M., ... & Gupta, G. R. (2019). Gender inequality and restrictive gender norms: framing the challenges to health. *The Lancet. 393* (10189): 2440-2454.

Hendriks, S. (2019). The role of financial inclusion in driving women's economic empowerment. *Development in Practice. 29* (8)*:* 1029-1038.

Hofacker, D., Stoilova, R. and Riebling, J. R. (2013). The gendered division of paid and unpaid work in different institutional regimes: Comparing West Germany, East Germany and Bulgaria. *European Sociological Review. 29* (2)*:* 192-209.

Ivanovic, B. (2007) *Classification Theory*. Belgrade: Institute for Industrial Economic.

Jeremic, V., Bulajic, M., Martic, M. and Radojicic, Z. (2011). A fresh approach to evaluating the academic ranking of world universities. *Scientometrics. 87*: 587-596.

Jeremic, V., Kostic-Stankovic, M., Markovic, A. and Martic, M. (2014). Towards a framework for evaluating scientific efficiency of world-class universities. *International Journal of Social, Management, Economics and Business Engineering. 8* (2): 590–595.

Kabeer, N. and Natali, L. (2013). Gender equality and economic growth: Is there a win - win?. *IDS Working Papers*. 2013(417): 1-58.

Klasen, S. (1999). *Does Gender Inequality Reduce Growth and Development? Evidence from Cross - Country Regressions*, Policy Research Report on Gender and Development. Washington DC: World Bank

Klasen, S. (2006). UNDP's gender-related measures: Some conceptual problems and possible solutions. *Journal of Human Development. 7* (2): 243–274.

Kline, R. (2005). Principles and Practice of Strucutral Equation Modeling 2nd Ed.

Kilincci, O. and Asli Onal, S. (2011). Fuzzy AHP approach for supplier selection in a washing machine company. *Expert Systems with Applications. 38* (8): 9656-9664.

Lauro, N. C., Grassia, M. G. and Cataldo, R. (2018). Model based composite indicators: New developments in partial least squares-path modeling for the building of different types of composite indicators. *Social Indicators Research. 135*: 421-455.

Leeves, G. D. and Herbert, R. (2014). Gender differences in social capital investment: Theory and evidence. *Economic Modelling. 37*: 377-385.

Magalhães-Timotio, J. G., Barbosa, F. V. and Ferreira, B. P. (2022). Constructing a composite financial inclusion index for Brazil. *Revista Gestão & Tecnologia. 22*(1): 168-192.

Maricic, M. and Jeremic, V. (2023). Imposing unsupervised constraints to the Benefit-of-the-Doubt (BoD) model. *METRON*, Online first: 1-38.

Maricic, M., Egea, J. A. and Jeremic, V. (2019). A hybrid enhanced Scatter Search—Composite I-Distance Indicator (eSS-CIDI) optimization approach for determining weights within composite indicators. *Social Indicators Research. 144:* 497-537.

Maricic, M. and Kostic-Stankovic, M. (2016). Towards an impartial responsible competitiveness index: A twofold multivariate I-distance approach. *Quality & Quantity. 50* (1)*:* 103-120.

Maricic, M., Bulajic, M., Dobrota, M., & Jeremic, V. (2016). Redesigning the global food security index: A multivariate composite I-distance indicator approach. *International Journal of Food and Agricultural Economics (IJFAEC), 4*(1), 69-86.

Markovic, M., Zdravkovic, S., Mitrovic, M. and Radojicic, A. (2016). An iterative multivariate post hoc I-distance approach in evaluating OECD Better Life Index. *Social Indicators Research*. *126*: 1-19.

Mateos de Cabo, R., Gimeno, R. and Nieto, M. J. (2011). Gender diversity on European banks' board of directors. *Journal of Business Ethics. 109*: 145-162.

Melyn, W. and Moesen, W. (1991). *Towards a Synthetic Indicator of Macroeconomic Performance: Unequal Weighting when Limited Information Is Available*. Leuven University, Working paper ID 26175691.

Montero, J. M., Chasco, C. and Larraz, B. (2010). Building an environmental quality index for a big city: A spatial interpolation approach combined with a distance indicator. *Journal of Geographical Systems, 12*(4), 435–459.

Munda, G. (2008). *Social Multi-Criteria Evaluation for a Sustainable Economy* (Vol. 17). Berlin: Springer.

OECD. (2005) *Handbook On Constructing Composite Indicators: Methodology And User Guide*. STD/DOC(2005)3

Papadimitriou, E., Norlen, H. and Del Sorbo, M. (2020). *JRC Statistical Audit of the 2020 Gender Equality Index*. European Commission, Joint Research Centre.

Permanyer. I. (2010). The measurement of multidimensional gender inequality: Continuing the Debate. *Social Indicators Research. 95*: 181-198.

Permanyer. I. (2013). Are UNDP indices appropriate to capture gender inequalities in Europe?. *Social Indicators Research. 100*: 927-950.

Peña, J. B. (1977). *Problemas de la medición del bienestar y conceptos afines* (Una aplicación al caso español). (Madrid: Instituto Nacional de Estadística (INE)).

Plantenga, J., Remery, C., Figueiredo, H. and Smith, M. (2009). Towards a European Union gender equality index. *Journal of European Social Policy. 19* (1): 19-33.

Radojicic, M., Savic, G. and Jeremic, V. (2018). Measuring the efficiency of banks: the bootstrapped i-distance gar dea approach. *Technological & Economic Development of Economy*, *24* (4)*: 1581-1605*.

Reichelt, M., Makovi, K. and Sargsyan, A. (2021). The impact of COVID-19 on gender inequality in the labor market and gender-role attitudes. *European Societies*. *23* (1)*:* S228-S245.

Rogge, N. and Van Nijverseel, I. (2019). Quality of life in the European Union: A multidimensional analysis. *Social Indicators Research. 141*(2): 765-789.

Saisana, M., Saltelli, A. and Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of Royal Statistical Society. 168* (2): 307-323.

Saranya, C. and Manikandan, G. (2013). A study on normalization techniques for privacy preserving data mining. *International Journal of Engineering and Technology (IJET). 5*(3): 2701-2704.

Savic, D., Jeremic, V. and Petrovic, N. (2016). Rebuilding the pillars of sustainable society index: A multivariate post hoc I-distance approach. *Problemy Ekorozwoju– problems of Sustainable Development. 12* (1)*:* 125-134D

Somarriba, N. and Peña, B. (2009). Synthetic indicators of quality of life in Europe. *Social Indicators Research, 94*(1), 115–133.

Tarantola, S. and Saltelli, A. (2007). Composite indicators: The art of mixing apples and oranges. Presented on Composite Indicators – Boon or Bane. Statistisches Bundesamt, Germany. JRC43341.

Trinchera, L., Russolillo, G. and Lauro, C. N. (2008). Using categorical variables in pls path modeling to build system of composite indicators. *Statistica Applicata, 20*(2): 309–330.

UN. (2012). Indicators of Gender Equality, *Working paper 21,* Economic Commission for Europe, Conference of European Statisticians, Switzerland, Geneva, 5th March 2012.

UN. (2014a). *World Survey on the Role of Women in Development 2014 - Gender Equality and Sustainable Development*. New York: United Nations. ISBN 978-92-1-130330-8.

UN. (2000). 55/2. United Nations Millennium Declaration. Available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N00/559/51/PDF/N0055951.pdf?OpenElement. Last access: 15/02/2023.

UNDP. (2023). Gender Development Index (GDI). Available at: https://hdr.undp.org/gender-development-index#/indicies/GDI. Last access: 20/01/2023.

Velluti, S. (2014). Gender regimes and gender equality measures in Central Eastern European Countries post-accession: The case of Hungary and Poland. *Journal of International and Comparative Social Policy. 30* (1)*:* 79-91.

Verbunt, P. and Rogge, N. (2018). Geometric composite indicators with compromise Benefit-of-the-Doubt weights. *European Journal of Operational Research. 264* (1)*:* 388-401.

WEF. (2022). The Global Gender Gap Report 2022. Available at: https://www.weforum.org/reports/global-gender-gap-report-2022/. Last access: 10/12/2022

Witkowska, D. (2013). Gender disparities in the labor market in the EU. *International Advances in Economic Research. 19*: 331-354.

Zelany, M. (1974). A concept of compromise solutions and the method of the displaced ideal. *Computers & Operations Research. 1*(3–4): 479–496

doi.org/10.26398/IJAS.248

BAYESIAN INFERENCE FOR EXPONENTIATED INVERTED WEIBULL DISTRIBUTION IN PRESENCE OF PROGRESSIVE TYPE-II CENSORING

Teena Goyal

Department of Mathematics & Statistics, Banasthali Vidyapith, Rajasthan, India

Piyush K Rai, Mahaveer S Panwar

Department of Statistics, Banaras Hindu University, Varanasi, India

Sandeep K Maurya [1]

Department of Statistics, Central University of South Bihar, Bihar, India

Abstract *The present article gives the point as well as interval estimates for the parameters and lifetime characteristics as reliability and hazard rate functions of the exponentiated inverted Weibull distribution in presence of progressive type-II censored data under classical and Bayesian approach. The point estimates under classical paradigm are obtained with the help of maximum likelihood estimation procedure and in case of Bayesian paradigm, gamma prior is used for both unknown parameters under squared error and linex loss functions. The Metropolis-Hasting algorithm is applied to generate MCMC samples from the posterior density. In case of interval estimation; bootstrap confidence intervals (Boot-t and Boot-p) and highest posterior density intervals for the unknown parameters are computed. The performance of these estimates are studied on the basis of their simulated risks and length of intervals. Additionally, a real dataset is used to illustrate the proposed censoring technique and a simulation study is used to support the given study.*

Keywords: *Progressive censoring, Bayes estimation, Loss function, Metropolis-Hastings algorithm, Simulated risk.*

1. Introduction

In life testing experiments, Weibull distribution is one of the most suitable, applicable and famous model among the other existing lifetime models. As a result of the wide diversity of the versatile mechanism, the two parameter Weibull distribution is used on a large scale in the field of survival and reliability theory; especially for the non-censored data. Generally, there are three parameters involved in the above distribution namely; scale, shape and location parameters. The estimation of these parameters can be possible by using the different methods available in the statistical literature as graphically and analytically. Jiang and

[1] sandeepmaurya.maurya48@gmail.com

Murthy (1999) has introduced an approach to characterize the parameters of exponentiated Weibull distribution graphically. Analytical methods include maximum likelihood estimation, least squares estimation and method of moments, etc. The analytical techniques of estimation of the parameters are more accurate and reliable than the graphical methods (see Dolas et al. (2014)). After that, Mudholkar and Srivastava (1993) has proposed generalization of the Weibull distribution as exponentiated Weibull distribution and studied its statistical properties. Later, Flaih et al. (2012) has proposed a generalization of inverted Weibull (IW) distribution, named as exponentiated inverted Weibull (EIW) distribution, by adding one more shape parameter exponentially in IW distribution.

Let X be a random variable, said to follow EIW distribution if its probability density function (PDF) is given as:

$$f(x;\theta,\beta)=\theta\beta x^{-(\beta+1)}(e^{-x^{-\beta}})^{\theta}; \qquad x>0, \quad \theta>0, \quad \beta>0 \tag{1}$$

here, θ and β both are the shape parameters. If we put $\theta=1$, it will convert to its baseline inverse Weibull distribution and if we put $\beta=1$ then, it becomes the exponentiated inverted exponential distribution. The reliability function of the EIW distribution is given by

$$R(t;\theta,\beta)=1-(e^{-t^{-\beta}})^{\theta}; \qquad t>0, \quad \theta>0, \quad \beta>0$$

and its associated hazard rate is

$$h(t;\theta,\beta)=\frac{\theta\beta t^{-(\beta+1)}(e^{-t^{-\beta}})^{\theta}}{1-(e^{-t^{-\beta}})^{\theta}}; \qquad t>0, \quad \theta>0, \quad \beta>0.$$

Flaih et al. (2012) mentioned that the shape of the PDF of the EIW distribution are uni-model and hazard rate function has upside-down bathtub nature. Singh et al. (2002) has discussed the estimation of parameters for the exponentiated Weibull family under linex loss function and the same distribution is studied by Singh et al. (2005) for the censored data. Kundu and Howlader (2010a) has discussed the inferences and prediction of the inverse Weibull distribution in the case of censored data. Flaih et al. (2012) has discussed the model selection between EIW and IW distributions.Later, Ahmad et al. (2014) has explained the estimation of EIW distribution under asymmetric loss functions.

The case of censored data do arise when a researcher may receive incomplete or partially known data. Specially in reliability and life-testing experiments, in which items are either lost or removed from experiment before its failure, intentionally or unintentionally. For example, in an experiment, if an individual gives up from the experiment, accidental breakage occurs or some abrupt circumstances arises like unavailability of testing facilities etc. in such cases, the complete sample information are difficult to be found. These type of datasets are know as censored data. In general, there are two conventional censoring schemes named, type-I and type-II censoring schemes.

The type-I censoring scheme, due to time constraint, also known as time censoring. Under this scheme, the number of failure observations is random and may vary from experiment to experiment but the time duration of the study is fixed in advanced for each of the experiment. This censoring scheme has advantage of saving time duration of the experiment. While, in the type-II censoring, investigator fixes the number of observations before the experiment started. Such type censoring is also known as failure censoring, since the number of the observations is prefixed. Here, duration of the life-test is a random variable i.e. it may vary from experiment to experiment whereas, the number of observation is fix known constant, thus, it ensures the availability of a fixed number of observations for the study. See Ng et al. (2006), Balakrishnan et al. (2007), Kundu and Howlader (2010b), Joarder et al. (2011), Dey and Kundu (2012), Han and Kundu (2015), Prajapati et al. (2020) and Goyal et al. (2019) etc. for more details about censoring scheme and estimation of the parameter for censored data. The type-II censoring ensures about the number of observations, it guarantees the desired efficiency of the inferential procedure. But these two censoring schemes do not allow the dropping or removal of any experimental unit before their failure. Later Balakrishnan and Sandhu (1995) has discussed the algorithm of a more advanced censoring scheme called, progressive type-II censoring (PTIIC) scheme, which allows the flexibility of removals. PTIIC is the generalization of failure censoring (type-II) scheme. Initially, the PTIIC scheme has been discussed by Herd (1956), even though he referred to them as “multi-censored samples”. After that, the importance and applicability of the progressive censoring scheme have been discussed by Cohen (1963) and Viveros and Balakrishnan (1994) obtained the interval estimation of lifetime under PTIIC scheme. Later Balakrishnan et al. (2003), Kundu (2008) and Kundu and Biswabrata (2009) have discussed the scheme for different distributions. Almetwally et al. (2023) have discussed the Bayesian analysis under progressive type -II censoring for unit- Weibull distribution. For more literature, reader may refers to Aggarwala and Balakrishnan (1998), Ng and Chan (2007), Raqab et al. (2010) and El-Din and Shafay (2013) etc.

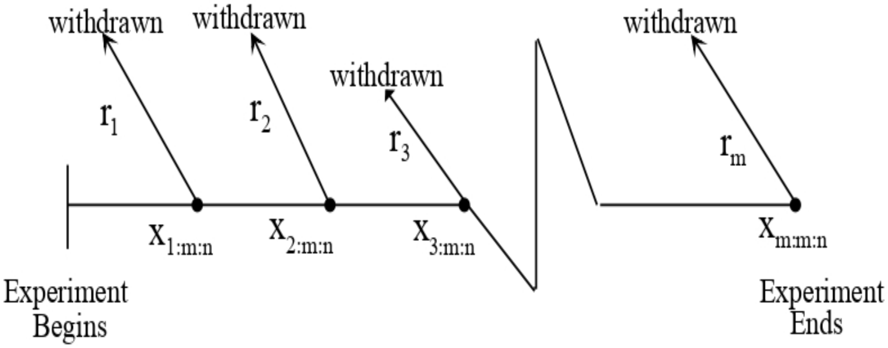

Figure 1: Schematic representation of progressive Type-II censoring scheme.

1.1. Progressive Censoring & its Likelihood Function

In the PTIIC scheme, along with the failure items, removals are also play an important role. The scheme is discussed below.

1. Suppose 'n' units are placed in a life testing experiment at time zero (starting point of time) and 'm' failure times are going to be observed with pre-decided removal scheme $r = (r_1, r_2, ..., r_m)$. Here all $r_i's$ $(i = 1, 2, \cdots, m)$ are positive integers.

2. At the time of first failure, r_1 of the surviving units randomly selected from the remaining '$n-1$' units and removed. At the time of second failure, r_2 of the surviving units are randomly selected from the resting '$n-2-r_1$' items and removed from the life-test experiment.

3. After that, at the time of m^{th} failure, all the waiting units $r_m = n - r_1 - r_2 - ... - m$ are removed and then the experiment is stopped.

Here, the observed failure times are denoted by $x_{i:m:n}$, where i denotes the i^{th} failure time, m denotes the total number of observation required (prefixed) and n denotes the total number of items placed on life test. Thus, the observed sample information in PTIIC scheme is $x_{1:m:n}, x_{2:m:n}, \cdots, x_{m:m:n}$. A pictorial representation of PTIIC scheme is given in Figure 1.

The likelihood function in the PTIIC scheme, (see Balakrishnan and Sandhu (1995)) is as follows;

$$L(x_i \mid \theta, \beta) = C \prod_{i=1}^{m} f(x_i \mid \theta, \beta)[1 - F(x_i \mid \theta, \beta)]^{r_i} \tag{2}$$

where C is the constant, formulated as:

$$C = n \times (n - r_1 - 1) \times \cdots \times (n - r_1 - r_2 - \cdots - r_m - 1 - m + 1). \tag{3}$$

The main emphasis of this paper is to test the efficacy of Bayes estimates for the parameters of the EIW distribution based on PTIIC. Motivated by these literature, here, we are trying to find better estimator for the parameter of EIW distribution using squared error and linex loss functions. The rest of the article is organized as follows: Section 2, deals with the classical estimation of parameters and Section 3, discusses the technique of Bayesian estimation for the parameters with very short description of prior, loss functions, posteriors, and M-H algorithm. Algorithm for a generation of the sample from PTIIC scheme is introduced in Section 4. In Section 5, techniques of parametric bootstrap confidence interval and HPD interval are discussed to construct the CIs for the unknown parameters. A brief study on simulation is discussed in Section 6. Particular real dataset is analyzed in the Section 7, and the conclusions of the present paper are commented in the last Section 8.

2. Classical Estimation

In this section, we have discussed the MLE of the parameters θ and β based on the data observed under the PTIIC scheme (as discussed in the Subsection 1.1).

2.1. Maximum Likelihood Estimator

Let $x = \{x_{1:m:n}, x_{2:m:n}, \cdots, x_{m:m:n}\}$ be a random sample from the EIW distribution with the PDF given in the equation (1). Then, the likelihood function by using the equation (2) is

$$l(\theta, \beta \mid x) = C \prod_{i=1}^{m} \theta \beta {x_i}^{-(\beta+1)} (e^{-{x_i}^{-\beta}})^{\theta} [1 - (e^{-{x_i}^{-\beta}})^{\theta}]^{r_i}.$$

And, the log of likelihood function is

$$L(\theta, \beta \mid x) = \log C + \sum_{i=1}^{m} \log \left[\theta \beta {x_i}^{-(\beta+1)} (e^{-{x_i}^{-\beta}})^{\theta} [1 - (e^{-{x_i}^{-\beta}})^{\theta}]^{r_i} \right]. \tag{4}$$

Where constant C is given in equation (3). Therefore, the ML estimator of the parameter θ and β can be obtained by differentiating the log likelihood function with respect to the corresponding parameters and equating to zero, we get

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=1}^{n} \left[\frac{1}{\theta} - {x_i}^{-\beta} + r_i \frac{(e^{-{x_i}^{-\beta}})^{\theta} {x_i}^{-\beta}}{\{1 - (e^{-{x_i}^{-\beta}})^{\theta}\}} \right] = 0.$$

$$\frac{\partial \log L}{\partial \beta} = \frac{n}{\beta} - \sum_{i=1}^{n} \log x_i + \theta \sum_{i=1}^{n} \frac{\log x_i}{{x_i}^{\beta}} - \sum_{i=1}^{n} r_i \frac{(e^{-{x_i}^{-\beta}})^{\theta} \theta {x_i}^{-\beta} \log x_i}{\{1-(e^{-{x_i}^{-\beta}})^{\theta}\}} = 0.$$

Since the above likelihood equations are not in close form and thus can not be solved it analytically. Therefore, to obtain the the solution from these likelihood equations, we have used Newton-Raphson method based on an iterative procedure. After getting the ML estimator of the parameters, the estimated reliability $\hat{R}_{ML}$ and hazard rate function $\hat{h}_{ML}$ at specific time t can be obtained by using the in-variance property of ML estimator. Thus,

$$\hat{R}_{ML} = 1-(e^{-t^{-\hat{\beta}_{ML}}})^{\hat{\theta}_{ML}}; \qquad t>0, \quad \theta>0, \quad \beta>0.$$

And

$$\hat{h}_{ML} = \frac{\hat{\theta}_{ML}\hat{\beta}_{ML} t^{-(\hat{\beta}_{ML}+1)}(e^{-t^{(-\hat{\beta}_{ML})}})^{\hat{\theta}_{ML}}}{1-(e^{-t^{-\hat{\beta}_{ML}}})^{\hat{\theta}_{ML}}}; \qquad t>0, \quad \theta>0, \quad \beta>0.$$

Here, $\hat{\theta}_{ML}$ and $\hat{\beta}_{ML}$ are the ML estimator of the parameters θ and β respectively.

3. Bayes Method of Estimation

In this section we have considered another approach namely Bayesian method. The reason behind the consideration is that, in the Bayesian approach added more flexible and accurate result as it incorporates prior knowledge with the sample in-formation. Another advantage of the consideration is that the Bayesian approach provides more appropriate results in small as well in large sample.

In Bayesian paradigm, posterior distribution is an effect of two components namely, a prior distribution and the likelihood function, calculated from the statistical model for the observed data. The prior distribution is the distribution of the parameter assumed before the data observed. The choice of the prior distribution may not be easily determined. For the selection of the prior distribution, one can see Berger and Sun (1993), Raqab and Madi (2005) and Singh et al. (2016). There are mainly two different categorization to the prior distribution of parameters defined as proper and improper prior. Another way to defined priors are based on information available in advance and called as informative and non-informative prior. Here, we use the prior distribution for θ and β as $Gamma(a,b)$ and $Gamma(c,d)$ respectively to obtain the posterior distribution.

The choice of the hyper-parameters of the priors (θ,β) are based on the information available in term of prior mean and prior variance. This chosen value of the hyper-parameter may be taken in a way that if we choose any two independent information as prior mean and variance of the priors (θ,β), then $(\mu_1 = a/b, \sigma_1 = a/b^2)$ and $(\mu_2 = c/d, \sigma_2 = c/d^2)$ respectively. Here, μ_1 and μ_2 are the true mean value of the parameter (θ,β) respectively and σ_1 and σ_2 are the true variance of the parameter (θ,β) respectively. Now by using this information, the hyper parameters can be easily evaluated from this relation, $(a = \mu_1/\sigma_1, b = \mu_1^2/\sigma_1)$ and $(c = \mu_2/\sigma_2, d = \mu_2^2/\sigma_2)$ respectively. See Kundu (2008), Singh et al. (2013), Dey et al. (2016), Singh et al. (2016) and El-Sherpieny et al. (2022) for more details about the choice of the hyper-parameters. Now, the joint prior distribution of θ and β is given as

$$\pi(\theta,\beta) \propto \theta^{a-1}\beta^{c-1}e^{-b\theta-d\beta}; \qquad (\theta,\beta) > 0, (a,b,c,d) > 0. \tag{5}$$

3.1. Loss Function

“A loss function is a function that maps an event or the values of one or more variables on a real number intuitively representing some cost associated with the even". In Bayesian statistics, a loss function is used for the estimation of parameters. Here we have considered two different widely used loss functions namely Squared Error Loss Function (SELF) and Linex Loss Function (LLF).

1. *Squared Error Loss Function:* It is a commonly used symmetric loss function, defined as

   $$L(\hat{\theta}_{BS},\theta) = (\hat{\theta}_{BS} - \theta)^2$$

   where $\hat{\theta}_{BS}$ is the Bayes estimator under SELF for the given parameter θ.

2. *Linex Loss Function:* This loss function is an asymmetric loss function. Zellner (1986) proposed this loss function for the estimation and prediction of a scaler parameter. The form of LLF is given as

   $$L(\hat{\theta}_{BL},\theta) = e^{\delta(\hat{\theta}_{BL}-\theta)} - \delta(\hat{\theta}_{BL} - \theta) - 1$$

   where $\delta \neq 0$ is a constant which determines the shape of the loss function. In particular, the LLF increases almost linearly for negative error and almost exponentially for positive error. Thus, under this loss function, over estimation is considered to be more serious than the under estimation. The behavior of the LLF for the small values of δ, is approximately same as the SELF.

3.2. Posterior Probability Density Function

Let $x = \{x_{1:m:n}, x_{2:m:n}, \cdots, x_{m:m:n}\}$ be a random sample from EIW distribution and the parameters θ and β have prior probabilities $\pi(\theta)$ and $\pi(\beta)$ respectively. Then by using equations (4) and (5), the joint posterior density function is given as:

$$\begin{aligned} \pi(\theta,\beta \mid x) &= \frac{\pi(\theta,\beta)L(\theta,\beta \mid x)}{\int\limits_0^\infty\int\limits_0^\infty \pi(\theta,\beta)L(\theta,\beta \mid x)d\theta d\beta}; \qquad \theta > 0, \beta > 0 \\ &\propto \theta^{a-1}\beta^{c-1}e^{-b\theta-d\beta}\prod_{i=1}^{m}\theta\beta x_i{}^{-(\beta+1)} \\ &\times (e^{-x_i{}^{-\beta}})^{\theta}[1-(e^{-x_i{}^{-\beta}})^{\theta}]^{r_i}. \end{aligned} \tag{6}$$

Now, the Marginal posterior densities of the parameters θ and β can be obtained by integrating the equation (6) with respect to β and θ respectively. And it can be written as equation (7) and equation (8) respectively as

$$\begin{aligned} \pi(\theta \mid x_i) &= \int_0^\infty \pi(\theta,\beta \mid x_i)d\beta \\ &\propto \int\limits_0^\infty \theta^{a-1}\beta^{c-1}e^{-b\theta-d\beta}\prod_{i=1}^{m}\theta\beta x_i{}^{-(\beta+1)} \\ &\times (e^{-x_i{}^{-\beta}})^{\theta}[1-(e^{-x_i{}^{-\beta}})^{\theta}]^{r_i}d\beta. \end{aligned} \tag{7}$$

$$\begin{aligned} \pi(\beta \mid x_i) &= \int_0^\infty \pi(\theta,\beta \mid x_i)d\theta \\ &\propto \int\limits_0^\infty \theta^{a-1}\beta^{c-1}e^{-b\theta-d\beta}\prod_{i=1}^{m}\theta\beta x_i{}^{-(\beta+1)} \\ &\times (e^{-x_i{}^{-\beta}})^{\theta}[1-(e^{-x_i{}^{-\beta}})^{\theta}]^{r_i}d\theta. \end{aligned} \tag{8}$$

3.2.1. Bayes Estimator under SELF

The Bayes estimator under the SELF is nothing but the posterior mean of the corresponding parameters. Let it is denoted by $\hat{\theta}_{BS}$ and. Therefore, the Bayes estimator of the parameter θ can be obtained as:

$$\begin{aligned}\hat{\theta}_{BS} &= E(\hat{\theta}) = \int_0^\infty \theta\pi(\theta \mid x_i)d\theta \\ &\propto \int_0^\infty \theta \int_{i=0}^{\infty} \theta^{a-1}\beta^{c-1}e^{-b\theta-d\beta}C\prod_{i=1}^{m}\theta\beta \\ &\times\ x_i^{-(\beta+1)}(e^{-x_i^{-\beta}})^{\theta}[1-(e^{-x_i^{-\beta}})^{\theta}]^{r_i}d\beta d\theta.\end{aligned}$$

Similarly, the Bayes estimator of the parameter β can be obtained as:

$$\begin{aligned}\hat{\beta}_{BS} &= E(\hat{\beta}) = \int_0^\infty \beta\pi(\beta \mid x_i)d\beta \\ &\propto \int_0^\infty \beta \int_0^\infty \theta^{a-1}\beta^{c-1}e^{-b\theta-d\beta}C\prod_{i=1}^{m}\theta\beta \\ &\times\ x_i^{-(\beta+1)}(e^{-x_i^{-\beta}})^{\theta}[1-(e^{-x_i^{-\beta}})^{\theta}]^{r_i}d\theta d\beta.\end{aligned}$$

3.2.2. Bayes Estimator under LLF

Since, the Bayes estimator for the parameter θ under LLF (say $\hat{\theta}_{BL}$) is defined as $\left[\frac{-1}{a}\log E(e^{-a\theta})\right]$. Thus, for the considered distribution, the expression is given as:

$$\begin{aligned}\hat{\theta}_{BL} &= \frac{-1}{a}\log\left[\int_0^\infty e^{-a\theta}\pi(\theta \mid x_i)d\theta\right] \\ &\propto \frac{-1}{a}\log\int_0^\infty e^{-a\theta}\int_0^\infty \theta^{a-1}\beta^{c-1}e^{-b\theta-d\beta} \\ &\times \prod_{i=1}^{m}\theta\beta x_i^{-(\beta+1)}(e^{-x_i^{-\beta}})^{\theta}[1-(e^{-x_i^{-\beta}})^{\theta}]^{r_i}d\beta d\theta.\end{aligned}$$

Similarly, the Bayes estimator for the parameter β is:

$$\begin{aligned}
\hat{\beta}_{BL} &= \frac{-1}{a}\log\left[\int_{0}^{\infty} e^{-a\beta}\pi(\beta \mid x_i)d\beta\right] \\
&\propto \frac{-1}{a}\log\int_{0}^{\infty} e^{-a\beta}\int_{0}^{\infty}\theta^{a-1}\beta^{c-1}e^{-b\theta-d\beta} \\
&\times \prod_{i=1}^{m}\theta\beta {x_i}^{-(\beta+1)}(e^{-{x_i}^{-\beta}})^{\theta}\left[1-(e^{-{x_i}^{-\beta}})^{\theta}\right]^{r_i} d\theta d\beta.
\end{aligned}$$

3.3. MCMC Simulation

The expressions for the Bayes estimators under SELF and LLF are not in closed form. So it needed some algorithm to draw sample and find its estimates. One of the famous and efficient technique of doing so is Markov Chain Monte Carlo (MCMC) method. One may refer Gilks et al. (1995), Gelfand (1996), Dagpunar (2007), Marin and Robert (2007), Chen et al. (2012) and Robert and Casella (2013) for more about MCMC technique. The integral involved in Bayes estimators do not solve analytically. In such a situation, MCMC methods namely Metropolis-Hastings (M-H) algorithm (see Hastings (1970)) can be effectively used.

To obtain the MCMC samples from the posterior probability $\pi(\theta \mid data)$, using the Metropolis-Hastings (M-H) algorithm, we have considered a normal distribution as the proposal density i.e. $N(\mu, \Sigma)$ where Σ is the variance-covariance matrix. It may be the point here that, if we generate observation from the normal distribution, we may get negative values also which are not possible as the parameters under consideration are positive valued. Therefore, we take the absolute value of generated observation. The M-H algorithm starts with an initial value of the parameter say θ^0 and specified a rule for simulating the t^{th} value in the sequence θ^t given the $(t-1)^{st}$ value in the sequence θ^{t-1}. This rule consists of a proposal density which simulates a candidate value say θ^* and acceptance probability say P. This algorithm can be described as follows:

1. Set the initial guess of parameter θ, say θ^0 from uniform $U(0,1)$.

2. Simulate a candidate value θ^* from a proposal density $p(\theta^* \mid \theta^{t-1})$.

3. Compute the ratio R= $\frac{\pi(\theta^*)p(\theta^{t-1}|\theta^*)}{\pi(\theta^{t-1})p(\theta^*|\theta^{t-1})}$.

4. Compute acceptance probability $P = \min\{R, 1\}$.

5. Take a sample value θ^t, such that, $\theta^t = \theta^*$ with probability P; otherwise $\theta^t = \theta^{t-1}$.

After getting MCMC samples from posterior distribution, we can find the Bayes estimates for the parameters in the following way

$$E(\theta|data) = \frac{1}{N - N_0} \sum_{i=N_0+1}^{N} \theta_i$$

where N_0 is burn-in period of Markov chain and N be the sufficiently large number of replications. In using the above algorithm, the problem arises how to choose the initial guess. Here, we propose to the use of ML estimate of the parameter θ, obtained by using the method described in subsection 2.1, as the initial value for MCMC processes. The choice of covariance matrix Σ is also an important issue, one can follow Ntzoufras (2011) for more details. One choice for Σ would be the asymptotic variance-covariance matrix $I^{-1}(\hat{\theta})$. While generating M-H samples by taking $\Sigma = I^{-1}(\hat{\theta})$, we noted that the acceptance rate for such a choice of Σ is about 30%. By acceptance rate, we mean the proportion of times a new set of values is generated at the iteration stages. It is well known that if the acceptance rate is low, a good strategy is to run a small pilot run using diagonal Σ as a rough estimate of the correlation structure for the target posterior distribution and then re-run the algorithm using the corresponding estimated variance-covariance matrix; for more detail see (Gelman et al., 2013, pp. 334-335), Kaushik et al. (2017) and Maurya et al. (2017).

4. Algorithm for Sample Generation under PTIIC Scheme

We have used the following steps to generate a PTIIC sample from the EIW distribution. The steps are:

1. Specify the values of n, m, $\underset{\sim}{\alpha} = (\theta, \beta)$ and $r = (r_1, r_2, ..., r_m)$.

2. Generate m i.i.d. random numbers $w_1, w_2, ..., w_m$ from uniform $U(0,1)$ distribution.

3. Set $V_i = w_i^{1/(i + \sum_{j=m-i+1}^{m} r_j)}$; for $i = 1, 2, ..., m$.

4. Set $U_{i:m:n} = 1 - V_m V_{m-1} ... V_{m-i+1}$ for $i = 1, 2, ..., m$. Then $U_{1:m:n}, U_{2:m:n}, ...,$ $U_{m:m:n}$ are the required PTIIC sample from the uniform $U(0,1)$ distribution.

5. Finally, set $x_{i:m:n} = F^{-1}(U_{i:m:n})$, for $i = 1, 2, ..., m$, where $F^{-1}(\cdot)$ is the inverse distribution function of EIW distribution.

Then, $x_{1:m:n}, x_{2:m:n}, \ldots, x_{m:m:n}$ are the required n random PTIIC sample from the EIW distribution.

5. Interval Estimation

In this section we have computed confidence intervals for the parameters of the EIW distribution under classical and the Bayesian setup. In classical setup, we have calculated parametric boot strep intervals namely; Boot-p and Boot-t. While in the Bayesian setup, we have calculated the highest posterior density (HPD) intervals. The details about these intervals are given below.

5.1. Bootstrap Confidence Interval

Sometime the class intervals based on the asymptotic property or the normal theory assumption do not perform good for small samples. In that situation, the use of bootstrap methods, one can obtain the accurate intervals without using the normal theory assumption. The bootstrap methods make computer-based adjustments to the standard intervals endpoints and surely improve the coverage accuracy by an order of magnitude, at least asymptotically. Here, we have discussed two types of CIs using bootstrap method. The parametric percentile bootstrap (Boot-p) suggested by Efron (1982) and parametric studentized bootstrap (Boot-t), suggested by Hall (1988). See Efron (1992) and DiCiccio and Efron (1996) for more details about bootstrap CIs.

5.1.1. Parametric Boot-p

An algorithm for the Boot-p CIs is as follows:

1. Assemble the PTIIC data and obtain ML estimators for the parameters θ & β, denoted as $\hat{\theta}_{ML}$ & $\hat{\beta}_{ML}$.
2. Generate a PTIIC sample by using ML estimators of the parameters based on pre-specified removal scheme $r = (r_1, r_2, ..., r_m)$.
3. Generate B number of bootstrap samples from the above generated samples.
4. Obtain ML estimators for each B bootstrap sample, denoted as $\left\{\hat{\theta}_1^*, \hat{\beta}_1^*\right\}$, $\left\{\hat{\theta}_2^*, \hat{\beta}_2^*\right\}$, ..., $\left\{\hat{\theta}_B^*, \hat{\beta}_B^*\right\}$.
5. Arrange these generated samples in ascending orders as $\left\{\hat{\theta}_{(1)}^*, \hat{\theta}_{(2)}^*, ..., \hat{\theta}_{(B)}^*\right\}$ and $\left\{\hat{\beta}_{(1)}^*, \hat{\beta}_{(2)}^*, ..., \hat{\beta}_{(B)}^*\right\}$.

A pair of $100(1-\alpha)\%$ Boot-p CIs for θ & β are given by $\left[\hat{\theta}^*_{(B\alpha/2)}, \hat{\theta}^*_{(B(1-\alpha/2))}\right]$ and $\left[\hat{\beta}^*_{(B\alpha/2)}, \hat{\beta}^*_{(B(1-\alpha/2))}\right]$ respectively.

5.1.2. Parametric Boot-t

The algorithm for generating p-Boot is very simple, though, if the sample size is very small then, percentile approach is not so much accurate. Thus, in this condition, studentized t bootstrap (Boot-t) approach can be used. It gives more accuracy to results than the percentile approach. The algorithm of the Boot-t CIs is just an extension of the algorithm of p-Boot.

5. Repeat step 1-4 as in Boot-p approach.

6. Compute, standard errors of the parameters, denoted as $\{s\hat{e}_1{}^*(\theta), s\hat{e}_1{}^*(\beta)\}$, $\{s\hat{e}_2{}^*(\theta), s\hat{e}_2{}^*(\beta)\}$, $\cdots$, $\{s\hat{e}_B{}^*(\theta), s\hat{e}_B{}^*(\beta)\}$.

7. Compute, statistics $z^*_B(\theta) = \frac{\hat{\theta}^*_B - \hat{\theta}_{ML}}{s\hat{e}_B{}^*(\theta)}$ and $z^*_B(\beta) = \frac{\hat{\beta}^*_B - \hat{\beta}_{ML}}{s\hat{e}_B{}^*(\beta)}$.

8. Arrange $z^*_B(\theta)$ and $z^*_B(\beta)$ in ascending orders, denoted as $z^*_{(B)}(\theta)$ and $z^*_{(B)}(\beta)$. A pair of $100(1-\alpha)\%$ Boot-t CIs for θ & β are given by

$$\left[\hat{\theta}_{ML} - z^*_{(B(1-\alpha/2))}(\theta) * \hat{se}(\theta),\ \hat{\theta}_{ML} + z^*_{(B\alpha/2)}(\theta) * \hat{se}(\theta)\right]$$

and

$$\left[\hat{\beta}_{ML} - z^*_{(B(1-\alpha/2))}(\beta) * \hat{se}(\beta),\ \hat{\beta}_{ML} + z^*_{(B\alpha/2)}(\beta) * \hat{se}(\beta)\right]$$

respectively.

5.2. Highest Posterior Density Interval

The HPD credible intervals (see Box and Tiao (1973) and Chen and Shao (1999)) of the parameter $\underset{\sim}{\alpha} = (\theta, \beta)$ are obtained on the basis of ordered MCMC samples of α as $\underset{\sim}{\alpha}_{(1)}, \underset{\sim}{\alpha}_{(2)}, \cdots, \underset{\sim}{\alpha}_{(N)}$. After that, $100(1-\alpha)\%$ credible interval for the parameter $\underset{\sim}{\alpha}$ is obtained as $((\underset{\sim}{\alpha}_{(1)}, \underset{\sim}{\alpha}_{[(1-\alpha)N]+1}), \cdots, (\underset{\sim}{\alpha}_{[N\alpha]}, \underset{\sim}{\alpha}_N).)$ Where $[Y]$ denotes the largest integer less than or equal to Y. The, HPD credible intervals provides shortest length CIs. See also Edwards et al. (1963), Ng et al. (2006), Kundu and Howlader (2010b) and Singh et al. (2013) etc.

6. Simulation Study

In this section we have performed simulation study based on the PTIIC samples for EIW distribution and estimates the parameters of the model under the above discussed classical and Bayesian methods. We have also calculated the CIs for the model parameters along with the model survival and hazard rate functions. We have calculated here the simulated risk function also. Performance of Bayes estimators and ML estimators are examined based on the simulation. The steps involve to perform the study are enumerated as follows.

1. Generate PTIIC samples using the algorithm discussed in the Section 3 for particular values of n, m, θ, β and r.

2. The ML estimators of the parameters, reliability and hazard rate function have been computed for these particular values.

3. In case of Bayesian analysis, we have assumed that both the parameters have gamma prior. The chosen values of the hyper-parameters are taken as $a = 0.2$, $b = 0.2$, $c = 0.2$ & $d = 0.2$, as particular case. The reason behind the choice of this hyper-parameter is to the consideration of informative prior. Also, the choice of hyper-parameters for the gamma prior should be guided by a combination of prior knowledge. One can choose large variance prior in case of lack of prior knowledge or it may be appropriate to choose hyper parameters that result in relatively flat or non-informative priors. As, we know that, in EIW distribution, both of the parameters play the role of shape parameters and both are sensitive for the shape of the distribution. So here, we have chosen same combination for choice of this hyper-parameter, as true mean 1 and true variance 5 for both of the hyper-parameters. Also, for the considered combination, when mean>variance, both the gamma priors of the parameters covers wide variation. For more details, see Section 3.

4. M-H algorithm of MCMC technique has been used to generate posterior samples.

5. From these simulated posterior samples, the Bayes estimators of the parameters, survival and hazard rate under the assumption of the above prior using SELF and LLF have been obtained.

6. Only one choice of loss function parameter δ is considered (δ = 0.1 as particular case) for LLF.

7. Boot-p and Boot-t CIs have obtained under classical set-up. And in Bayesian paradigm, we have constructed 95% HPD CI for both the parameters.

8. The values of the estimates, survival and hazard rate at time $t = 1$ have been reported in the Tables 5-7. Here the arbitrary chosen true value of the parameters (θ, β) are taken as $(2, 2)$. Table 8, shows the risks of these parameters under different estimation techniques and CIs for these estimators are mentioned in Tables 9-10.

9. The estimates of the parameters θ & β for $m = 32$ for varying combinations of values of the parameters as (θ, β) = (1.2, 2), (3, 2), (2, 0.5) & (2, 1.5) with their associated risks and estimators of survival and hazard rate functions in presence of different removal schemes under PTIIC as R_1, R_2, R_3 & R_4 have been tabulated from Tables 11-13.

The Figure 2 gives an idea about the quantiles of the parameters θ and β respectively. These are based on the MCMC samples, which explains the probability $(P(X \leq x) > 0.1, 0.5, 0.9)$, where X is a random variable. In our case, it is for θ and β parameters and x is the particular values of the MCMC sample.

For simulation study, we have generated different random samples of size n = 50, m = 20, 30, 32 and 40 with the parameters $\theta = 2$ and $\beta = 2$ from EIW PTIIC and taken different censoring schemes R_1,R_2,R_3 and R_4. The complete schemes along with these parameters values are given in Table 4. The simulation results for these schemes are given in the Tables 5-13.

Under different censoring schemes (see Table 4) p^*q (means number p is repeated q times) and for different parameter values, we conclude the following:

1. From the Table 8, the risks of the Bayes estimators are lesser as compared to the risk of the ML estimators. This table also shows that the risk of the ML estimator for the parameter θ is less than the parameter β for all the considered value of m and considered different removable schemes. But no such pattern found in case of risk of the Bayes estimators under different loss functions except $m = 20$, which shows the reverse result as in classical ML estimation under different removable schemes.

2. From the Table 8, we observed that the risk of the Bayes estimators of the parameters, under LLF, is consistently smaller than the risk of the estimators under SELF.

3. This table also shows that, for all the considered values of m, the Bayes risk under LLF are concentrated and converses to zero. So, this method may be accepted for this distribution.

4. From Tables 9 and 10, we see that the length of HPD intervals is smaller than the length of Boot-t CI and Boot-t is smaller than Boot-p CI for both the parameters θ and β.

5. From Table 12, we may conclude that the Bayes estimators have lesser risk as compared to classical estimation for different combinations of the parameters θ and β. One point is also noticeable that, for the EIW distribution, both the parameters (θ ,β) are play the role of shape parameter and both are sensitive.

6. This table also shows that, for any parameter (θ or β) smaller risk is associated with smaller value of parameter and vice-versa.

7. This table also shows that the under R_2 scheme, for all the combination of the parameters, risk under classical estimators are maximum for both the parameters.

8. This table also shows that, for all the considered parameters (θ or β) the Bayes risks under LLF are concentrated and converses to zero. So, this method may be accepted for this distribution.

7. Real Data Analysis

Here we have considered a real dataset of the remission time (in months) of 128 bladder cancer patients data, to show the applicability of the considered model in classical as well as Bayesian context in complete as well as in censored case. The dataset was reported by Lee and Wang (2003), and is given in Table 1.

The ML estimate of the parameters (θ and β), survival and hazard rate function based on the complete sample ($n = 128$) are obtained as $\hat{\theta}_{ML} = 2.4262$, $\hat{\beta}_{ML} = 0.7551$, $\hat{S}_{ML}(t = 1) = 0.9116$ & $\hat{h}_{ML}(t = 1) = 0.1776$ respectively. The Bayes estimate for the parameters θ and β, survival and hazard rate under SELF are $\hat{\theta}_{BS} = 2.1961$, $\hat{\beta}_{BS} = 0.7370$, $\hat{S}_{BS}(t = 1) = 0.8888$, $\hat{h}_{BS}(t = 1) = 0.2026$ respectively and under LLF estimates are $\hat{\theta}_{BL} = 2.1941$, $\hat{\beta}_{BL} = 0.7370$, $\hat{S}_{BL}(t = 1) = 0.8885$, $\hat{h}_{BL}(t = 1) = 0.2028$ respectively.

A PTIIC sample of size $m = 80$ is selected randomly from the complete sample of size $n = 128$ with the censoring scheme $R = (0^*32, 3^*16, 0^*32)$. The point and the interval estimates of the parameters are given in the Table 2. This table, shows the ML estimates and the Bayesian estimates based on MCMC technique using SELF and LLF for both parameters θ & β along with the interval estimates of both the parameters using bootstrap confidence intervals (Boot-p and Boot-t) technique and highest posterior density intervals.

The point and the interval estimates of the survival and hazard rate function at different times $t = 3,6$ and 12 are given in Table 3. From this table, one can observed that the length of the interval become shorter with respect to increase in time for all the considered times. And also, the length of the intervals are shortest in case of HPD intervals (here LL stands for lower limit and UL stands for upper limit of the intervals).

8. Conclusion

In this paper, we have proposed point as well as interval estimation under classical and Bayesian context of the parameters for the exponentiated inverse Weibull distribution. We have also estimated the survival and hazard rate functions of the considered model under progressive type-II censoring using maximum likelihood estimation and Bayesian analysis using gamma prior under SELF and LLF. Both the classical and Bayesian analyses methods have their own advantages and limitations, and this depends on factors such as the availability of prior information, the desire for uncertainty quantification. We have also obtained the confidence intervals for the parameters using parametric bootstrap methods (namely Boot-t and Boot-p) and Bayesian HPD intervals. We have taken a simulation study by using MCMC technique to compute the point estimations and their corresponding confidence intervals. From a simulated study, we can conclude that the Bayes estimators with an informative gamma prior may be used particu-larly when prior information is known. We also find that the Bayes risks is always smaller than the classical risks. Also, the Bayes risks under LLF for all the consid-ered parameters (θ or β) values and m more concentrated and converses to zero. So, one can also this method while dealing with EIW distribution. The perfor-mance of HPD intervals seems comparatively good because the computed risks are considerably smaller as compared to classical method as well as the length of the HPD intervals are minimum with respect to the bootstrap intervals.

In the simulation study, the results of estimated parameters and their confidence intervals as Boot-p, Boot-t and HPD intervals have also provided. The characteristics based on the functional form of parameters like survival and hazard rate functions with their confidence intervals have also provided under different censoring schemes.

The results obtained under this study may be motivate to researchers of the field of statistical inference to consider the facts for the better application of EIW distribution in real life testing circumstances.

Figure 2: Cumulative quantile plot for the parameters θ and β.

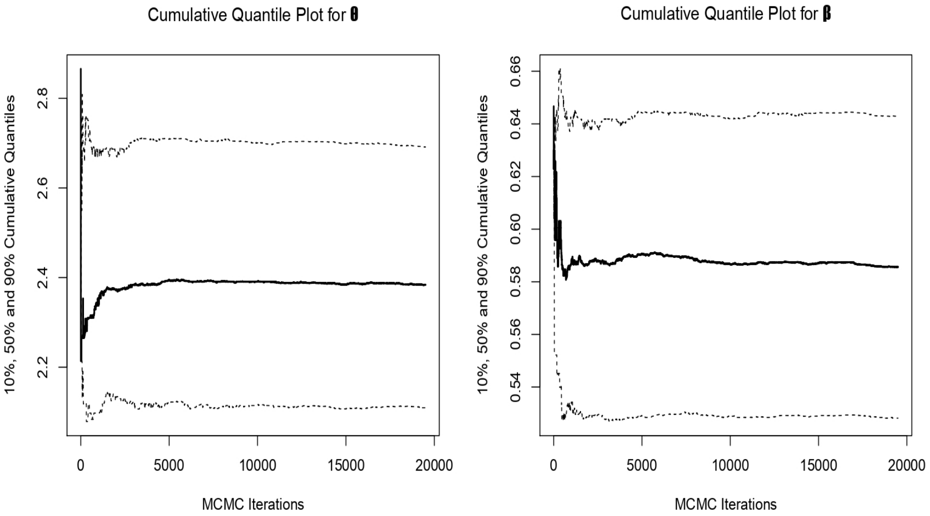

Table 1: Remission times (in months) of 128 bladder cancer patients.

0.08	2.09	3.48	4.87	6.94	8.66	13.11	23.63	0.2	2.23
3.52	4.98	6.97	9.02	13.29	0.4	2.26	3.57	5.06	7.09
9.22	13.8	25.74	0.5	2.46	3.64	5.09	7.26	9.47	14.24
25.82	0.51	2.54	3.7	5.17	7.28	9.74	14.76	6.31	0.81
2.62	3.82	5.32	7.32	10.06	14.77	32.15	2.64	3.88	5.32
7.39	10.34	14.83	34.26	0.9	2.69	4.18	5.34	7.59	10.66
15.96	36.66	1.05	2.69	4.23	5.41	7.62	10.75	16.62	43.01
1.19	2.75	4.26	5.41	7.63	17.12	46.12	1.26	2.83	4.33
5.49	7.66	11.25	17.14	79.0	51.35	2.87	5.62	7.87	11.64
17.36	1.4	3.02	4.34	5.71	7.93	11.79	18.1	1.46	4.4
5.85	8.26	11.98	19.13	1.76	3.25	4.5	6.25	8.37	12.02
2.02	3.31	4.51	6.54	8.53	12.03	20.28	2.02	3.36	6.76
12.07	21.73	2.07	3.36	6.93	8.65	12.63	22.69.		

Table 2: Point and interval estimates of the parameters (θ, β) under different techniques for the real dataset.

	ML	Boot-t	Boot-p	SELF	LLF	HPD
θ	2.7130	(1.7675, 3.5487)	(2.0886, 4.7906)	2.4000	2.3974	(1.9633, 2.8526)
β	0.6134	(0.4326, 0.7373)	(0.5079, 0.893)	0.5858	0.5857	(0.5004, 0.672)

Table 3: Point and interval estimates of survival and hazard rate functions under different techniques and time points for the real dataset.

	ML	Boot-t	Boot-p	SELF	LLF	HPD	ML	Boot-t	Boot-p	SELF	LLF	HPD
time	$\hat{S}(t)$	length	length	$\tilde{S}(t)$	$\tilde{S^*}(t)$	length	$\hat{h}(t)$	length	length	$\tilde{h}(t)$	$\tilde{h^*}(t)$	length
t=3	0.2509	0.127	0.1366	0.2834	0.2837	0.0679	0.2828	0.2295	0.3322	0.2462	0.2459	0.1178
t=6	0.4050	0.0551	0.0513	0.4316	0.4320	0.0245	0.0924	0.0577	0.0728	0.0820	0.0819	0.0295
t=12	0.5539	0.0196	0.0404	0.5713	0.5716	0.017	0.0302	0.0132	0.0137	0.0273	0.0273	0.0066

Table 4: Different censoring schemes (CS) considered for simulation study.

n	m	R_1	R_2	R_3	R_4
50	20	0*5,2*15	0*5,4*7,2*1,0*7	0*1,6*2,0*2,2*5,4*2,0*8	0*1,6*2,0*2,2*4,0*6,2*5
	30	0*20,2*10	0*10,2*10,0*10	4*5,0*25	2*5.0*10,2*5
	32	2*8, 0*23, 2*1	1*18, 0*14	3*6, 0*26	0*23, 1*3, 4*3, 1*3
	40	0*35,2*5	1*10,0*30	2*5,0*35	0*10,2*5,0*25

Table 5: Simulation study of estimated values of θ for varying sample sizes under different estimation methods

		R_1	R_2	R_3	R_4
m=20	$\hat{\theta}_{ML}$	2.8684	3.3939	3.1899	2.692
	$\hat{\theta}_{BS}$	1.8727	2.0429	1.9924	1.8326
	$\hat{\theta}_{BL}$	1.8683	2.0375	1.9871	1.8284
m=30	$\hat{\theta}_{ML}$	2.3942	2.827	2.4023	2.3008
	$\hat{\theta}_{BS}$	1.7508	1.9373	1.7648	1.7136
	$\hat{\theta}_{BL}$	1.7472	1.9328	1.761	1.7101
m=32	$\hat{\theta}_{ML}$	2.3779	2.5968	2.3734	2.3533
	$\hat{\theta}_{BS}$	1.761	1.8608	1.7586	1.7371
	$\hat{\theta}_{BL}$	1.7573	1.8567	1.7548	1.7336
m=40	$\hat{\theta}_{ML}$	2.1382	2.2624	2.2165	2.3674
	$\hat{\theta}_{BS}$	1.6555	1.7226	1.6972	1.7817
	$\hat{\theta}_{BL}$	1.6525	1.7193	1.6939	1.7781

Table 6: Simulation study of estimated values of β for varying sample sizes under different estimation methods

		R_1	R_2	R_3	R_4
m=20	$\hat{\beta}_{ML}$	2.7385	2.5188	2.5785	2.6505
	$\hat{\beta}_{BS}$	1.6956	1.5711	1.6714	1.7423
	$\hat{\beta}_{BL}$	1.6902	1.5666	1.6666	1.7369
m=30	$\hat{\beta}_{ML}$	2.4031	2.4664	2.3547	2.3168
	$\hat{\beta}_{BS}$	1.7898	1.8201	1.831	1.7793
	$\hat{\beta}_{BL}$	1.7861	1.8164	1.8273	1.7757
m=32	$\hat{\beta}_{ML}$	2.3483	2.4667	2.3656	2.3701
	$\hat{\beta}_{BS}$	1.8391	1.8957	1.8581	1.8
	$\hat{\beta}_{BL}$	1.8357	1.8921	1.8546	1.7965
m=40	$\hat{\beta}_{ML}$	2.159	2.2993	2.214	2.3785
	$\hat{\beta}_{BS}$	1.7642	1.879	1.8202	1.9276
	$\hat{\beta}_{BL}$	1.7615	1.8761	1.8174	1.9246

Table 7: Estimated values of survival and hazard rate functions at specific time $t = 1$ when true values of parameters considered $\theta = 2$ and $\beta = 2$

m	RS	$\hat{R}_{ML}$	$\hat{R}_{BS}$	$\hat{R}_{BL}$	$\hat{h}_{ML}$	$\hat{h}_{BS}$	$\hat{h}_{BL}$
20	R_1	0.9432	0.8463	0.8456	0.473	0.5767	0.5765
	R_2	0.9664	0.8703	0.8696	0.297	0.4781	0.4785
	R_3	0.9588	0.8636	0.8629	0.3532	0.5258	0.5261
	R_4	0.9323	0.84	0.8393	0.5185	0.6081	0.6079
30	R_1	0.9088	0.8264	0.8257	0.5777	0.6585	0.6585
	R_2	0.9408	0.8559	0.8553	0.4387	0.5936	0.5942
	R_3	0.9095	0.8288	0.8281	0.5629	0.6675	0.6679
	R_4	0.8998	0.8198	0.8192	0.5935	0.6703	0.6704
32	R_1	0.9073	0.8281	0.8275	0.5708	0.6722	0.6725
	R_2	0.9255	0.8445	0.8438	0.5157	0.6498	0.6503
	R_3	0.9068	0.8277	0.8271	0.5768	0.6802	0.6805
	R_4	0.9049	0.824	0.8234	0.5859	0.668	0.6682
40	R_1	0.8821	0.809	0.8084	0.6168	0.6895	0.6898
	R_2	0.8959	0.8214	0.8208	0.6044	0.7038	0.7042
	R_3	0.891	0.8168	0.8162	0.6003	0.6929	0.6933
	R_4	0.9063	0.8316	0.831	0.5823	0.6953	0.6958

Table 8: Risk of the estimated values for the parameters θ & β under different estimation methods when their values are $\theta = 2$ and $\beta = 2$

m	RS	$risk(\hat{\theta}_{ML})$	$risk(\hat{\theta}_{BS})$	$risk(\hat{\theta}_{BL})$	$risk(\hat{\beta}_{ML})$	$risk(\hat{\beta}_{BS})$	$risk(\hat{\beta}_{BL})$
20	R_1	1.2831	0.0551	0.0003	0.7833	0.132	0.0007
	R_2	2.6207	0.0424	0.0002	0.4964	0.2356	0.0012
	R_3	2.0357	0.0468	0.0002	0.5653	0.1625	0.0008
	R_4	0.8551	0.0673	0.0003	0.6019	0.1057	0.0005
30	R_1	0.3226	0.0906	0.0005	0.2521	0.0759	0.0004
	R_2	0.9662	0.0438	0.0002	0.3484	0.0814	0.0004
	R_3	0.3582	0.0945	0.0005	0.2091	0.0647	0.0003
	R_4	0.2453	0.1122	0.0006	0.185	0.0816	0.0004
32	R_1	0.3146	0.0916	0.0005	0.1971	0.059	0.0003
	R_2	0.5852	0.0623	0.0003	0.3152	0.0532	0.0003
	R_3	0.318	0.0957	0.0005	0.2143	0.0567	0.0003
	R_4	0.2893	0.0988	0.0005	0.2212	0.0716	0.0004
40	R_1	0.1062	0.1382	0.0007	0.0751	0.0805	0.0004
	R_2	0.1979	0.1065	0.0005	0.1527	0.0474	0.0002
	R_3	0.1597	0.1197	0.0006	0.0999	0.0599	0.0003
	R_4	0.2713	0.0784	0.0004	0.2058	0.039	0.0002

Table 9: CIs for estimated values of θ under different estimation methods

m	RS	Boot-p	Boot-t	HPD
20	R_1	(2.8578, 11.1488)	(2.6288, 6.1536)	(1.3261, 2.4467)
	R_2	(3.7472, 18.0335)	(3.7067, 10.1156)	(1.4354, 2.6929)
	R_3	(3.4901, 13.8735)	(3.3244, 8.5)	(1.3958, 2.6221)
	R_4	(2.7133, 8.034)	(2.366, 5.0645)	(1.2945, 2.3969)
30	R_1	(2.0618, 5.1748)	(1.8705, 3.9422)	(1.2531, 2.2653)
	R_2	(2.7957, 8.0421)	(2.6447, 6.0912)	(1.3757, 2.5225)
	R_3	(2.2813, 4.6984)	(1.8374, 3.4712)	(1.2465, 2.3016)
	R_4	(2.042, 4.5805)	(1.7029, 3.3854)	(1.22, 2.2205)
32	R_1	(2.1664, 4.5572)	(1.824, 3.4918)	(1.2536, 2.2879)
	R_2	(2.4742, 5.755)	(2.2791, 4.7645)	(1.3244, 2.4148)
	R_3	(2.2539, 4.5638)	(1.858, 3.504)	(1.2478, 2.2867)
	R_4	(1.9942, 4.8306)	(1.8056, 3.7792)	(1.2455, 2.2427)
40	R_1	(1.7312, 3.5425)	(1.5848, 3.0472)	(1.1956, 2.1346)
	R_2	(1.9907, 3.9013)	(1.7448, 3.2471)	(1.235, 2.2288)
	R_3	(1.9355, 3.6744)	(1.6093, 2.973)	(1.2116, 2.1957)
	R_4	(2.0731, 4.3427)	(1.9664, 3.8268)	(1.2814, 2.3017)

Table 10: CIs for estimated values of β under different estimation methods

m	RS	Boot-p	Boot-t	HPD
20	R_1	(3.1851, 6.4733)	(2.077, 4.0564)	(1.0733, 2.3297)
	R_2	(2.3496, 4.8528)	(1.9902, 3.9582)	(1.0065, 2.1425)
	R_3	(2.218, 4.4424)	(2.379, 4.5623)	(1.0882, 2.2594)
	R_4	(2.6085, 5.1819)	(2.2945, 4.4122)	(1.126, 2.3731)
30	R_1	(2.8568, 4.9921)	(1.6458, 2.892)	(1.2644, 2.3197)
	R_2	(2.4732, 4.1915)	(2.1725, 3.5995)	(1.2992, 2.3392)
	R_3	(1.6349, 2.9493)	(2.5221, 4.5128)	(1.3172, 2.349)
	R_4	(2.3164, 3.9038)	(1.7931, 3.0022)	(1.2702, 2.2951)
32	R_1	(1.9933, 3.2423)	(2.3841, 3.8558)	(1.3405, 2.3472)
	R_2	(2.2417, 3.4711)	(2.6081, 3.9883)	(1.3807, 2.4114)
	R_3	(1.7294, 2.9539)	(2.5262, 4.2396)	(1.3539, 2.3694)
	R_4	(2.728, 4.793)	(1.5904, 2.7698)	(1.2914, 2.3134)
40	R_1	(2.3153, 3.6335)	(1.4744, 2.3389)	(1.3206, 2.2107)
	R_2	(1.9488, 2.9533)	(2.2802, 3.448)	(1.4162, 2.3473)
	R_3	(1.7508, 2.7971)	(2.0975, 3.3296)	(1.3706, 2.2749)
	R_4	(2.1957, 3.3477)	(2.3853, 3.5826)	(1.4553, 2.4036)

Table 11: Estimates of parameters (θ, β) for various choice of parameters under different estimation methods for $m = 32$ failures

θ	β	RS	$\hat{\theta}_{ML}$	$\hat{\theta}_{BS}$	$\hat{\theta}_{BL}$	$\hat{\beta}_{ML}$	$\hat{\beta}_{BS}$	$\hat{\beta}_{BL}$
1.5	2	R_1	1.688	1.4188	1.4165	0.1152	0.0381	0.0002
	2	R_2	1.8131	1.4962	1.4936	0.2111	0.0397	0.0002
	2	R_3	1.6819	1.412	1.4097	0.1115	0.0382	0.0002
	2	R_4	1.6885	1.4149	1.4127	0.119	0.0376	0.0002
3	2	R_1	3.8175	2.2328	2.2261	1.2067	0.6187	0.0031
	2	R_2	4.2918	2.3469	2.3394	2.4905	0.4562	0.0023
	2	R_3	3.8472	2.2421	2.2354	1.3377	0.6097	0.003
	2	R_4	3.7751	2.1794	2.173	1.1533	0.6995	0.0034
2	0.5	R_1	2.3617	1.8129	1.809	0.3032	0.0889	0.0004
	0.5	R_2	2.562	1.9253	1.9209	0.5415	0.0703	0.0003
	0.5	R_3	2.3462	1.7956	1.7916	0.3092	0.1006	0.0005
	0.5	R_4	2.3561	1.8219	1.8181	0.2799	0.0775	0.0004
2	1.5	R_1	2.3704	1.7738	1.7701	0.3032	0.0868	0.0004
	1.5	R_2	2.5964	1.8849	1.8807	0.5864	0.0609	0.0003
	1.5	R_3	2.3752	1.7754	1.7715	0.3486	0.1012	0.0005
	1.5	R_4	2.3506	1.7606	1.7571	0.3006	0.0946	0.0005

Table 12: Risk of the estimated values for the parameters (θ, β) for various choice of parameters under different estimation methods for $m = 32$ failures

θ	β	RS	$risk(\hat{\theta}_{ML})$	$risk(\hat{\theta}_{BS})$	$risk(\hat{\theta}_{BL})$	$risk(\hat{\beta}_{ML})$	$risk(\hat{\beta}_{BS})$	$risk(\hat{\beta}_{BL})$
1.5	2	R_1	0.1152	0.0381	0.0002	0.1984	0.0463	0.0002
	2	R_2	0.2111	0.0397	0.0002	0.2992	0.0477	0.0002
	2	R_3	0.1115	0.0382	0.0002	0.1912	0.0507	0.0003
	2	R_4	0.119	0.0376	0.0002	0.2282	0.051	0.0003
3	2	R_1	1.2067	0.6187	0.0031	0.2047	0.129	0.0006
	2	R_2	2.4905	0.4562	0.0023	0.3155	0.1182	0.0006
	2	R_3	1.3377	0.6097	0.003	0.1924	0.1377	0.0007
	2	R_4	1.1533	0.6995	0.0034	0.2064	0.1716	0.0009
2	0.5	R_1	0.3032	0.0889	0.0004	0.0119	0.0045	2.00E-05
	0.5	R_2	0.5415	0.0703	0.0003	0.0197	0.0067	3.00E-05
	0.5	R_3	0.3092	0.1006	0.0005	0.0121	0.0048	2.00E-05
	0.5	R_4	0.2799	0.0775	0.0004	0.0133	0.0041	2.00E-05
2	1.5	R_1	0.3032	0.0868	0.0004	0.1142	0.0262	0.0001
	1.5	R_2	0.5864	0.0609	0.0003	0.1638	0.027	0.0001
	1.5	R_3	0.3486	0.1012	0.0005	0.1198	0.0272	0.0001
	1.5	R_4	0.3006	0.0946	0.0005	0.1189	0.0296	0.0001

Table 13: Estimated values of survival and hazard rate functions at specific time $t = 1$ for various choice of parameters and for $m = 32$ failures

θ	β	RS	$\hat{R}_{ML}$	$\hat{R}_{BS}$	$\hat{R}_{BL}$	$\hat{h}_{ML}$	$\hat{h}_{BS}$	$\hat{h}_{BL}$
1.5	2	R_1	0.8151	0.758	0.7574	0.9002	0.8769	0.8764
	2	R_2	0.8369	0.776	0.7754	0.8692	0.8645	0.8642
	2	R_3	0.814	0.7563	0.7558	0.8961	0.8778	0.8773
	2	R_4	0.8152	0.7571	0.7565	0.9094	0.8629	0.8623
3	2	R_1	0.978	0.8928	0.8921	0.2021	0.4485	0.4498
	2	R_2	0.9863	0.9043	0.9036	0.1465	0.4209	0.4223
	2	R_3	0.9787	0.8938	0.8931	0.1954	0.4425	0.4437
	2	R_4	0.9771	0.8869	0.8862	0.2081	0.4475	0.4487
2	0.5	R_1	0.9057	0.8368	0.8362	0.1435	0.1877	0.188
	0.5	R_2	0.9228	0.8542	0.8535	0.1319	0.1812	0.1816
	0.5	R_3	0.9043	0.834	0.8333	0.1448	0.1909	0.1912
	0.5	R_4	0.9052	0.8383	0.8377	0.1453	0.1834	0.1838
2	1.5	R_1	0.9066	0.8303	0.8297	0.4314	0.5262	0.5267
	1.5	R_2	0.9255	0.8482	0.8475	0.3841	0.5006	0.5013
	1.5	R_3	0.907	0.8306	0.8299	0.4314	0.5286	0.5291
	1.5	R_4	0.9047	0.8281	0.8275	0.4387	0.5169	0.5172

REFERENCES

Aggarwala, R. and Balakrishnan, N. (1998). Some properties of progressive censored order statistics from arbitrary and uniform distributions with applications to inference and simulation. *Journal of Statistical Planning and Inference*, 70(1):35–49.

Ahmad, A., Ahmad, S. and Ahmed, A. (2014). Bayesian estimation of exponentiated inverted Weibull distribution under asymmetric loss functions. *Journal of Statistics Applications and Probability*, 4(1):183–192.

Almetwally, E. M., Jawa, T. M., Sayed-Ahmed, N., Park, C., Zakarya, M. and Dey, S. (2023). Analysis of unit-Weibull based on progressive type-ii censored with optimal scheme. *Alexandria Engineering Journal*, 63:321–338.

Balakrishnan, N., Kannan, N., Lin, C.-T., and Ng, H. K. T. (2003). Point and interval estimation for Gaussian distribution, based on progressively type-ii censored samples. *IEEE Transactions on Reliability*, 52(1):90–95.

Balakrishnan, N., Kundu, D., Ng, K. T. and Kannan, N. (2007). Point and interval estimation for a simple step-stress model with type-ii censoring. *Journal of Quality Technology*, 39(1):35–47.

Balakrishnan, N. and Sandhu, R. (1995). A simple simulation algorithm for generating progressive type ii censored samples. *The American Statistician*, 49(2):229–230.

Berger, J. O. and Sun, D. (1993). Bayesian analysis for the poly-Weibull distribution. *Journal of the American Statistical Association*, 88(424):1412–1418.

Box, G. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Massachusetts.

Chen, M. H. and Shao, Q. M. (1999). Monte Carlo estimation of Bayesian credible and hpd intervals. *Journal of Computational and Graphical Statistics*, 8(1):69–92.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2012). *Monte Carlo Methods in Bayesian Computation*. Springer Science & Business Media.

Cohen, A. (1963). Progressively censored samples in life testing. *Technometrics*, 5(3):327–339.

Dagpunar, J. S. (2007). *Simulation and Monte Carlo: With Applications in Finance and MCMC*. John Wiley & Sons.

Dey, A. K. and Kundu, D. (2012). Discriminating between the Weibull and lognormal distributions for type-ii censored data. *Statistics*, 46(2):197–214.

Dey, S., Dey, T. and Luckett, D. J. (2016). Statistical inference for the generalized inverted exponential distribution based on upper record values. *Mathematics and Computers in Simulation*, 120:64–78.

DiCiccio, T. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228.

Dolas, D., Jaybhaye, M. and Deshmukh, S. (2014). Estimation the system reliability using Weibull distribution. *International Proceedings of Economics Development and Research*, 75(29):144–148.

Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM

Efron, B. (1992). *Bootstrap Methods: Another Look at the Jackknife*. Springer.

El-Din, M. M. and Shafay, A. R. (2013). One-and two-sample Bayesian prediction intervals based on progressively type-ii censored data. *Statistical Papers*, 54(2):287–307.

El-Sherpieny, E.-S. A., Almetwally, E. M. and Muhammed, H. Z. (2022). Bivariate Weibull-g family based on copula function: Properties, Bayesian and non-Bayesian estimation and applications. *Statistics, Optimization & Information Computing*, 10(3):678–709.

Flaih, A., Elsalloukh, H., Mendi, E. and Milanova, M. (2012). The exponentiated inverted Weibull distribution. *Applied Mathematics and Information Sciences*, 6(2):167–171.

Gelfand, A. E. (1996). Model determination using sampling-based methods. *Markov Chain Monte Carlo in Practice*, pages 145–161.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A, and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. CRC press.

Goyal, T., Rai, P. K. and Maurya, S. K. (2019). Classical and Bayesian studies for a new lifetime model in presence of type-ii censoring. *Communications for Statistical Applications and Methods*, 26(4):385–410.

Hall, P. (1988). Theoretical comparision of bootstrap confidence intervals. *The Annals of Statistics*, 16(3):927–953.

Han, D. and Kundu, D. (2015). Inference for a step-stress model with competing risks for failure from the generalized exponential distribution under type-ii censoring. *IEEE Transactions on Reliability*, 64(1):31–43.

Hastings, W. K. (1970). Montecarlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Herd, R. (1956). Estimation of the parameters of a population from a multi-censored sample. *Ph.D. Thesis, Iowa State College, Ames, Iowa*.

Jiang, R. and Murthy, D. (1999). The exponentiated Weibull family: A graphical approach. *IEEE Transactions on Reliability*, 48(1):68–72.

Joarder, A., Krishna, H. and Kundu, D. (2011). Inferences on Weibull parameters with conventional type-i censoring. *Computational Statistics & Data Analysis*, 55(1):1–11.

Kaushik, A., Singh, U. and Singh, S. K. (2017). Bayesian inference for the parameters of Weibull distribution under progressive type-i interval censored data with beta-binomial removals. *Communications in Statistics-Simulation and Computation*, 46(4):3140–3158.

Kundu, D. (2008). Bayesian inference and life testing plan for the Weibull distribution in presence of progressive censoring. *Technometrics*, 50(2):144–154.

Kundu, D. and Biswabrata, P. (2009). Bayesian inference and life testing plans for generalized exponential distribution. *Science in China Series A: Mathematics*, 52(6):1373–1388.

Kundu, D. and Howlader, H. (2010a). Bayesian inference and prediction of the inverse Weibull distribution for type-ii censored data. *Computational Statistics and Data Analysis*, 54(6):1547–1558.

Kundu, D. and Howlader, H. (2010b). Bayesian inference and prediction of the inverse Weibull distribution for type-ii censored data. *Computational Statistics & Data Analysis*, 54(6):1547–1558.

Lee, E. T. and Wang, J. (2003). *Statistical Methods for Survival Data Analysis*, volume 476. John Wiley & Sons.

Marin, J.-M. and Robert, C. (2007). *Bayesian Core: a Practical Approach to Computational Bayesian Statistics*. Springer Science & Business Media.

Maurya, S. K., Kaushik, A., Singh, S. K. and Singh, U. (2017). A new class of exponential transformed Lindley distribution and its application to yarn data. *International Journal of Statistics and Economics*, 18(2):135–151.

Mudholkar, G. S. and Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, 42(2):299–302.

Ng, H., Kundu, D., and Balakrishnan, N. (2006). Point and interval estimation for the two-parameter Birnbaum–Saunders distribution based on type-ii censored samples. *Computational Statistics & Data Analysis*, 50(11):3222–3242.

Ng, H. K. T. and Chan, P. S. (2007). Comments on Progressive censoring methodology: An appraisal. *Test*, 16(2):287–289.

Ntzoufras, I. (2011). *Bayesian Modeling Using WinBUGS*, volume 698. John Wiley & Sons.

Prajapati, D., Mitra, S. and Kundu, D. (2020). A new decision theoretic sampling plan for exponential distribution under type-i censoring. *Communications in Statistics-Simulation and Computation*, 49(2):453–471.

Raqab, M. Z., Asgharzadeh, A. and Valiollahi, R. (2010). Prediction for Pareto distribution based on progressively type-ii censored samples. *Computational Statistics & Data Analysis*, 54(7):1732–1743.

Raqab, M. Z. and Madi, M. T. (2005). Bayesian inference for the generalized exponential distribution. *Journal of Statistical computation and Simulation*, 75(10):841–852.

Robert, C. and Casella, G. (2013). *Monte Carlo Statistical methods*. Springer Science & Business Media.

Singh, S. K., Singh, U. and Kumar, M. (2013). Estimation of parameters of exponentiated Pareto model for progressive type-ii censored data with binomial removals using Markov chain Monte Carlo method. *International Journal of Mathematics & Computation*, 21(4):88–102.

Singh, S. K., Singh, U. and Kumar, M. (2016). Bayesian estimation for Poisson-exponential model under progressive type-ii censoring data with binomial removal and its application to ovarian cancer data. *Communications in Statistics-Simulation and Computation*, 45(9):3457–3475.

Singh, U., Gupta, P. K. and Upadhyay, S. K. (2002). Estimation of exponentiated Weibull shape parameters under linex loss function. *Communications in Statistics - Simulation and Computation*, 31(4):523–537.

Singh, U., Gupta, P. K. 2and Upadhyay, S. K. (2005). Estimation of parameters for exponentiated Weibull family under type-ii censoring scheme. *Computational Statistics & Data Analysis*, 48(3):509–523.

Viveros, R. and Balakrishnan, N. (1994). Interval estimation of parameters of life from progressively censored data. *Technometrics*, 36(1):84–91.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Communications in Statistics - Simulation and Computation*, 81(394):446–451.