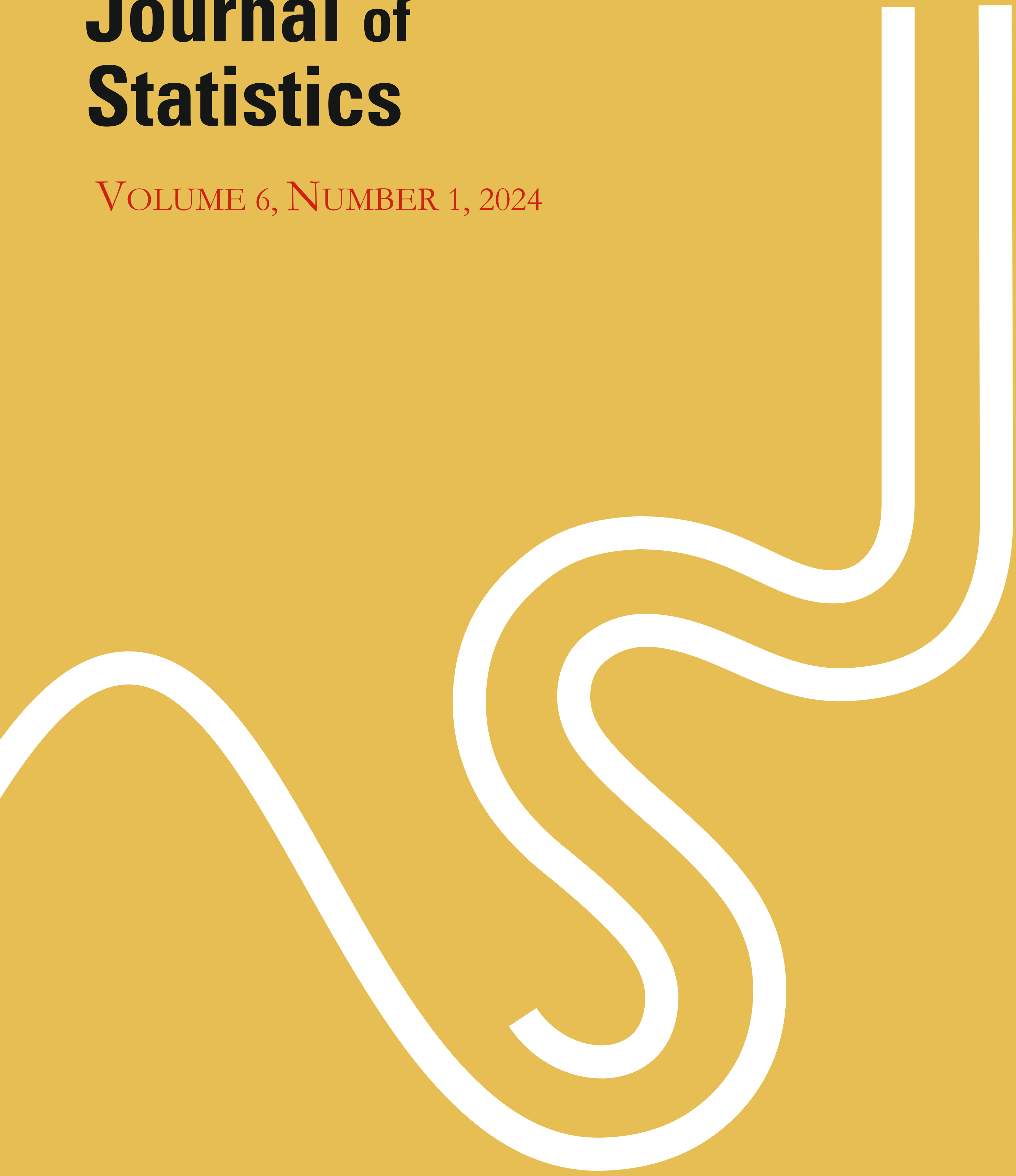
Spanish Journal of Statistics

Volume 6, Number 1, 2024



NIPO: 222-24-003-9
ISSN: 2695 - 9070
DOI: https://doi.org//10.37830/SJS.2024.1

Published by the Spanish National Statistical Institute

SPANISH JOURNAL OF STATISTICS

VOLUME 6, NUMBER 1, 2024

Contents

Spanish Journal of Statistics
Vol. 6 No. 1, 2024, Pages 5 – 6
doi: https://doi.org/10.37830/SJS.2024.1.01

Editorial

Presentation of Volume 6, 1, 2024

José María Sarabia

Editor-in-Chief Spanish Journal of Statistics

Dear readers and dear members of the statistical community,

It is a pleasure for me to present Volume 6, 1 corresponding to the year 2024. This volume is composed of three articles: one article in the official statistics section and two articles in the general statistics section.

The first article is entitled: “Measuring tourism using mobile network data" and its authors are Belén González Olmos and María Velasco Gimeno, from the Spanish Statistical Office, INE. In Spain, basic tourism statistics are the responsibility of the INE and are traditionally based on surveys. In recent years, due to the challenges associated with collecting data from individuals, especially during the COVID-19 pandemic, national statistical offices have explored access to data generated by the private sector using two different approaches: based on a specific agreement or taking advantage of a regulatory framework. In this article, the Spanish experience in using mobile phone positioning data is explained. It is important to highlight that the use of this source of information allows obtaining new products with a granularity in terms of origin/destination of tourists that would be impossible to achieve using traditional techniques, without increasing the cost of the statistics and the burden on the informant. The results obtained are published as experimental statistics, but the final objective is to integrate them with traditional tourism surveys.

The next two papers are presented in the general section. The second paper is titled, “ĂIJFinding most nearly compatible conditionals under a finite discrete set-up: An overview and recent developments" by Indranil Ghosh, University of North Carolina, USA. The paper is devoted to the topic of conditional specification of discrete distributions. When modeling complicated real-life scenarios, one of the goals is to capture the observed dependence. The paper provides an overview of a variety of divergence measures including, but not limited to, the Kullback-Leibler divergence measure, the power divergence statistic, the Hellinger distance along with some recently developed divergence measures and their role in addressing various compatible conditions in order to find the most compatible one for a finite discrete case, and also in identifying compatibility under conditional and marginal information under some additional information in the form of marginal and/or conditional summary. The author provides some numerical examples to illustrate each of the scenarios.

The third paper is titled, “Census-based comparability of data on literacy processes in western Europe, by José Manuel Gutiérrez, from Universidad de Salamanca. The author presents a comparative picture of the literacy processes in Western Europe on the eve of and during the Second

Industrial Revolution, taking censual literacy rates as a yardstick to measure and compare literacy in different countries. Censual data are obtained and analysed from the original source. If only partial or insufficient censual data are available, literacy is assessed as if given by full censual data. A set of comparable literacy data is built. Four literacy groups result. The area of Western Europe where mass literacy was first achieved was the German-speaking or culturally highly Germanised zone. Britain and Sweden turn out to be in the same cluster as France. The periphery of Western Europe shows the well-known pattern of delayed literacy development.

Finally, I would like to thank all the authors of this volume for choosing our journal as a means of disseminating their research. I appreciate the work of the editors and reviewers, who contribute to maintaining a high standard of scientific quality.

SPANISH JOURNAL OF STATISTICS
Vol. **6** No. 1 2024, Pages 7–19
doi:https://doi.org/10.37830/SJS.2024.1.02

OFFICIAL STATISTICS

Measuring tourism using mobile network data

Belén González Olmos[1], María Velasco Gimeno[2]

[1] National Statistics Institute (Spain), belen.gonzalez.olmos@ine.es
[2]National Statistics Institute (Spain), maria.velasco.gimeno@ine.es

Received: February 29, 2024. Accepted: October 25, 2024.

Abstract: IIn Spain, Tourism basic statistics are responsibility of the National Statistical Institute (INE) and traditionally are based on surveys. In recent years, due to the challenges associated with the collection of data from persons, especially during the COVID-19 pandemic, national statistics offices have explored the access to data generated by private sector using two different approaches: based on either a specific arrangement or taking advantage of a regulatory framework. In this article, the Spanish experience in the use of mobile phones positioning data is explained.

INE uses mobile phone positioning data as auxiliary information to tourism surveys with the objective of improving the geographical breakdown of tourism figures. The use of this source of information allows obtaining new products with a granularity in terms of origin/destination of tourists that would be impossible to achieve using traditional techniques, without increasing the cost of the statistics and the burden on the informant. Nonetheless, it poses some challenges in terms of quality assurance and sustainable access. The results are published as experimental statistics, but the ultimate aim is to integrate them with traditional tourism surveys.

Keywords: tourism statistics, official statistics, Big Data, modernization, mobile network data, granularity, quality, integration of sources

MSC: 6201, 6211

1 Introduction

The production model of official statistics in the near future has to be adapted to the new situations of competition, data availability and user requirements. Improvements in the production model have a direct impact on improving the quality of statistical products of all kinds. Currently, the national statistical system is still mainly based on traditional surveys; however, the use of administrative registers and the use of Big Data in statistical production have been incorporated into the various operations for several years now.

Published by the Spanish National Statistical Institute

The use of new available data sources (private databases, Big Data, ...) in the compilation of official statistics is a path that has already been undertaken. The use of these new sources of information opens up new possibilities to compile statistics more quickly, with a greater geographical and functional disaggregation and to address the study of emerging phenomena in a shorter period of time. Moreover, they represent a fundamental way to reduce the response burden on informants, as they are based on existing and available information.

Given the need to study in depth the quality, procedures, applicable statistical techniques, etc., of these sources, the strategy followed by the European Statistical System (ESS), which has already been joined by different national statistical offices such as the INE, as well as Eurostat, is to disseminate these operations under the name of experimental statistics. The idea is to be able to test and try out these new data sources without the restrictions and limitations imposed by official statistics. In this way, the organization is acquiring the necessary knowledge to make these statistics official in the near future.

Experimental statistics use new data sources and methods in an effort to better respond to the needs of our users in a timely manner. The contents they present are considered experimental because they have not yet reached sufficient maturity in terms of reliability, stability or data quality to be included in official statistics. Nevertheless, the available results are offered to users for their use and evaluation, due to the relevance they may have and as a means to improve the products themselves by gathering the opinion of the final users of the information.

Until an experimental statistic has reached sufficient stability and maturity, it will not be proposed for inclusion as official, and therefore will not be included in the National Statistical Plan.

In the case of mobile network data, the main problem countries are encountering is access to those data. They are strongly protected by national and international legal regulations, due to highly sensitive issues of confidentiality and personal privacy. In addition, there are economic interests to commercially exploit them by the phone companies, so access to the data is the first major problem to be addressed. The objective is, in the short term, to have access to specific datasets for use in research activities and, in the long term, to investigate the feasibility of sustained access under standard production conditions, as well as the required characteristics of such data (what information it needs to contain).

In relation to access, there is no guiding principle or golden rule for success, as the situation is markedly different per country (different legal regulations, different contents in the datasets offered) and per mobile network operator (different company structures and different commercial interests), and the role that national data protection agencies can play is important.

The generation of mobile network data has one characteristic that specifically affects processing: the data are not generated with a specific metadata structure for statistical purposes. Moreover, it is characterized by the fact that the data does not contain information from the data provider but from third parties (the customers) and that it is information that plays a central role in the provider's business.

These data (as with many other Big Data sources) have been generated for purposes very different from statistical production, even before their potential use for statistical purposes has been identified. Therefore, there is no proper metadata structure included in the generation process.

The raw data are extremely technical and some of them are only stored temporarily, which makes pre-processing necessary to generate exploitable data for statistical purposes (microdata).

Statistical microdata can be analysed in many different ways and the purpose for which they are used will condition the aggregation process required. The experience described above has been oriented towards a particular type of aggregated data: those that provide counts of individuals from a given target population (general population, tourists, resident tourists, travelers, ...) by territorial cells and time interval.

Aggregate data must be somehow linked to the target population, i.e. an inference exercise must be carried out, given that NIS only access data from some telephone companies. The problem is that, because of the way the data have been generated and collected, there is a need for a new methodology because the traditional sample design methodology is no longer applicable.

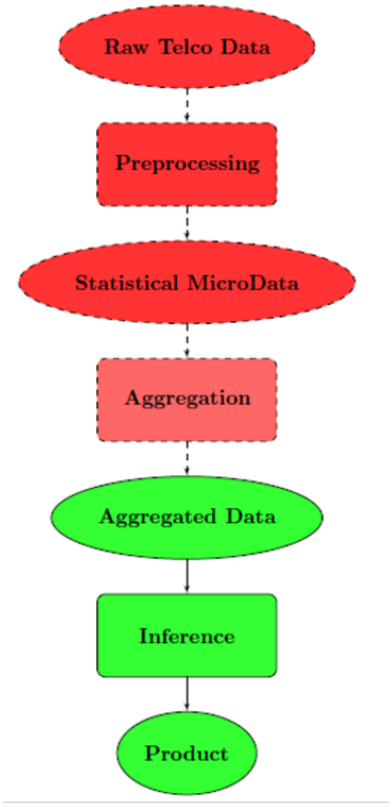


Figure 1: Mobile network data process in mobile phone companies.

The process can be represented as in the Figure 1 (the oval elements indicate data sets and the rectangular ones indicate steps in the process; in green the phases in which the National Statistical Institutes have access to the data -the three at the bottom- and in red those that are exclusively accessed by the source -the three at the top-, although the aggregation phase will depend on each specific case).

As can be seen, there is no access to the raw telecommunications data originating in the network, nor even to the pre-processing to produce statistical microdata or even to the next stage of aggregation.

The contents of this paper are the following. In Section 2 the experimental statistic to measure tourism flows with mobile phone positioning data is presented. Section 3 goes into more detail on methodological aspects such as the definitions and processing data. Section 4 presents how the results are disseminated in INE website with maps, and some examples of the usefulness of this information. Finally in Section 5 some conclusions are included.

2 Experimental statistics: Measurement of tourism using mobile network data

Since 2015, the Spanish National Statistics Institute (INE) has been responsible for the statistical operation Resident Tourism Survey (RTS) and the Border Tourism Movements Survey (FRONTUR). Both operations are based, in a large part of their elaboration, on surveys aimed at individuals and households. While the main objective of the former is to estimate the number of trips made by Spanish residents both within and outside Spain, that of the latter is to estimate trips made by non-residents in Spanish territory.

In both cases the cost of collecting questionnaires is high and the results obtained do not provide a good geographical granularity (Autonomous Community at most) because they do not have the necessary sample support. Regarding data availability, FRONTUR disseminates its results around one month after the end of the reference month. The ETR, on the other hand, publishes its results quarterly three months after the end of the reference quarter, which significantly reduces its timeliness. For example, data for the summer season are released around Christmas, three months after the end of the season.

Fortunately, technology is evolving, and many tools or devices have entered the daily lives of citizens. This fact, combined with the decreasing prices of using these devices and the continuous growth of the capacity to process and analyse the immense volumes of data (Big Data) is creating a whole new range of data sources that cannot be ignored by the official statistics system.

Due to this, INE, in contact with the three large mobile companies in Spain, has developed a project for the exploitation of aggregated mobile telephone data from which the movements of resident and foreign tourists can be known, breaking down the information by Autonomous Community, provinces and municipalities through which their trips take place. In the same way, the countries to which tourists resident in Spain travel to when they go abroad are also known.

The complexity of the data captured by an antenna requires specialised processing to transform them into a set of information valid for statistical processing. This process has been carried out by the three mobile telephone operators, through their algorithms. This work has required continuous interaction by the INE over two years with the three operators for the definition of the algorithms, the detailed analysis of the results, the detection of systematic differences and the drawing up of conclusions that have enabled the transformation of mobile data to be adapted to international definitions of tourism.

The improvements seen in this project compared to traditional surveys are:

- The pool of individuals available to the operators is much larger than the traditional survey samples. Considering that a large percentage of the population has a mobile phone and that the main operators have around 25% market share each, the sample of each of them can be around 3 orders of magnitude larger than that of the surveys.
- With a much larger number of individuals for whom travel information is available, the geographical breakdowns provided are much broader. Traditional surveys provide information for both resident tourism (ETR) and non-resident tourism (FRONTUR) at Autonomous Community level. With this new source, information is provided up to municipal level.
- In the case of resident tourism, the temporal availability improves considerably. The ETR publishes quarterly data three months after the end of the reference quarter, which means that the data for the first month of the published quarter is published with a time lag of 5 months. With this study, this information is published with only a one-month lag.

- With respect to outbound tourism, the ETR provides annual information for the four main countries receiving resident tourism, as well as for groupings of these countries. Through this study, monthly information is provided for all countries with a minimum number of trips.
- Regarding inbound tourism, FRONTUR provides monthly information for a few countries individually and for groupings of these. With this study, monthly information is given for practically all countries (if information is available for a minimum number of trips).
- Subjectivity and errors that may be introduced by informants (actively or passively) when providing information in the questionnaire are eliminated.

On the other hand, as mentioned above, this data source does not allow us to obtain qualitative variables such as type of accommodation used, reason for the trip, form of organization or mode of booking, so it is still necessary to maintain some kind of field operation to obtain them.

Due to the wide coverage of this project, which includes domestic, inbound and outbound tourism, 3 independent experimental statistics are carried out. The first publication was made in May 2022, including data from July 2019. Since then, monthly publications are made, around 35 days after the reference month.

- Inbound tourism (Spain, 2022b): The number of trips, overnight stays and the corresponding average stay of tourists coming to Spain from any country are published. They are provided by municipalities, provinces and Autonomous Communities. This is a great improvement in terms of the geographical breakdown provided by FRONTUR.
- Outbound tourism (Spain, 2022c): The number of trips made by residents in Spain to foreign countries is published, as well as the associated overnight stays and average duration. This is broken down by municipalities, provinces and Autonomous Communities of residence, and for all countries to which residents have made trips. It offers an improvement over the ETR in terms of both geographical breakdown and timeliness.
- Domestic tourism (Spain, 2022a): This is published for residents in Spain, the number of trips outside the province of residence, as well as their overnight stays and average duration. The geographical breakdown level, both by origin and destination, is Autonomous Community, province and municipality. As with outbound tourism, it offers an improvement over the ETR in terms of both geographical breakdown and timeliness.

The study period for these three operations is the month. In the first dissemination, the months from July 2019 to April 2022 were published. Since then, they have been updated in monthly basis.

In parallel to these publications, INE is working to integrate this information with the ETR and FRONTUR, which would be the final objective of these experimental statistics. In this way, it will be possible to provide leading indicators that will make it possible to provide travel information much earlier than in the current deadlines. Special improvement will be obtained in the data on domestic and foreign trips (ETR) as the information will be published in less than a month and without waiting the current 3-5 months (depending on the month) for the information to be available. On the other hand, this integration will bring granularity to the surveys, as well as a reduction in the size of the samples and questionnaires, which will reduce the burden on the respondent, as well as the cost of the operations.

3 Methodology

3.1 Adjustment of definitions

For the development of this project, the definitions and concepts of traditional surveys, which follow international methodologies and standards, have been adjusted and adapted to the information available to mobile companies. By way of example, the adaptation of the main concepts is shown below:

- Trip in the field of tourism statistics (UNWTO, 2008): Tourism trips are all trips with a main destination outside the person's place of usual residence, involving an overnight stay outside the place of usual residence and lasting less than one year, provided that the main purpose of the trip, including business, leisure or other personal reasons, is other than employment in a company established in the place visited. They are outward and return journeys and end when the person returns to his/her place of usual residence.
- Journey adapted to mobile network data: A tourism trip is considered to have taken place when a mobile phone has been detected for a longer period between 22:00 and 06:00 in a municipality or country other than that of usual residence and, in addition, has also been captured the following day (from 06:00 onwards) in that municipality. The journey ends when the mobile phone is again detected for a longer period between 22:00 and 06:00 in the municipality or country of residence.
- Overnight stay in the field of tourism statistics: Number of consecutive nights that a person spends in a municipality or country other than that of residence as part of a trip.
- Overnight stay adapted to mobile network data: Number of consecutive nights in which a mobile phone has been detected longer between 22:00 and 06:00 in a municipality or country other than the municipality or country of residence and has also been captured the following day (from 06:00 onwards) in that municipality or country.
- Main destination of the trip in the field of tourism statistics: This is the place where the respondent has spent the greatest number of nights.
- Main destination adapted to mobile network data: This is the place where a mobile phone has made the most overnight stays as part of a trip (according to the definition).

3.2 Data collection and integration

INE receives the tabulated and aggregated data prepared by each company. It does not have individual device information, only receives the aggregated information provided by the mobile telephone operators. It receives both raw data and data aggregated to the population.

Depending on the type of tourism (inbound, outbound or domestic), the operators' files are processed and integrated in different ways. To mention one of them as example, the procedure for outbound tourism is detailed below.

The first step consists of processing the operators' files. The files sent to INE monthly contain information on trips and overnight stays abroad, by Autonomous Community, province and municipality of origin, made by resident tourists. INE carries out a prior filtering of these files, in order to suitably adapt their format before processing the information they contain.

The second step is the estimation of the totals by country. For this purpose, the trips to the corresponding country provided by the three operators in the Autonomous Community files are added up. Since the three operators do not cover 100% of mobile telephony users in Spain, correction factors are estimated to bring this sum to the total population. These factors vary according to the number of operators providing data for each country and are estimated by quarter using market share data from the CNMC (Spanish National Markets and Competition Commission). Similarly, total overnight stays are estimated for each country. Average durations are calculated as the quotient of trips and overnight stays for each country estimated independently.

Finally, these totals must be distributed by Autonomous Community, province and municipality of origin. This process is followed for the Autonomous Communities (ccaa, from now on):

- For each country, the percentage of trips (and overnight stays) from each autonomous community is determined for each of the operators.
- For each ccaa-country crossing, the average of the three percentages obtained in the previous point is calculated.
- The averages calculated are adjusted so that their total sum per country is 100%.
- These percentages are applied to the estimate of trips (and overnight stays) for each country, thus distributing them by ccaa.

The distribution of totals by province and municipality is done in a similar way.

More information on the different methodologies for each type of tourism (domestic, outbound and inbound) can be found in the technical projects on the website: https://www.ine.es/experimental/turismo_moviles/experimental_turismo_moviles.htm.

4 Dissemination of results

The variables published are the number of tourists and the overnight stays and average duration associated with their trips.

The geographical breakdown variables are countries, Autonomous Communities, provinces and municipalities. The time disaggregation variable is the month.

For the publication of the results, tables are used, where you can select the different variables that you want to consult, as well as infographics that allow you to select the countries, ccaa, provinces and municipalities on different maps, to evaluate the complete series or the data on trips in a specific month.

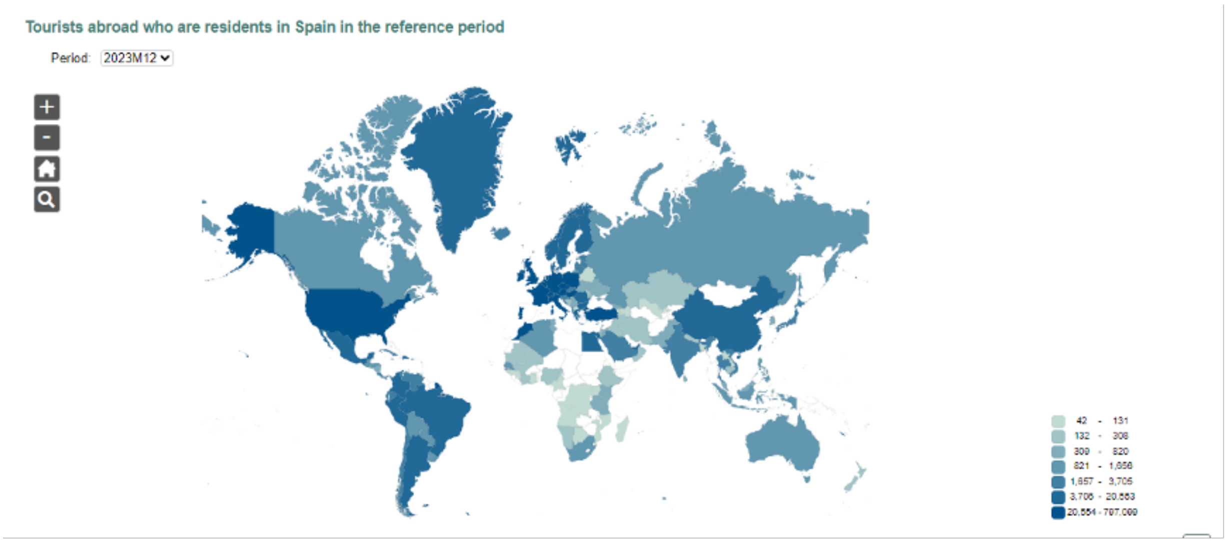
In the case of outbound tourism, the following maps are presented in the infographic:

- World map with trips in each month to each destination country (see Figure 2).
- Map and line graph for each continent (see Figure 3).
- Map and graph for Autonomous Communities, provinces and municipalities. When an area is selected, the 10 most visited countries are displayed in the graph (see Figure 4).

Similarly for inbound tourism, the following maps are available:

- World map with monthly trips from each country (see Figure 5):
- Map and line graph from each continent (see Figure 6).
- Map and graph for Autonomous Regions, provinces and municipalities. By selecting an area, the graph shows the 10 countries of origin with the highest number of tourists (see Figure 7).

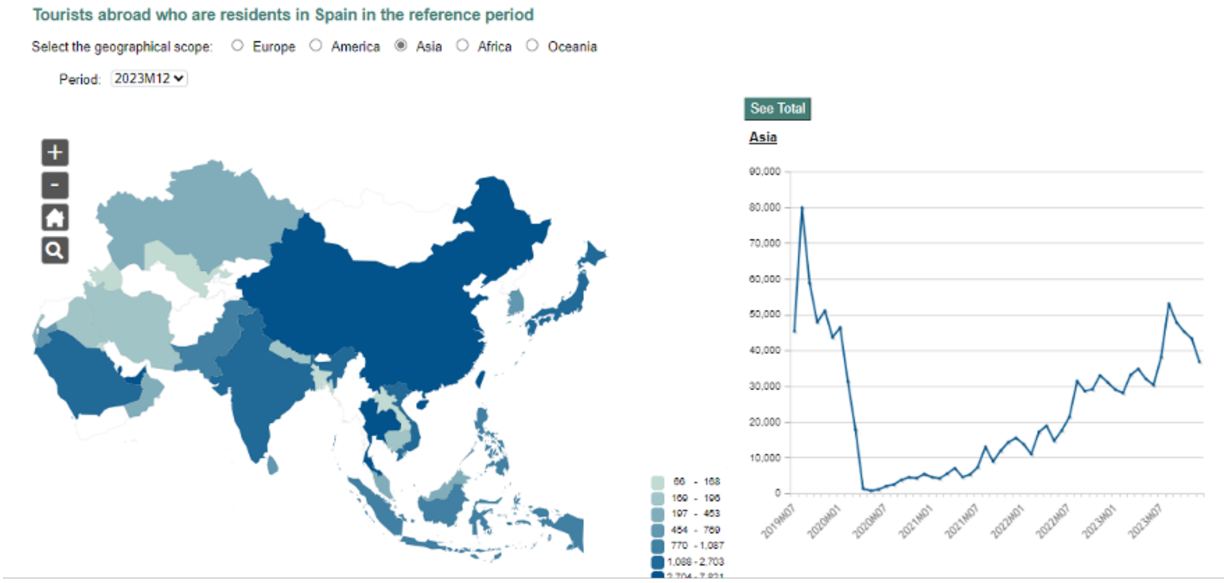
This dissemination is complemented with specific infographics for the annual data where, for example, the countries that receive the highest percentage of Spanish tourists in summer or winter (outbound tourism) can be visualized (see Figure 8) or the municipalities with the most tourists by country of origin (inbound tourism, see Figure 9)



Tourists abroad who are residents in Spain in the reference period

Period: 2023M12

Figure 2: World map for outbound tourism data



Tourists abroad who are residents in Spain in the reference period

Select the geographical scope: ○ Europe ○ America ● Asia ○ Africa ○ Oceania

Period: 2023M12

Figure 3: Outbound tourism data by destination continents.

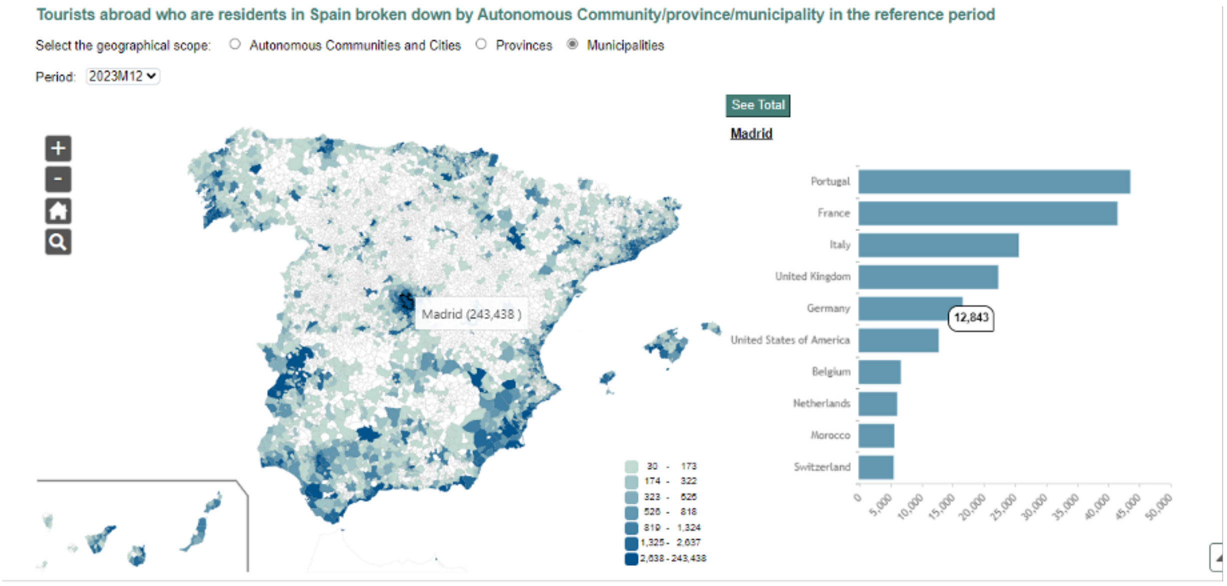
4.1 Cases of use

Two simple cases of us that can be carried out with the published data are shown below.

Objective: To analyse the evolution of the Asian countries most visited by residents in Spain in summer.

After downloading[1] the information on the number of monthly trips to the different countries on the Asian continent in the summer months (July, August and September) for the available years (2019-2023, see Figure 10).

[1] Available data: https://www.ine.es/dynt3/inebase/es/index.htm?padre=8578&capsel=8580

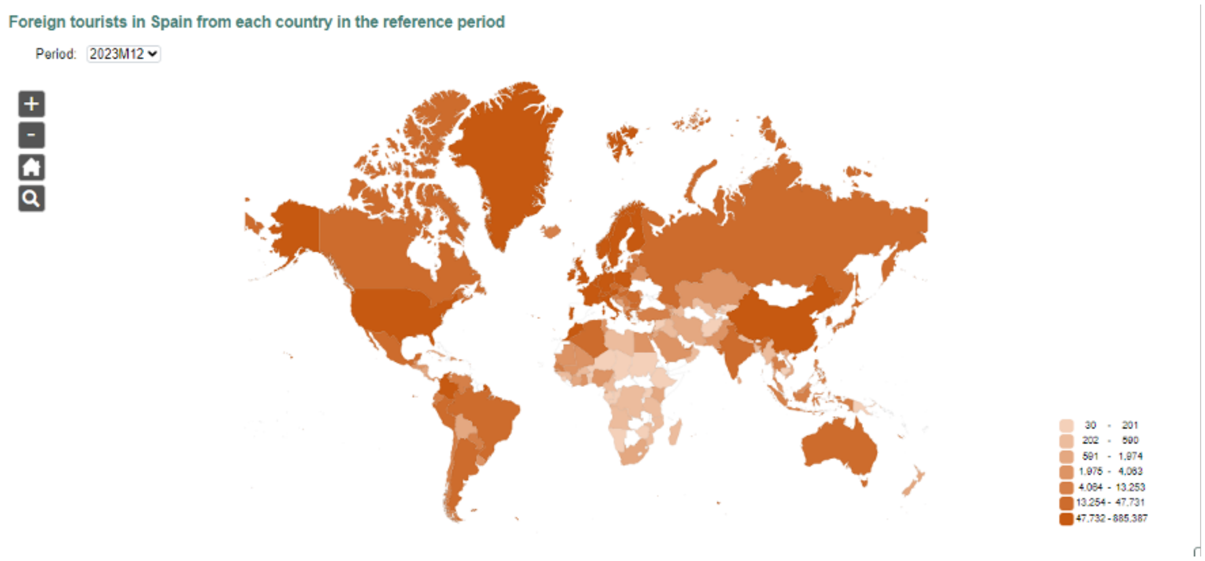


Tourists abroad who are residents in Spain broken down by Autonomous Community/province/municipality in the reference period

Select the geographical scope: ○ Autonomous Communities and Cities ○ Provinces ● Municipalities

Period: 2023M12

Figure 4: Outbound tourism data by origin of trips.



Foreign tourists in Spain from each country in the reference period

Period: 2023M12

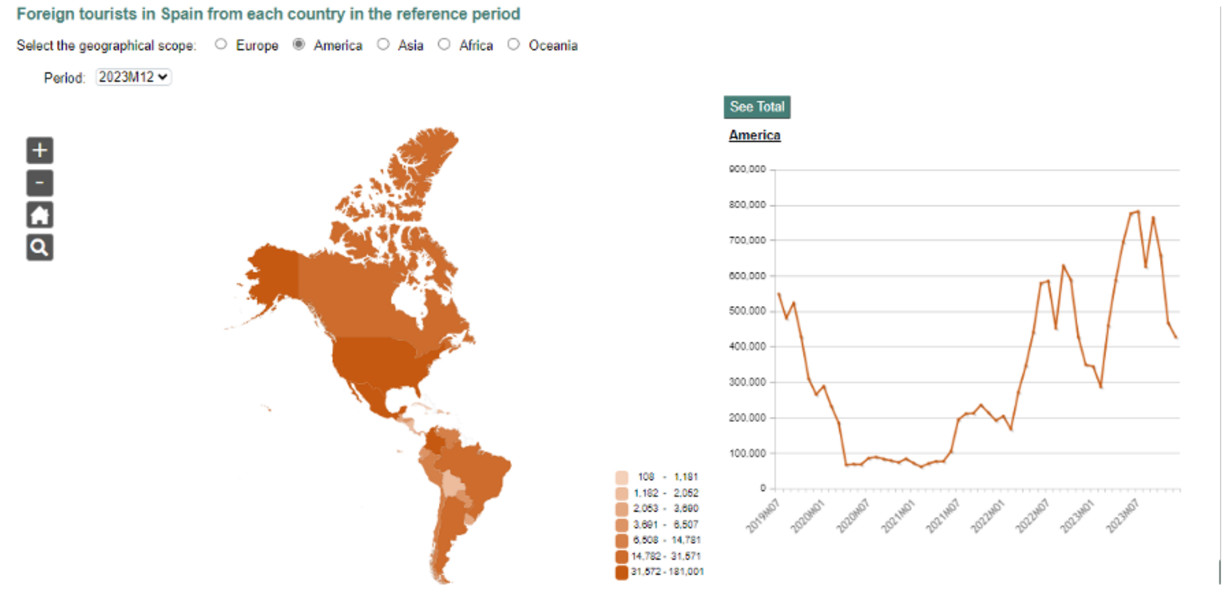
Figure 5: World map for inbound tourism data.

The cases where a "." appears are because there are less than 30 trips and are hidden for statistical secrecy. For simplicity, it is assumed that there have been no trips in these cases.

Aggregating the data for the three summer months of each year and sorting by number of trips will result in the most visited Asian countries in the summer of each year (see Figure 11).

Some conclusions or conjectures could be drawn from these results:

- Tourist countries like Japan or Thailand disappeared from the Top-5 during 2020 and 2021 (Japan also in 2022), because of travel restrictions due to the pandemic.

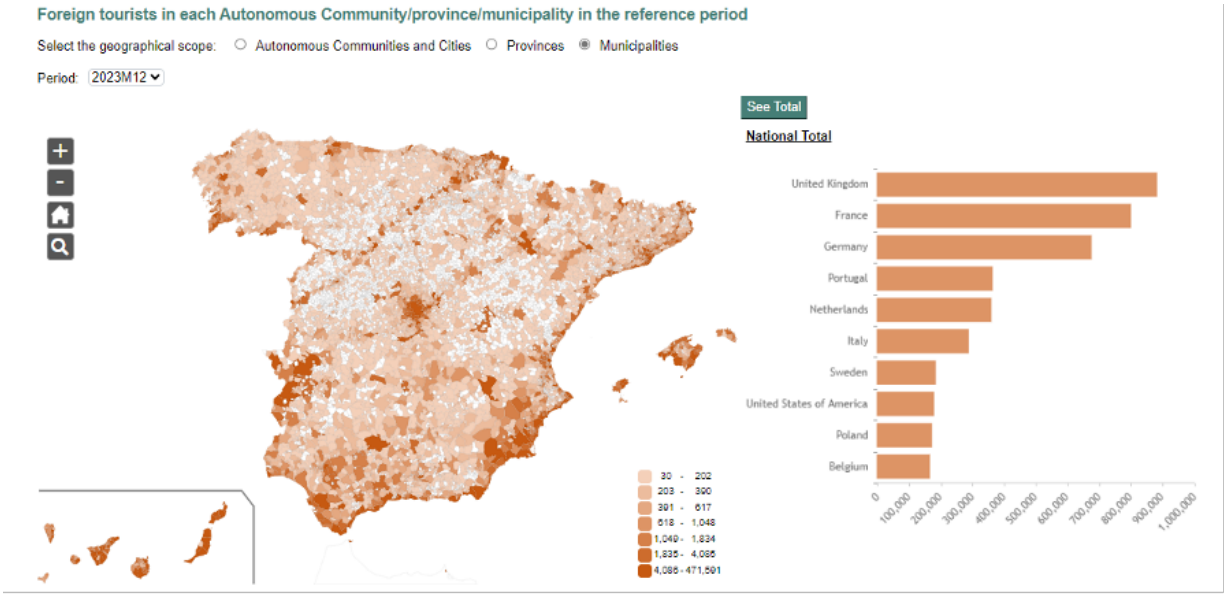


Foreign tourists in Spain from each country in the reference period

Select the geographical scope: ○ Europe ● America ○ Asia ○ Africa ○ Oceania

Period: 2023M12

Figure 6: Inbound tourism data by continents.



Foreign tourists in each Autonomous Community/province/municipality in the reference period

Select the geographical scope: ○ Autonomous Communities and Cities ○ Provinces ● Municipalities

Period: 2023M12

Figure 7: Inbound tourism data by destination.

- Maldives appears for the first time in the ranking in 2021. This may indicate that it has become fashionable among Spanish people or that perhaps they have put direct and cheaper flights to the country or had fewer travel restrictions.
- The year 2023 can be considered the year of normality in terms of international travel by Spaniards, as the ranking of the top 10 Asian countries is the same as in 2019 (except for changes in positions).

Objective: To identify favorite destination municipalities according to country of origin.

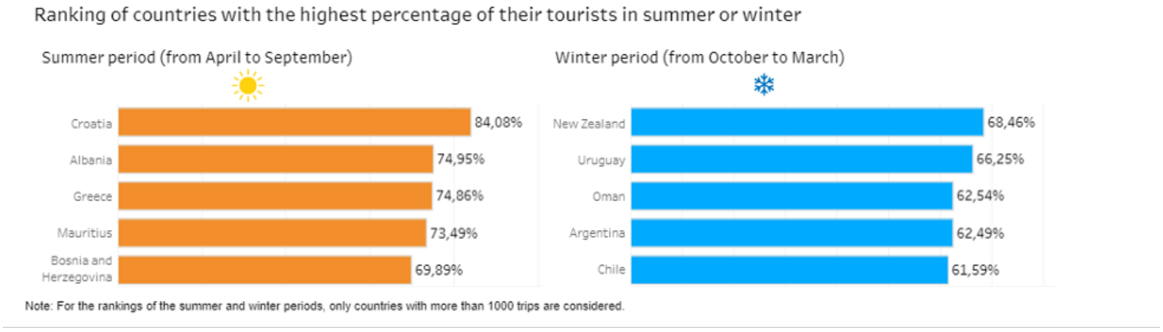


Figure 8: Ranking of countries with the highest percentage of tourist in summer or winter season.

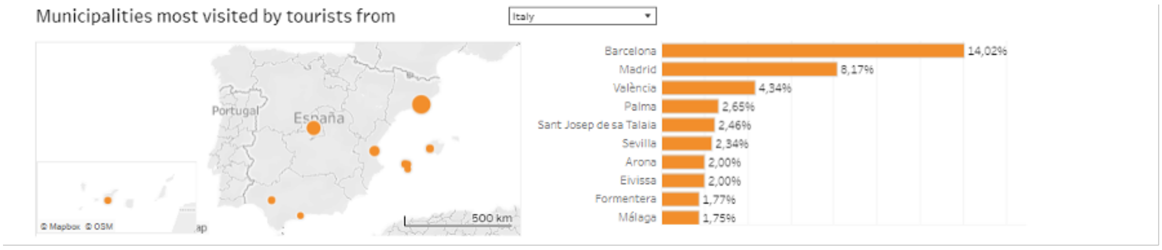


Figure 9: Municipalities most visited by country.

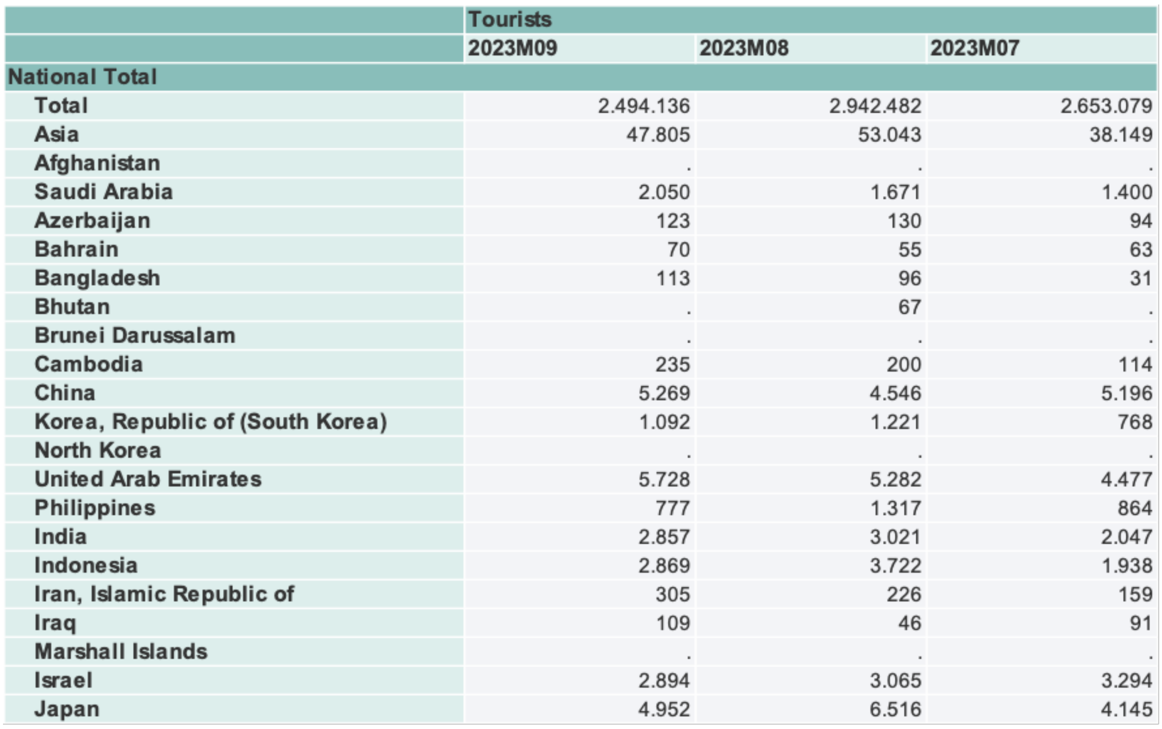
Using monthly tourist data per destination municipality, broken down by continent and country of residence (for this example, data from August 2023[2] was used), it is possible to reveal the different preferences according to nationality. Another possible exercise along these lines would be to look at the different preferences of tourists from the same country throughout the year (see Figure 12).

5 Conclusions

After the experience gained in these years of work in this experimental statistic that we have just described, using cell phone positioning data, and in general, with the use of new sources of information, we can conclude that:

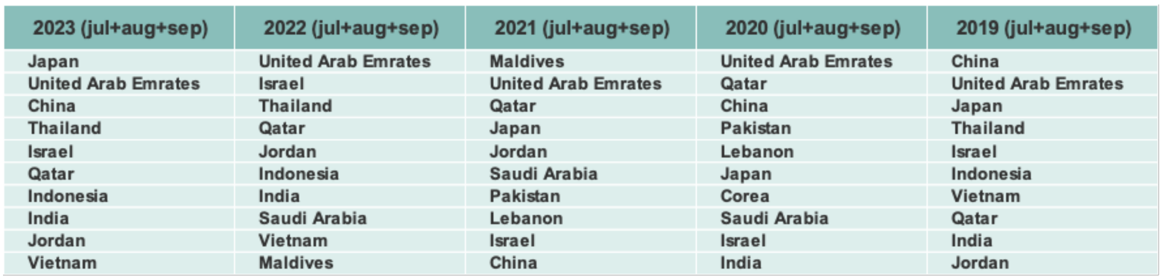
- It is important to use these sources of information, with rigor, within the framework of official statistics. The advantages they offer are many, among which stand out the more than significant increase in the granularity of the results obtained (difficult to obtain with traditional methods), the reduction of the burden on the informant or the improvement in the timeliness of the final results.
- In order to work accurately with these new sources of information it is very important to collaborate with the owners of these databases, especially if they have to make specific treatments to respond to statistical needs (adaptation of definitions, adjustment of fields of study, etc.). They must be aware of the relevance of the use of the information and there must be maximum collaboration between the statistical offices and the database owners.
- In general, the new data sources do not give a total answer or do not provide the complete information that in terms of variables, for example, of the traditional surveys, so the way to

[2] https://ine.es/experimental/turismo_moviles/exp_tmov_receptor_mun_2023.xlsx



	Tourists		
	2023M09	2023M08	2023M07
National Total			
Total	2.494.136	2.942.482	2.653.079
Asia	47.805	53.043	38.149
Afghanistan	.	.	.
Saudi Arabia	2.050	1.671	1.400
Azerbaijan	123	130	94
Bahrain	70	55	63
Bangladesh	113	96	31
Bhutan	.	67	.
Brunei Darussalam	.	.	.
Cambodia	235	200	114
China	5.269	4.546	5.196
Korea, Republic of (South Korea)	1.092	1.221	768
North Korea	.	.	.
United Arab Emirates	5.728	5.282	4.477
Philippines	777	1.317	864
India	2.857	3.021	2.047
Indonesia	2.869	3.722	1.938
Iran, Islamic Republic of	305	226	159
Iraq	109	46	91
Marshall Islands	.	.	.
Israel	2.894	3.065	3.294
Japan	4.952	6.516	4.145

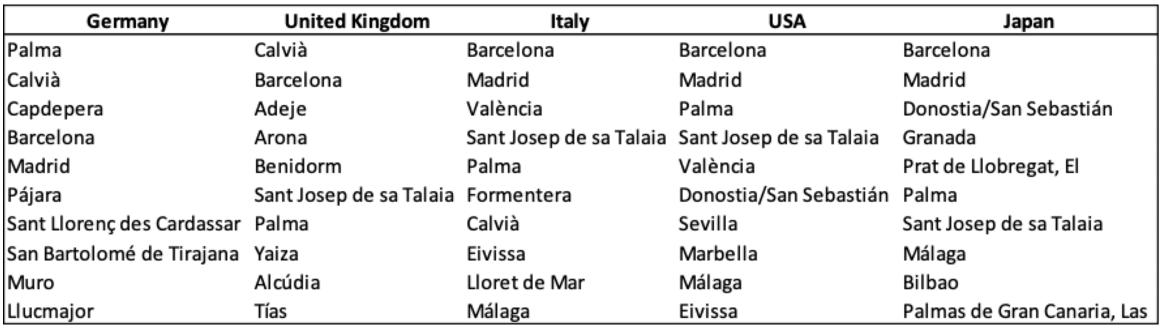
Figure 10: Trips to Asian countries



2023 (jul+aug+sep)	2022 (jul+aug+sep)	2021 (jul+aug+sep)	2020 (jul+aug+sep)	2019 (jul+aug+sep)
Japan	United Arab Emrates	Maldives	United Arab Emrates	China
United Arab Emrates	Israel	United Arab Emrates	Qatar	United Arab Emrates
China	Thailand	Qatar	China	Japan
Thailand	Qatar	Japan	Pakistan	Thailand
Israel	Jordan	Jordan	Lebanon	Israel
Qatar	Indonesia	Saudi Arabia	Japan	Indonesia
Indonesia	India	Pakistan	Corea	Vietnam
India	Saudi Arabia	Lebanon	Saudi Arabia	Qatar
Jordan	Vietnam	Israel	Israel	India
Vietnam	Maldives	China	India	Jordan

Figure 11: Ranking of most visited Asian countries by Spanish populations

follow integration of different sources. An example will be in the inbound tourism survey where the information provided by mobile telephony will serve to estimate the number of tourists visiting Spain, knowing the countries of origin and their places of destination, and with a survey it will be possible to know the characteristics of their trips such as the purpose of trips, type of accommodation or the tourism expenditure.



Germany	United Kingdom	Italy	USA	Japan
Palma	Calvià	Barcelona	Barcelona	Barcelona
Calvià	Barcelona	Madrid	Madrid	Madrid
Capdepera	Adeje	València	Palma	Donostia/San Sebastián
Barcelona	Arona	Sant Josep de sa Talaia	Sant Josep de sa Talaia	Granada
Madrid	Benidorm	Palma	València	Prat de Llobregat, El
Pájara	Sant Josep de sa Talaia	Formentera	Donostia/San Sebastián	Palma
Sant Llorenç des Cardassar	Palma	Calvià	Sevilla	Sant Josep de sa Talaia
San Bartolomé de Tirajana	Yaiza	Eivissa	Marbella	Málaga
Muro	Alcúdia	Lloret de Mar	Málaga	Bilbao
Llucmajor	Tías	Málaga	Eivissa	Palmas de Gran Canaria, Las

Figure 12: Ranking of most visited Spanish municipalities by tourist country of residence

References

Spain, NSI (2022a). Measurement of domestic tourism using mobile phone positioning. Technical report, Spain: National Statistics Institute.

Spain, NSI (2022b). Measurement of inbound tourism using mobile phone positioning. Technical report, Spain: National Statistics Institute.

Spain, NSI (2022c). Measurement of outbound tourism using mobile phone positioning. Technical report, Spain: National Statistics Institute.

UNWTO (2008). International recommendations for tourism statistics 2008. Technical report, Madrid, Spain: United Nations World Tourism Organization.

SPANISH JOURNAL OF STATISTICS
Vol. **6** No. 1 2024, Pages 21–38
doi:https://doi.org/10.37830/SJS.2024.1.03

REGULAR ARTICLE

Finding most nearly compatible conditionals under a finite discrete set-up: An overview and recent developments

Indranil Ghosh

Department of Mathematics and Statistics, University of North Carolina, ghoshi@uncw.edu

Received: February 26, 2024. Accepted: July 31, 2024.

Abstract:

In modeling complicated real-life scenarios, one objective is to capture the dependence being observed. Consequently, conditional specification is a worthy alternative to the joint-distribution models. Since its' inception, the use of divergence measures have been instrumental in determining the closeness between two probability distributions, especially when joint distributions are specified by the corresponding conditional distributions. Conditional specification of distributions is a developing area with several applications. This work gives an overview of a variety of divergence measures including, but not limited to, Kullback-Leibler divergence measure, Power-divergence statistic, Hellinger distance along with some newly developed divergence measures and its role in addressing various compatible conditions in search for a most-nearly compatible for a finite discrete case, and also identifying compatibility under conditional and marginal information under some additional information in the form of marginal and/or conditional summary. Finally, we provide some numerical examples to illustrate each of the scenarios.

Keywords: ncompatible conditionals, divergence measures, iterative algorithm, conditional specification, near compatibility

MSC: 62H05, 62E17

1 Introduction

The problem of determining whether two families of conditional distributions are compatible or minimally incompatible has been considered by several authors and the problem is well established in the literature. For an excellent survey on this topic, an interested reader is referred to the scholarly works by Arnold and Press (1989) and Arnold et al. (1999) and the references cited therein. A non-exhaustive list of pertinent references can be cited as follows, for example, in the works by Gelman and Speed (1993), and Arnold and Gokhale (1994, 1998). Arnold et al. (1992) provided a useful survey of distributions being obtained in such a fashion. Several alternative approaches exist in the literature with regard to the problem of determining the possible compatibility of two families of conditional distributions, for example in the works of Arnold and Press (1989); Arnold and Gokhale (1994); Cacoullos and Papageorgiou (1983); Wesolowski (1996). In addition, the problem of determining most nearly compatible distributions, in the absence of compatibility, has been addressed (Arnold and Gokhale, 1998; Arnold et al., 1999, 2001). In this paper, our our main objective is concentrated on cases in which the conditional specifications are incompatible. In addition, we envision a scenario in which case, from our informed expert and/or practitioner who is working in this field has provided a set of additional information in the form of conditional moments/percentiles; marginal moments etc. We want to examine to what extent such amount of additional information is compatible with the given two conditional probability matrices in search for a most nearly compatible (equivalently minimally incompatible) probability distribution. It is safe to say that the problem has been explored by Arnold et al. (2001) in which the authors derived this problem as a set of non-linear equations involving some constraints.

Our search for a compatible P in terms of equations subject to inequality constraints is based on the fact that we really need to find one compatible marginal, say that corresponding to the random variable X, and we consider the fact that when this is combined with B will give us P. However, in this paper, we look at a different objective which is not discussed in Arnold et al. (2001). Here, we explore the applicability of several measures of divergence (alias pseudo-distance measures) in finding a most nearly compatible distributions by incorporating the additional sets of information along with the complete specification of two conditionals. For an excellent survey on the use of divergence measures in various aspects of distribution theory and associated statistical inference, one is suggested to take a look at the book by Pardo (2006).

In particular, we examine the relative performance of these measures of divergence based on at what stage of iterative algorithm in search for a most nearly compatible P, the adopted procedure converges based on a user defined level of precision which is described later. Needless to say, compatible conditional and marginal specifications of distributions are of fundamental importance in modeling scenarios. Moreover in Bayesian prior elicitation contexts, inconsistent conditional specifications are to be expected. In such situations interest will center on most nearly compatible distributions.

The remainder of the paper is organized as follows. In Section 2, we provide some basic preliminaries regarding compatibility of two discrete conditionals. Section 3 deals with various necessary conditions for compatibility. In Section 4, we discuss the role of pseudo-distance measures in identifying a most nearly compatible probability distribution starting from two given conditional probability matrices under a finite discrete set-up. In Section 5, various methods of finding most nearly compatible distributions are discussed. Section 6 provides an overview on the topic of using pseudo-divergence measures in the presence of additional marginal and/or conditional information. Several illustrative examples are provided in Section 7. Finally, some concluding remarks are presented in Section 8.

2 Basic preliminaries

Let A and B be two $(I \times J)$ matrices with non-negative elements such that $\sum_{i=1}^{I} a_{ij} = 1, \forall j = 1, \ldots, J$ and $\sum_{j=1}^{J} b_{ij} = 1, \forall i = 1, 2, \ldots, I$. Without loss of generality, it can be assumed that $I \leq J$. Matrices A and B are said to form a compatible conditional specification for the distribution of (X, Y) if there exists some $(I \times J)$ matrix P with non-negative entries p_{ij} and with $\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = 1$ such that, for every (i,j), $a_{ij} = \frac{p_{ij}}{p_{.j}}$ and $b_{ij} = \frac{p_{ij}}{p_{i.}}$, where $p_{i.} = \sum_{j=1}^{J} p_{ij}$ and $p_{i.} = \sum_{i=1}^{I} p_{ij}$. If such a matrix P exists, then, if we assume that $p_{ij} = P(X = x_i, Y = y_j)$, $i = 1, 2, \cdots, I$, $j = 1, 2, \cdots, J$, we will have $a_{ij} = P(X = x_i | Y = y_j)$, $i = 1, 2, \cdots, I$, $j = 1, 2, \cdots, J$, and $b_{ij} = P(Y = y_j | X = x_i)$, $i = 1, 2, \cdots, I$, $j = 1, 2, \cdots, J$. Equivalently, A and B are compatible if there exist stochastic vectors $\underline{\tau} = (\tau_1, \tau_2, \cdots, \tau_J)$ and $\underline{\eta} = (\eta_1, \eta_2, \cdots, \eta_I)$ such that

$$a_{ij}\tau_j = b_{ij}\eta_i,$$

for every (i, j). In the case of compatibility, $\underline{\eta}$ and $\underline{\tau}$ can be readily interpreted as the resulting marginal distributions of X and Y, respectively. For any probability vector $\eta = (\eta_1, \eta_2, \ldots, \eta_I)$, $p_{ij} = b_{ij}\eta_i$ is a probability distribution on the IJ cells. So, the conditional probability matrix, denoted by A, and its elements (a_{ij}) will be given by

$$a_{ij} = \frac{p_{ij}}{\sum_{s=1}^{I} p_{sj}} = \frac{b_{ij}\eta_i}{\sum_{s=1}^{I} b_{sj}\eta_s}, \tag{1}$$

for every i, j. If A and B are compatible, then

$$a_{ij}\sum_{s=1}^{I} b_{sj}\eta_s = b_{ij}\eta_i.$$

We then have

$$\tau_j = \sum_{s=1}^{I} b_{ij}\eta_s, \forall j = 1, \ldots, J.$$

In this case, the expressions given in (1) can be rewritten as

$$a_{ij}\sum_{s=1}^{I} b_{sj}\eta_s - b_{ij}\eta_i = 0.$$

3 Compatibility conditions

Conditions for compatibility are listed in the following theorems which are due to Arnold and his co-authors.

Suppose that A and B have identical incidence sets then they are compatible if and only if either of the following two conditions hold.

(a) There exist stochastic vectors $\vec{\tau} = (\tau_1, \tau_2, ..., \tau_I)$ and $\vec{\eta} = (\eta_1, \eta_2, ..., \eta_J)$ such that $\eta_j a_{ij} = (\tau_i b_{ij})$,$\forall i, j$. In the case of compatibility, the vectors $\vec{\tau}$ and $\vec{\eta}$ can readily interpreted being proportional to the marginal distributions of X and Y respectively.

(b) There exists vectors $\vec{u}$ and $\vec{v}$ for which $d_{ij} = \frac{a_{ij}}{b_{ij}} = u_i v_j$,$\forall i, j \in N$.

This suggests the use of log-linear models to fit the matrix D. Indeed, if the log-linear model has all interactions equal to zero, then we have compatibility. Otherwise, A and B are incompatible.
If $N = \{1, 2, ...I\} \times \{1, 2, ...J\}$, i.e; if all the entries in A and B are positive, then we have the following theorem given by due to Arnold and Gokhale (1994).

1. A and B are compatible iff they have identical uniform marginal representations(UMRs) (Mosteller, 1968).
2. A and B are compatible iff all cross product ratios of A are identical to those of B.

Note: Some restrictions on the common incidence set of A and B is necessary for the above theorem. For example if we consider

$$A = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{pmatrix}$$

and $$B = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 0 & 1/3 & 2/3 \\ 2/3 & 0 & 1/3 \end{pmatrix}$$

It may be verified here here that A and B have equal cross product ratios(there are no positive 2×2 submatrices)and have identical uniform marginal representations but A and B are not compatible. Compatibility of A and B of course does not confirm a unique compatible matrix P. The simplest sufficient condition is positivity, i.e; $(a_{ij}b_{ij}) \geq 0$ and $\forall i, j$.

4 Measures of divergence

In this section, we list several useful divergence measures which will be utilized in this paper for finding the ϵ-compatible distributions under the finite discrete set-up. In addition, we provide some useful relationships among these divergence measures. Some of these results have been independently derived and discussed in Ghosh and Sunoj (2024) and Borzadaran and Amini (2010) in the context of copula-based divergence measures. We begin our discussion with the power divergence statistics as a measure of divergence, for pertinent details, see Cressie and Read (1984). A divergence measure between two probability distributions $\underline{p}$ and $\underline{q}$ (which are of the same dimension) returns a measure of similarity or distance between them. It is non-negative. It measures the divergence between the population distribution $\underline{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)$ and the uniform distribution $\left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$, where a value closer to zero represents a wider divergence from the uniform distribution. A natural generalization, when considered in this way, is to define a measure of divergence between two general distributions. This concept was first considered by Kullback (1959) in his directed divergence measure. It was followed up by Arnold and Gokhale (1994, 1998) while considering minimum incompatibility via the K-L criterion. It is of the form

$$K\left(\underline{p} : \underline{q}\right) = \sum_{i=1}^{k} p_i \log_2 \left(\frac{p_i}{q_i}\right), \tag{2}$$

where $\underline{p}$ and $\underline{q}$ are two discrete probability distributions defined on the $(k-1)$ dimensional simplex

$$\Delta_k = \left\{ \underline{\pi} : \pi_i \geq 0; i = 1, \ldots, k; \sum_{i=1}^{k} \pi_i = 1 \right\}.$$

Here, we adopt the convention that $p_i \log_2 \left(\frac{p_i}{q_i}\right) = 0$ when $p_i = 0$ and for any $0 \leq q_i \leq 1$. A family of power divergence statistics indexed by $\lambda \in \mathbb{R}$ for $\underline{p} = (p_1, p_2, \ldots, p_k)$, $\underline{q} = (q_1, q_2, \ldots, q_k)$ can be defined as

$$I^{\lambda}\left(\underline{p}:\underline{q}\right)=\frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{k}p_i\left[\left(\frac{p_i}{q_i}\right)^{\lambda}-1\right] \tag{3}$$

with the convention $p_i=0$ whenever $q_i=0$. Note that (3) generalizes (2) in the same way the Rényi entropy (Rényi, 1961) generalizes the Shannon entropy (Shannon, 1951).

1. Considering the fact that a matrix can be written as an array of column vectors, we define the power divergence statistic for matrices A and B as:

$$\begin{aligned} D_1 &= I^{\lambda}\left(p_{ij}:a_{ij}p_{\cdot j}\right)+I^{\lambda}\left(p_{ij}:b_{ij}p_{i\cdot}\right)\\ &= \frac{1}{\lambda(\lambda+1)}\left[\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\left(\left(\frac{p_{ij}}{a_{ij}p_{\cdot j}}\right)^{\lambda}-1\right)+\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\left(\left(\frac{p_{ij}}{b_{ij}p_{i\cdot}}\right)^{\lambda}-1\right)\right], \end{aligned}$$

where $\lambda\in\mathbb{R}$ is a parameter. The power divergence statistic is undefined for $\lambda=-1$ or $\lambda=0$. However, if we define these two cases as continuous limits of D_1 for $\lambda\to-1$ and $\lambda\to0$, then D_1 is continuous in λ.
The name power divergence derives from the fact that the statistic D_1 measures the divergence of p_{ij} from $(a_{ij}p_{\cdot j})$ and $(b_{ij}p_{i\cdot})$ through a weighted sum of powers of the terms $\left(\frac{p_{ij}}{a_{ij}p_{\cdot j}}\right)$ and $\left(\frac{p_{ij}}{b_{ij}p_{i\cdot}}\right)$ for all $(i,j)\in N$. We want to minimize D_1 with respect to $\sum\sum_{(i,j)\in N}p_{ij}=1$.

Note: On the choice of λ

In the power divergence statistic, λ is a parameter that can take any real value. A natural question that arises here is: what should be the optimum choice of λ? There are some conflicting recommendations regarding which value of λ results in the optimal test statistic. In all our examples of iterative study discussed in Section 4 later, we find that the rate of convergence is very slow for most values of λ. For example, for $\lambda=0.2,0.3$ and 0.5, the iterative procedure for the divergence measure D_{λ} converges at $n=20,27$ and 34, respectively. For negative choices of λ, D_1 is quite big, and moreover the resulting matrix is not a probability matrix. A future work will focus on providing practical guidelines about how to choose λ and also to investigate the sensitivity of solutions in addition to the rate of convergence) when different values of λ's are used in its' permissible range. In the next, we provide a collection of divergence measures which has been utilized to obtain the ϵ-compatible distribution(s) under the finite discrete set-up. For pertinent details, see Ghosh (2011), Ghosh and Balakrishnan (2015), Ghosh and Nadarajah (2017) and the references cited therein.

2. Modified Renyi's divergence measure, see Ghosh (2011)

$$D_2=\frac{1}{(\alpha-1)}\left[\sum_{i=1}^{I}\sum_{j=1}^{J}(a_{ij}p_{.j})^{-1}\log\left(\frac{p_{ij}}{a_{ij}p_{.j}}\right)^{\alpha}+\sum_{i=1}^{I}\sum_{j=1}^{J}(b_{ij}p_{i.})^{-1}\log\left(\frac{p_{ij}}{b_{ij}p_{i.}}\right)^{\alpha}\right] \tag{4}$$

Note: Nadarajah and Zografos (2003); Zografos and Nadarajah (2005) provided a useful review of Renyi's entropy for different univariate and k-variate random variables.

3. χ^2 measure of divergence
   It is defined as

$$D_3 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\left(\frac{p_{ij}}{a_{ij}p_{.j}}\right)^2\right]a_{ij}p_{.j} + \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\left(\frac{p_{ij}}{b_{ij}p_{i.}}\right)^2\right]b_{ij}p_{i.} \tag{5}$$

4. First new measure of divergence (see, Ghosh (2011))

$$D_4 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\left(\frac{p_{ij}}{a_{ij}p_{.j} + b_{ij}p_{i.}} - 1\right)^2\right]^{\lambda}, \tag{6}$$

   where $\lambda > 0$ is a constant.

5. Second new measure of divergence (see, Ghosh (2011))

$$D_5 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sqrt{p_{ij}} - \sqrt{a_{ij}p_{.j}}\right)^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sqrt{p_{ij}} - \sqrt{b_{ij}p_{.j}}\right)^2. \tag{7}$$

Note: It is to be noted that if the two conditional matrices A and B are compatible then each of these measures will be equal to zero.

5 Available methods of obtaining minimally incompatible distributions

In this section, we describe the idea of minimal incompatibility of two given conditional distributions, and then explain some methods of finding minimally incompatible distributions. For pertinent details, see Arnold et al. (1999).

5.1 ϵ-Compatibility

Suppose, we do not insist on precise compatibility, and instead wish to have p_{ij} to be approximately consistent with two given conditional probability matrices A and B. Let W be a weight matrix that represents the relative importance of accuracy in determining the probabilities p_{ij} for each (i,j). For a given weight matrix W which might be uniform, i.e., $w_{ij} = 1, \forall\,(i,j)$ if all pairs (i,j) were equally important, we may consider the following strategies expressed as non-linear and linear programming problems.

(i) **First method:** Find a matrix P, with $p_{ij} \geq 0 \quad \forall\,(i,j)$, such that

$$\left|p_{ij} - a_{ij}\sum_{i=1}^{I}p_{ij}\right| \leq \epsilon w_{ij} \quad \forall(i,j) \in N,$$

$$\left|p_{ij} - b_{ij}\sum_{j=1}^{J}p_{ij}\right| \leq \epsilon w_{ij} \quad \forall(i,j) \in N,$$

with the linear constraint $\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij} = 1$.

(ii) **Second method:** Second method: Seek two probability vectors η and τ such that $|a_{ij}\eta_j - b_{ij}\tau_i| \le \epsilon w_{ij} \quad \forall\,(i,j))\,,\quad \sum_j \eta_j = 1,\quad \sum_i \tau_i = 1,\quad$ and $\tau_i \ge 0,\quad \eta_j \ge 0, \forall\,(i,j) \in N.$

(iii) **Third method:** Find a (marginal) probability vector $\tau \ge 0$, such that and $\tau_i \ge 0, \forall i$.

Clearly, the above methods introduce three different concepts of ϵ-compatibility. If we use Method 1, and if A and B are ϵ-compatible, then the matrix P^* which satisfies Eq. (1) will be said to be most nearly compatible. If we use Method 2 and if A and B are ϵ-compatible, then a reasonable choice for a most nearly compatible matrix P^* will be

$$P^* = \frac{a_{ij}\eta_j^* + b_{ij}\tau_j^*}{2},$$

where η_j^* and τ_j^* satisfy Eq.(2). Finally, if we use Method 3 and if A and B are ϵ-compatible, then a plausible choice for a most nearly compatible P^* will be $P^* = (b_{ij}\tau_i^*)$, where τ_i^* satisfies Eq.(3).

6 Pseudo-divergence measures under additional information

Until now, we have discussed the power divergence statistic as a measure of divergence to obtain minimally incompatible (or equivalently ε-compatible) joint probability distributions from the set of two conditionals. Here, we want to find a procedure from which we would like to get the joint probability distribution from the two conditionals but with some additional information provided on the marginal and conditional probabilities and expectations, i.e., we want to see whether a given set of constraints involving marginal and conditional probabilities and expectations of functions are compatible or minimally incompatible. The finite discrete case (the main focus of the paper) may be viewed as one involving solutions of linear equations in restricted domains. We will consider cases where the given conditional probabilities and expectations are specified. Cases of imprecise specification will be considered later on. So far, in all divergence criteria we minimized the given function based on only one linear constraint: $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$. Instead, Suppose we are given (by our well-informed expert engaged in this study) the following set of marginal and conditional information (one may call this a set of precise information):

1. $P\left(\underline{X} \in A_i\right) = \delta_i$ for specified sets $A_1, A_2, \ldots, A_{n_1}$,
2. $P\left(\underline{X} \in B_i \middle| \underline{X} \in C_i\right) = \eta_i$, $i = 1, 2, \ldots, n_2$ for specified sets of $B_1, B_2, \ldots, B_{n_2}$, and $C_1, C_2, \ldots, C_{n_2}$,
3. $E\left(\epsilon_j\left(\underline{X}\right)\right) = \xi_j$, $j = 1, 2, \ldots, n_3$ for specified functions $\epsilon_1, \epsilon_2, \ldots, \epsilon_{n_3}$,
4. $E\left(\varphi_i\left(\underline{X}\right) \middle| \phi_i\left(\underline{X}\right) = \lambda_i\right) = \omega_i$, $i = 1, 2, \ldots, n_4$ for specified functions $\varphi_1, \varphi_2, \ldots, \varphi_{n_4}$ and specified constants $\lambda_1, \lambda_2, \ldots, \lambda_{n_4}$,
5. $P\left(\nu_i\left(\underline{X}\right) \in E_i \middle| \gamma_i\left(\underline{X}\right) \in F_i\right) = \beta_i$, $i = 1, 2, \ldots, n_5$ for specified functions $\nu_1, \nu_2, \ldots, \nu_{n_5}$ and specified sets $E_1, E_2, \ldots, E_{n_5}$ and $F_1, F_2, \ldots, F_{n_5}$.

Note that the above sets of information can be rewritten as follows.

- $P\left(\underline{X} \in A_i\right) = \sum_{\underline{X} \in A_i} p\left(\underline{X}\right) = \delta_i$,
- $P\left(\underline{X} \in B_i \middle| \underline{X} \in C_i\right) = \eta_i$ if and only if $\sum_{\underline{X} \in B_i \cap C_i} p\left(\underline{X}\right) - \eta_i \sum_{\underline{X} \in C_i} p\left(\underline{X}\right) = 0$,

- $E\left(\epsilon_j\left(\underline{X}\right)\right) = \sum_{\underline{X}} \epsilon_j\left(\underline{X}\right) p\left(\underline{X}\right) = \xi_j$,
- $E\left(\varphi_i\left(\underline{X}\right) \middle| \phi_i\left(\underline{X}\right) = \lambda_i\right) = \omega_i$ if and only if $\sum_{\phi_i(\underline{X})=\lambda_i} \varphi_i\left(\underline{X}\right) p\left(\underline{X}\right) - \omega_i \sum_{\phi_i(\underline{X})=\lambda_i} p\left(\underline{X}\right) = 0$,
- $P\left(\nu_i\left(\underline{X}\right) \in E_i \middle| \gamma_i\left(\underline{X}\right) \in F_i\right) = \beta_i$ if and only if $\sum_{\nu_i(\underline{X})\in E_i \cap \gamma_i(\underline{X})\in F_i} p\left(\underline{X}\right) - \beta_i \left(\sum_{\nu_i(\underline{X})\in F_i} p\left(\underline{X}\right) \right) = 0.$

Thus, if we arrange the values of the joint density $p\left(\underline{X}\right)$ of $\underline{X}$ as a vector of dimension $\Omega = \text{card}(X_1) \times \text{card}(X_2) \times \cdots \times \text{card}(X_k)$, where $\underline{X} = (X_1, X_2, \ldots, X_k)$, then we can write every piece of information given above in the form:

$$M\underline{p} = \underline{\theta}, \tag{8}$$

where the matrix M in Eq.(8) is of order $(r+1) \times \Omega$, assuming r pieces of information are given and rank $r + 1 \leq \Omega$. The $\vec{\theta}$ is of order $(r+1) \times 1$. Both M and θ are assumed to be known. The "natural" constraint $\underline{p} \cdot \vec{1} = 1$ is incorporated in Eq.(8) by letting the first row of M consist of all unit elements and the first element of θ equal to unity. The system in Eq.(8) is assumed to be consistent in the sense that there exists a positive probability vector satisfying (1). If $r + 1$ is large, it is highly unlikely that $r+1$ pieces of information will be compatible with the given information, in the sense that Eq.(8) has a solution $\underline{p}^*$ with non-negative coordinates adding up to one. In general, it would be more rational to seek approximate equality in Eq.(8) subject to $\underline{p} \geq 0$ and $M\underline{p} = \underline{\theta}$. In other words, we are seeking an almost compatible distribution.

6.1 Power divergence statistic under conditional and marginal information

Our search for a most nearly compatible distribution (equivalently ε compatible) $\underline{p}$ can be viewed as a problem of minimizing $D\left(M\underline{p}, \underline{\theta}\right)$ for a suitable distance measure D subject to the restriction that $\underline{p} \geq 0$ and $M\underline{p} = \underline{\theta}$. One such reasonable distance measure is the power divergence statistic. The determined minimum value of the objective function, in each of the examples, described later, provides a measure of incompatibility of the given information.

In this case, we have $P^{I \times J} = \left(\underline{p}_1, \underline{p}_2, \ldots, \underline{p}_I\right)^{1\times I}$, where $\underline{p}_1 = (p_{11}, p_{12}, \ldots, p_{1J})^{1\times J}$, $\underline{p}_2 = (p_{21}, p_{22}, \ldots, p_{2J})^{1\times J}$, and so on up to $\underline{p}_I = (p_{I1}, p_{I2}, \ldots, p_{IJ})^{1\times J}$, and we have the linear restriction of the form

$$\sum_{u=1}^{I} M_{tu}\underline{p}_u = \theta_t,$$

for $t = 1, 2, \ldots, (r+1)$. The power divergence statistic (PDS) in this case reduces to

$$D_1\left(\underline{p}\right) = \frac{1}{\lambda(\lambda+1)} \sum_{u=1}^{I} \left[\underline{p}_u \left(\left(\frac{\underline{p}_u}{\underline{a}_u p_{\cdot j}} \right)^{\lambda} - 1 \right) + \underline{p}_u \left(\left(\frac{\underline{p}_u}{\underline{b}_u p_{i \cdot}} \right)^{\lambda} - 1 \right) \right].$$

Now we consider the following Lagrangian function

$$F = D_1\left(\underline{p}\right) + \sum_{t=1}^{r+1} \tau_t \left(\sum_{u=1}^{I} M_{tu} p_u - \theta_t \right),$$

where τ_t, $t = 1, 2, \ldots, (r+1)$ are $(r+1)$ Lagrangian multipliers. To minimize F, we consider simultaneous solution of

$$\frac{\partial F}{\partial \underline{p}_u} = 0. \tag{9}$$

Consequently, the optimal value of $\underline{p}_u$ is

$$\underline{p}_u^* = \frac{\left(\left(\frac{1}{(\underline{a}_u p_{\cdot j})^{\lambda}} + \frac{1}{(\underline{b}_u p_{i\cdot})^{\lambda}} \right)^{\frac{1}{\lambda}} \right)^{-1}}{\left(\sum_{u \in N} \left\{ \left(\frac{1}{(\underline{a}_u p_{\cdot j})^{\lambda}} + \frac{1}{(\underline{b}_u p_{i\cdot})^{\lambda}} \right)^{\frac{1}{\lambda}} \right\}^{-1} \right)^{-1}}.$$

For an iterative study, we consider the following

$$\underline{p}_u^{n+1} = \frac{\left(\frac{1}{\left(\underline{a}_u p_{\cdot j}^n\right)^{\lambda}} + \frac{1}{\left(\underline{b}_u p_{i\cdot}^n\right)^{\lambda}} \right)^{\frac{1}{\lambda}}}{\sum_{u \in N} \left(\frac{1}{\left(\underline{a}_u p_{\cdot j}^n\right)^{\lambda}} + \frac{1}{\left(\underline{b}_u p_{i\cdot}^n\right)^{\lambda}} \right)^{\frac{1}{\lambda}}},$$

for $n = 0, 1, \ldots$ with the initial choice of $p_{ij}^{(0)} = \frac{1}{IJ}$ for all $(i, j) \in N$. We may use the stopping rule for this iterative algorithm as $\left| \frac{D_1^{(n+1)}}{D_1^{(n)}} - 1 \right| \leq 10^{-6}$. In all the examples we considered, our process was found to converge for a wide range of λ.

6.2 Kullback-Leibler divergence criterion under conditional and marginal information

In this case, the K-L divergence statistic is

$$D_2(\underline{p}) = \sum_{u=1}^{I} \left[\underline{a}_u \log \left(\frac{\underline{a}_u p_{\cdot j}}{\underline{p}_u} \right) + \underline{b}_u \log \left(\frac{\underline{b}_u p_{i\cdot}}{\underline{p}_u} \right) \right].$$

Again, we consider the following Lagrangian function

$$F_2 = D_2\left(\underline{p}\right) + \sum_{t=1}^{r+1} \tau_t \left(\sum_{u=1}^{I} M_{tu} p_u - \theta_t \right),$$

where τ_t, $t = 1, 2, \ldots, (r+1)$ are $(r+1)$ Lagrangian multipliers. To minimize F_2, we consider simultaneous solution of

$$\frac{\partial F_2}{\partial \underline{p}_u} = 0,$$

same as in (9). So, the optimal value of $\underline{p}_u$ is

$$\underline{p}_u^* = \frac{\left(\frac{\underline{a}_u + \underline{b}_u}{\frac{1}{p_{i\cdot}} + \frac{1}{p_{\cdot j}}} \right)}{\left(\sum_{u \in N} \frac{\underline{a}_u + \underline{b}_u}{\frac{1}{p_{i\cdot}} + \frac{1}{p_{\cdot j}}} \right)}.$$

For an iterative study, we consider the following

$$\underline{p}_u^{(n+1)} = \frac{\left(\frac{\underline{a}_u + \underline{b}_u}{\frac{1}{p^n_{i.}} + \frac{1}{p^n_{.j}}}\right)}{\left(\sum_{u \in N} \frac{\underline{a}_u + \underline{b}_u}{\frac{1}{p^n_{i.}} + \frac{1}{p^n_{.j}}}\right)}$$

for $n = 0, 1, \ldots$ with the initial choice of $p_{ij}^{(0)} = \frac{1}{IJ}$ for all $(i, j) \in N$. We use the following stopping rule $\left|\frac{D_2^{(n+1)}}{D_2^{(n)}} - 1\right| \leq 10^{-6}$. Here also our iterative algorithm is convergent.

6.3 Modified Renyi's measure of divergence under the marginal and conditional information

Proceeding as before, in this case, the statistic will be

$$D_3 = \frac{1}{(\alpha - 1)}\left[\sum_{i=1}^{I}\sum_{j=1}^{J} (\underline{a}_u p_{.j})^{-1} \log\left(\frac{p_{ij}}{\underline{a}_u p_{.j}}\right)^{\alpha} + \sum_{i=1}^{I}\sum_{j=1}^{J} (\underline{b}_u p_{i.})^{-1} \log\left(\frac{p_{ij}}{\underline{b}_u p_{i.}}\right)^{\alpha}\right]. \tag{10}$$

Next, we consider the following Lagrangian function

$$F_3 = D_3\left(\underline{p}\right) + \sum_{t=1}^{r+1} \tau_t \left(\sum_{u=1}^{I} M_{tu} p_u - \theta_t\right),$$

where τ_t, $t = 1, 2, \ldots, (r + 1)$ are $(r + 1)$ Lagrangian multipliers. Now, to minimize F_3, we consider simultaneous solution of

$$\frac{\partial F_3}{\partial \underline{p}_u} = 0,$$

same as in (9). Consequently, the optimal value of $\underline{p}_u$ is

$$\underline{p}_u^* = \frac{\frac{1}{\underline{a}_u p_{.j}} + \frac{1}{\underline{b}_u p_{i.}}}{\sum\sum_{(i,j)\in N}\left(\frac{1}{\underline{a}_u p_{.j}} + \frac{1}{\underline{b}_u p_{i.}}\right)}.$$

Subsequently, for an iterative study, we consider the following iterative algorithm

$$\underline{p}_u^{(n+1)} = \frac{\frac{1}{\underline{a}_u p_{.j}^{(n)}} + \frac{1}{\underline{b}_u p_{i.}^{(n)}}}{\sum\sum_{(i,j)\in N}\left(\frac{1}{\underline{a}_u p_{.j}^{(n)}} + \frac{1}{\underline{b}_u p_{i.}^{(n)}}\right)}.$$

for $n = 0, 1, \ldots$ with the initial choice of $p_{ij}^{(0)} = \frac{1}{IJ}$ for all $(i, j) \in N$. We use the following stopping rule $\left|\frac{D_3^{(n+1)}}{D_3^{(n)}} - 1\right| \leq 10^{-6}$. Here also our iterative algorithm is convergent based on all the empirical studies that we have made in this regard. A formal mathematical proof is still remains an open problem.

6.4 χ^2 divergence criterion under conditional and marginal information

In this case, our test statistic reduces to

$$D_4 = \sum \sum_{(i,j)\in N} \left[\left(\frac{p_{ij}}{\underline{a}_u p_{.j}} \right)^2 \right] \underline{a}_u p_{.j} + \sum \sum_{(i,j)\in N} \left[\left(\frac{p_{ij}}{\underline{b}_u p_{i.}} \right)^2 \right] \underline{b}_u p_{i.} \tag{11}$$

Next, we consider the following Lagrangian function

$$F_4 = D_4\left(\underline{p}\right) + \sum_{t=1}^{r+1} \tau_t \left(\sum_{u=1}^{I} M_{tu} p_u - \theta_t \right),$$

where τ_t, $t = 1, 2, \dots, (r+1)$ are $(r+1)$ Lagrangian multipliers. Now, to minimize F_4, we consider simultaneous solution of

$$\frac{\partial F_4}{\partial \underline{p}_u} = 0,$$

same as in (9). Consequently, the optimal value of $\underline{p}_u$ will be

$$\underline{p}_u^* = \left(\frac{1}{\underline{a}_u p_{.j}} + \frac{1}{\underline{b}_u p_{i.}} \right)^{-1} \left[\sum \sum_{(i,j)\in N} \frac{1}{\underline{a}_u p_{.j}} + \frac{1}{\underline{b}_u p_{i.}} \right]^{-1}$$

Consequently, an iterative algorithm for finding minimally compatible (alias ϵ-compatible) P would be to have

$$\underline{p}_u^{(n+1)} = \left(\frac{1}{\underline{a}_u p_{.j}^{(n)}} + \frac{1}{\underline{b}_u p_{i.}^{(n)}} \right)^{-1} \left[\sum \sum_{(i,j)\in N} \frac{1}{\underline{a}_u p_{.j}^{(n)}} + \frac{1}{\underline{b}_u p_{i.}^{(n)}} \right]^{-1},$$

for $n = 0, 1, \dots$ with the initial choice of $p_{ij}^{(0)} = \frac{1}{IJ}$ for all $(i,j) \in N$. We use the following stopping rule $\left| \frac{D_4^{(n+1)}}{D_4^{(n)}} - 1 \right| \leq 10^{-6}$. Here also our iterative algorithm is convergent based on all the empirical studies that we have made in this regard. A formal mathematical proof is still remains an open problem.

6.5 Divergence measure D_5 under conditional and marginal information

Here, our test statistic reduces to

$$D_5 = \sum \sum_{(i,j)\in N} \left[\left(\frac{p_{ij}}{\underline{a}_u p_{.j} + \underline{b}_u p_{i.}} - 1 \right)^2 \right]^{\lambda},$$

Next, we consider the following Lagrangian function

$$F_5 = D_5\left(\underline{p}\right) + \sum_{t=1}^{r+1} \tau_t \left(\sum_{u=1}^{I} M_{tu} p_u - \theta_t \right),$$

where τ_t, $t = 1, 2, \ldots, (r+1)$ are $(r+1)$ Lagrangian multipliers. Now, to minimize F_5, we consider simultaneous solution of

$$\frac{\partial F_5}{\partial \underline{p}_u} = 0,$$

same as in (9). Consequently, the optimal value of $\underline{p}_u$ will be

$$\underline{p}_u^* = \frac{(\underline{a}_u p_{.j} + \underline{b}_u p_{i.})^{1-\lambda^{-1}}}{\sum\sum_{(i,j)\in N} (\underline{a}_u p_{.j} + \underline{b}_u p_{i.})^{1-\lambda^{-1}}}$$

Based on the above optimal value, an iterative algorithm could be

$$\underline{p}_u^{(n+1)} \quad = \quad \frac{\left(\underline{a}_u p_{.j}^{(n)} + \underline{b}_u p_{i.}^{(n)}\right)^{1-\lambda^{-1}}}{\sum\sum_{(i,j)\in N} \left(\underline{a}_u p_{.j}^{(n)} + \underline{b}_u p_{i.}^{(n)}\right)^{1-\lambda^{-1}}},$$

for $n = 0, 1, \ldots$ with the initial choice $p_{ij}^{(0)} = \frac{1}{IJ}$ for all $(i, j) \in N$. We use the following stopping rule $\left|\frac{D_5^{(n+1)}}{D_5^{(n)}} - 1\right| \leq 10^{-6}$. Here also our iterative algorithm is convergent based on all the empirical studies that we have made in this regard. A formal mathematical proof is still remains an open problem.

6.6 Divergence measure D_6 under conditional and marginal information

Here, our test statistic reduces to

$$D_6 = \sum\sum_{(i,j)\in N} \left(\sqrt{p_{ij}} - \sqrt{\underline{a}_u p_{.j}}\right)^2 + \sum\sum_{(i,j)\in N} \left(\sqrt{p_{ij}} - \sqrt{\underline{b}_u p_{.j}}\right)^2. \tag{12}$$

Next, we consider the following Lagrangian function

$$F_6 = D_6\left(\underline{p}\right) + \sum_{t=1}^{r+1} \tau_t \left(\sum_{u=1}^{I} M_{tu} p_u - \theta_t\right),$$

where τ_t, $t = 1, 2, \ldots, (r+1)$ are $(r+1)$ Lagrangian multipliers. Now, to minimize F_6, we consider simultaneous solution of

$$\frac{\partial F_6}{\partial \underline{p}_u} = 0,$$

same as in (9). Consequently, the optimal value of $\underline{p}_u$ will be

$$\underline{p}_u^* = \frac{(\underline{a}_u p_{.j})^2 + (\underline{b}_u p_{.j})^2}{\sum\sum_{(i,j)\in N} \left\{(\underline{a}_u p_{.j})^2 + (\underline{b}_u p_{.j})^2\right\}}$$

Based on the above optimal value, an iterative algorithm could be

$$\underline{p}_u^{(n+1)} = \frac{\left(\underline{a}_u p_{.j}^{(n)}\right)^2 + \left(\underline{b}_u p_{.j}^{(n)}\right)^2}{\sum\sum_{(i,j)\in N}\left\{\left(\underline{a}_u p_{.j}^{(n)}\right)^2 + \left(\underline{b}_u p_{.j}^{(n)}\right)^2\right\}}$$

for $n = 0, 1, \ldots$ with the initial choice $p_{ij}^{(0)} = \frac{1}{IJ}$ for all $(i, j) \in N$. We use the following stopping rule $\left|\frac{D_6^{(n+1)}}{D_6^{(n)}} - 1\right| \leq 10^{-6}$. Here also our iterative algorithm is convergent based on all the empirical studies that we have made in this regard. A formal mathematical proof is still remains an open problem.

7 Illustrative Examples

In these illustrative examples, we consider conditional probability matrices that are incompatible in nature. These examples, although not taken from a real life scenario, are representative of the fact that given an additional set of precise information, whether the two conditional distributions are compatible or not, and in case they are not, can we find something close to what we call as ε-compatibility. Prominent real life scenarios in which this might be useful are Bayesian networks, model building in classical statistical settings, and elicitation and construction of multiparameter prior distributions in Bayesian scenarios. The dimensions of the matrices A and B are taken to be either 3 or 4 in Examples 1 to 5. The matrix M for each example was easily constructed using `Mathematica` software. The results of the iterative algorithm for the examples are shown in Tables 1 to 3.

- **Example 1.** In this example, we illustrate the above defined method in a simple case. Consider the set (X, Y) of two variables taking values $1, 2, 3, 4$. Let us consider the associated conditional probability matrices, where $I = 4$ and $J = 4$ and

$$A = \begin{pmatrix} 0.27 & 0.4 & 0 & 0.10 \\ 0.18 & 0.20 & 0.50 & 0.40 \\ 0.55 & 0.20 & 0.30 & 0.25 \\ 0 & 0.20 & 0.20 & 0.25 \end{pmatrix},$$

and

$$B = \begin{pmatrix} 0.15 & 0.28 & 0.35 & 0.22 \\ 0.45 & 0 & 0.25 & 0.30 \\ 0.50 & 0.17 & 0.20 & 0.13 \\ 0 & 0.55 & 0.20 & 0.30 \end{pmatrix}.$$

Here, A and B are incompatible since they do not share even a common incidence matrix.

Suppose that we have the following information (from our informed expert) :

  - $E\left(X^2\right) = 7.49$;
  - $P(Y = 3) = 0.38$;
  - $P\left(X^2 = 9 \middle| Y = 2\right) = 0.37$;
  - $P\left(Y^2 = 1 \middle| X = 2\right) = 0.53$.

Here, we have $\underline{p} = (p_{11}, p_{12}, \ldots, p_{44})$. In this case all the above information can be summarized by our M matrix given as follows.

$$M = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 4 & 4 & 4 & 4 & 9 & 9 & 9 & 9 & 16 & 16 & 16 & 16 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0.37 & 0 & 0 & 0 & 0.37 & 0 & 0 & 0 & -0.63 & 0 & 0 & 0 & 0.37 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.47 & 0.53 & 0.53 & 0.53 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Subsequently, $\underline{\theta} = (1, 7.49, 0.38, 0, 0)$. The iterative algorithm results are given in Table 1. In all the examples we considered, the constraints were approximated to a relative absolute error of 10^{-6}. The algorithm was found to converge for a wide range of values of λ.

- **Example 2.** In this example, we consider the set $\{X, Y\}$ of two variables taking values $1, 2, 3$. Let us consider two conditional probability matrices, where $I = 3$ and $J = 3$ and

$$A = \begin{pmatrix} 0.35 & 0.43 & 0 \\ 0 & 0.57 & 0.42 \\ 0.65 & 0 & 0.58 \end{pmatrix},$$

and

$$B = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & 0 & \frac{1}{4} \end{pmatrix}.$$

Here also, one can easily examine that the matrices A and B are incompatible. Suppose that we have the following information:

  - $E(X|Y = 2) = 1.5372$;
  - $P\left(X^2 = 1\middle| Y = 1\right) = 0.4235$;
  - $E\left(X^2\middle| Y^2 = 4\right) = 3.2953$;
  - $P(X < 3|Y > 2) = 0.4367$.

Here, we have $\underline{p} = (p_{11}, p_{12}, \ldots, p_{33})$. Subsequently, in this case, our M matrix is

$$M = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & -0.5372 & 0 & 0 & 0.4728 & 0 & 0 & 1.4728 & 0 \\ 0.5865 & 0 & 0 & -0.4235 & 0 & 0 & -0.4235 & 0 & 0 \\ 0 & -1.2953 & 0 & 0 & 1.6147 & 0 & 0 & 7.6147 & 0 \\ 0 & 0 & 0.5733 & 0 & 0 & 0.5733 & 0 & 0 & -0.4367 \end{pmatrix}.$$

We have $\underline{\theta} = (1, 0, 0, 0, 0)$. The iterative algorithm results are given in Table 1.

- **Example** 3. Let us consider two conditional probability matrices, where $I = 3$ and $J = 3$ and

$$A = \begin{pmatrix} \frac{2}{7} & \frac{3}{7} & 0 \\ 0 & \frac{4}{7} & \frac{6}{7} \\ \frac{5}{7} & 0 & \frac{1}{7} \end{pmatrix},$$

and

$$B = \begin{pmatrix} \frac{2}{5} & \frac{3}{5} & 0 \\ 0 & \frac{1}{3} & \frac{2}{3} \\ \frac{3}{5} & 0 & \frac{2}{5} \end{pmatrix}.$$

Suppose that we have the following information:

- $P\left(X^2 = 1 \middle| Y = 3\right) = 0$;
- $P\left(X^2 = 9 \middle| Y \geq 1\right) = 0.3956$;
- $E\left(X \middle| Y^2 = 4\right) = 1.3726$;
- $P(Y > 2|X < 3) = 0.6849$.

Here, we have $\underline{p} = (p_{11}, p_{12}, \ldots, p_{33})$. Also in this case our M matrix is

$$M = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3956 & 0 & 0 & -0.3956 & 0 & 0 & 0.6044 & 1 & 1 \\ 0 & -0.3726 & 0 & 0 & 0.6374 & 0 & 0 & 0 & 2.6374 \\ -0.6849 & -0.6849 & -0.6849 & -0.6849 & -0.6849 & -0.6849 & 1 & 1 & 0 \end{pmatrix}.$$

Here, $\underline{\theta} = (1, 0, 0, 0, 0)$. The iterative algorithm results are given in Table 3.

Criterion	Optimal value	Matrix P	No. of iterations
D_1	0.002353209	$\begin{pmatrix} 0.0610 & 0.0557 & 0.0484 & 0.0219 \\ 0.0837 & 0.0103 & 0.0485 & 0.0489 \\ 0.2069 & 0.0377 & 0.1111 & 0.0461 \\ 0.0000 & 0.0815 & 0.1345 & 0.0078 \end{pmatrix}$	8
D_2	0.003245132	$\begin{pmatrix} 0.0583 & 0.0571 & 0.0492 & 0.0227 \\ 0.0837 & 0.0132 & 0.0465 & 0.0489 \\ 0.2062 & 0.0352 & 0.1132 & 0.0460 \\ 0.0000 & 0.0821 & 0.1351 & 0.0084 \end{pmatrix}$	11
D_3	0.002129779	$\begin{pmatrix} 0.0681 & 0.0741 & 0.1590 & 0.0000 \\ 0.0841 & 0.0000 & 0.0419 & 0.0949 \\ 0.0000 & 0.0624 & 0.1004 & 0.1448 \\ 0.0389 & 0.0763 & 0.0547 & 0.0000 \end{pmatrix}$	10
D_4	0.005605187	$\begin{pmatrix} 0.0686 & 0.0711 & 0.1538 & 0.0000 \\ 0.0864 & 0.0000 & 0.0419 & 0.0919 \\ 0.0000 & 0.0635 & 0.1026 & 0.1438 \\ 0.0415 & 0.0781 & 0.0561 & 0.0000 \end{pmatrix}$	12
D_5	0.002219034	$\begin{pmatrix} 0.0682 & 0.0704 & 0.1523 & 0.0000 \\ 0.0868 & 0.0000 & 0.0420 & 0.0901 \\ 0.0000 & 0.0640 & 0.1049 & 0.1418 \\ 0.0426 & 0.0798 & 0.0571 & 0.0000 \end{pmatrix}$	10
D_6	0.001537571	$\begin{pmatrix} 0.0686 & 0.0705 & 0.1528 & 0.0000 \\ 0.0867 & 0.0000 & 0.0420 & 0.0915 \\ 0.0000 & 0.0637 & 0.1037 & 0.1432 \\ 0.0421 & 0.0787 & 0.0565 & 0.0000 \end{pmatrix}$	9

Table 1: Minimal (ϵ) incompatibility results for Example 1.

The small values of divergence in Tables 1 to 3 are quite encouraging. There is no evidence that D_1 decreases/increases with the dimension or the values in A and B. The nature of the results were similar for a wide range of other A, B and for A, B of higher dimensions. A similar approach in the case of continuous probability models still remains an open problem and will be taken up in a future article.

Criterion	Optimal value	Matrix P	No. of iterations
D_1	0.000291763	$\begin{pmatrix} 0.0924 & 0.1638 & 0.0000 \\ 0.0000 & 0.1468 & 0.2605 \\ 0.1030 & 0.0000 & 0.2335 \end{pmatrix}$	6
D_2	0.001796547	$\begin{pmatrix} 0.1113 & 0.1469 & 0.0000 \\ 0.0000 & 0.1734 & 0.1302 \\ 0.2013 & 0.0000 & 0.2365 \end{pmatrix}$	9
D_3	0.001796547	$\begin{pmatrix} 0.1142 & 0.1478 & 0.0000 \\ 0.0000 & 0.1737 & 0.1320 \\ 0.2009 & 0.0000 & 0.2316 \end{pmatrix}$	10
D_4	0.001652207	$\begin{pmatrix} 0.1107 & 0.1472 & 0.0000 \\ 0.0000 & 0.1726 & 0.1298 \\ 0.2013 & 0.0000 & 0.2384 \end{pmatrix}$	11
D_5	0.001079299	$\begin{pmatrix} 0.1104 & 0.1471 & 0.0000 \\ 0.0000 & 0.1723 & 0.1293 \\ 0.2013 & 0.0000 & 0.2396 \end{pmatrix}$	9
D_6	0.000609597	$\begin{pmatrix} 0.1103 & 0.1472 & 0.0000 \\ 0.0000 & 0.1721 & 0.1283 \\ 0.2013 & 0.0000 & 0.2398 \end{pmatrix}$	8

Table 2: Minimal incompatibility results for Example 2.

Criterion	Optimal value	Matrix P	No. of iterations
D_1	0.001992807	$\begin{pmatrix} 0.1052 & 0.1691 & 0.0000 \\ 0.0000 & 0.0587 & 0.07112 \\ 0.3062 & 0.0000 & 0.2895 \end{pmatrix}$	7
D_2	0.001453787	$\begin{pmatrix} 0.0921 & 0.1654 & 0.0000 \\ 0.0000 & 0.0632 & 0.0817 \\ 0.2931 & 0.0000 & 0.3045 \end{pmatrix}$	8
D_3	0.002309232	$\begin{pmatrix} 0.0961 & 0.2339 & 0.0000 \\ 0.0000 & 0.1799 & 0.0691 \\ 0.1194 & 0.0000 & 0.3014 \end{pmatrix}$	8
D_4	0.008158526	$\begin{pmatrix} 0.0904 & 0.2317 & 0.0000 \\ 0.0000 & 0.1721 & 0.0653 \\ 0.1224 & 0.0000 & 0.3182 \end{pmatrix}$	8
D_5	0.004180903	$\begin{pmatrix} 0.0860 & 0.2380 & 0.0000 \\ 0.0000 & 0.1885 & 0.1007 \\ 0.1109 & 0.0000 & 0.2759 \end{pmatrix}$	8
D_6	0.00251268	$\begin{pmatrix} 0.0936 & 0.2246 & 0.0000 \\ 0.0000 & 0.1808 & 0.0755 \\ 0.1317 & 0.0000 & 0.2935 \end{pmatrix}$	8

Table 3: Minimal incompatibility results for Example 3.

7.1 Some observations on the concept of ϵ-compatibility

The advantage of the definition of ϵ- compatibility utilized in this article is that the degree of incompatibility could be determined by standard linear programming techniques which has been advocated by Arnold et al. (2001). However, this simplicity comes at a cost. If the information is found to be, say, .0058 compatible it is difficult to interpret the meaning of the quantity $.0058$. It is obvious that 0-compatible means completely compatible and 0.01 compatible is better than 0.023 compatible but no interpretation of 0.01 or 0.02 seems available in the literature.

8 Concluding remarks

The problem of finding most nearly compatible distribution(s) starting from two given conditionals (that are incompatible) is not new in the literature. However, there is a scarcity of scholarly work on this topic when in addition to complete specification of two given conditional probability matrices, our informed expert has some additional information in the form of say, conditional percentiles and/or conditional moments etc., among others. Arnold et al. (2001) has provided a brief overview on the issue of finding minimally incompatible distribution in the presence of additional information. However, the role of various existing as well as comparatively newly defined pseudo-divergence measures in search for a minimally incompatible under the presence of additional information has not been adequately addressed. In this paper, we explore the relative performance (equivalently the applicability) of some of the well-known measures of divergence in finding a most nearly compatible distribution in the presence of additional information. The survey made in this paper is far from complete. Compatibility in higher dimensions, such as, given three conditional matrices, say X given Y and Z; Y given X and Z; and Z given X and Y in the presence of additional information (in terms of marginal/conditional moments, percentiles etc.) will be the subject matter of a separate article.

References

Arnold, B.C., E. Castillo, and J.M. Sarabia (1999). *Conditional Specification of Statistical Models*. New York: Springer Verlag.

Arnold, B.C., E. Castillo, and J.M. Sarabia (2001). Quantification of incompatibility of conditional and marginal information. *Communications in Statistics: Theory and Methods 30*, 381–395.

Arnold, B.C. and D.V. Gokhale (1994). On uniform marginal representations of contingency tables. *Statistics and Probability Letters 21*, 311–316.

Arnold, B.C. and D.V. Gokhale (1998). Distributions most nearly compatible with given families of conditional distributions. the finite discrete case. *Test 7*, 377–390.

Arnold, Barry C, Enrique Castillo, José-Mariá Sarabia, Barry C Arnold, Enrique Castillo, and José-Mariá Sarabia (1992). Conditional specification. In *Conditionally Specified Distributions*, pp. 1–6. Springer.

Arnold, Barry C and S James Press (1989). Compatible conditional distributions. *Journal of the American Statistical Association 84*(405), 152–156.

Borzadaran, GR Mohtashami and M Amini (2010). Information measures via copula functions. *Journal of Statistical Research Iran 7*, 47–60.

Cacoullos, Theophilos and H Papageorgiou (1983). Characterizations of discrete distributions by a conditional distribution and a regression function. *Annals of the Institute of Statistical Mathematics 35*, 95–103.

Cressie, Noel and Timothy RC Read (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology 46*(3), 440–464.

Gelman, A. and T.P. Speed (1993). Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society, Series B 55*, 185–188.

Ghosh, Indranil (2011). *Inference for the bivariate and multivariate hidden truncated Pareto (type II) and Pareto (type IV) distribution and some measures of divergence related to incompatibility of probability distribution*. University of California, Riverside.

Ghosh, I. and N. Balakrishnan (2015). Study of incompatibility or near compatibility of bivariate discrete conditional probability distributions through divergence measures. *Journal of Statistical Computation and Simulation 85*, 117–130.

Ghosh, I. and S. Nadarajah (2017). On the construction of a joint distribution given two discrete conditionals. *Studia Scientiarum Mathematicarum Hungarica 54*, 178–204.

Ghosh, Indranil and SM Sunoj (2024). Copula-based mutual information measures and mutual entropy: A brief survey. *Mathematical Methods of Statistics 33*(3), 297–309.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Mosteller, Frederick (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association 63*(321), 1–28.

Nadarajah, Saralees and Kostas Zografos (2003). Formulas for rényi information and related measures for univariate distributions. *Information Sciences 155*(1-2), 119–138.

Pardo, L. (2006). *Statistical inference based on divergence measures*. Boca Raton, USA: Chapman & Hall/CRC Press.

Rényi, Alfréd (1961). On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, Volume 4, pp. 547–562. University of California Press.

Shannon, Claude E (1951). Prediction and entropy of printed english. *Bell system technical journal 30*(1), 50–64.

Wesolowski, J (1996). A new conditional specification of the biva riate po isson cond it iona is distribution'. *Statistica Neerlandica 50*(3), 390–393.

Zografos, K and S Nadarajah (2005). Expressions for rényi and shannon entropies for multivariate distributions. *Statistics & Probability Letters 71*(1), 71–84.

Spanish Journal of Statistics
Vol. 6 No. 1 2024, Pages 39–59
doi:https://doi.org/10.37830/SJS.2024.1.04

Regular Article

Census-based comparability of data on literacy processes in Western Europe

José Manuel Gutiérrez

Universidad de Salamanca, jmgut@usal.es

Received: May 20, 2024. Accepted: December 4, 2024.

Abstract:

A comparative picture of the literacy processes in Western Europe on the eve of and during the Second Industrial Revolution is provided, taking censual literacy rates as a yardstick to measure and compare literacy in different countries. Censual data are obtained and analysed from the original source. If only partial or insufficient censual data are available, literacy is assessed as if given by full censual data. A set of comparable (as far as possible) literacy data is built. Four literacy groups result. The area of Western Europe where mass literacy was first achieved was the German-speaking or culturally highly Germanised zone. Britain and Sweden turn out to be in the same cluster as France. The periphery of Western Europe shows a well-known pattern of delayed literacy development.

Keywords: Historical censuses, literacy, nineteenth century, Western Europe

MSC: 91F10, 62P25

1 Introduction

This article provides a comparative picture of the literacy processes in Western Europe on the eve of and during the Second Industrial Revolution (generally dated between 1870 and 1914), taking as a yardstick "modern census" data. The aim is not the construction of explicative models, but the harmonization of data (as far as possible) to enable comparisons. The entire adult population is intended; in literacy processes, improvements in literacy of children or young men translate only gradually into the general literacy rates.

In order to estimate the literacy level of a population, we face considerable problems, conceptual and practical, especially when we go back in time[1]. The advent of "modern censuses" in the mid-19th century opened new possibilities for the measurement of literacy. In modern censuses, data were

[1] As for the pre-statistical age, the main tool of analysis is considering who could sign and who could not sign in documents (such as marriage certificates, deeds, wills, etc.), and even the quality of the signatures. Apart from the issue of how representative of the population is the sample in each case, the ability of an individual to write his/her name does

Published by the Spanish National Statistical Institute

obtained on all individuals present in the household on the specified census day. Information was self-reported by the household heads through household forms (later individual forms). A field force of professional enumerators was employed to assist in the process from house to house (especially if there was no one in the house who could write) and collect the forms.[2]

All modern censuses had a similar basic methodology, and thus comparisons between countries are made easier[3]. In this paper, we shall take censual literacy rates as a yardstick to measure and compare literacy in different countries. Censual data are obtained and analyzed from the original source, and the methodology of each census is considered. If only partial or insufficient censual data are available, we shall try to assess literacy as if given by full censual data.

Where modern censuses provide literacy data, some points are to be specified for their comparability: (1) the literacy criterion (i.e., when a person is considered literate); (2) the minimum age limit (i.e., the age from which literacy is considered, as the illiteracy of babies is irrelevant); (3) how persons unspecified for literacy are dealt with. The different timing of the censuses is also an issue, especially if there is a large temporal gap in the census data available (e.g., there are no French censual literacy data between 1872 and 1900).

Certainly, the most serious comparability problem arises when census data are lacking. There is a swathe of land in northern Europe, from the Netherlands to Sweden, including Great Britain, where literacy data obtained using modern statistical criteria and covering the whole population are not available, or they are available very late (as in Sweden)[4]. At any rate, we shall consider here three types of countries. Firstly, countries where we have census data for the whole country, sufficient for our purpose, although perhaps with some minor additional estimation work. Secondly, countries for which we have partial census data, but which allow us to make well-founded extrapolations. Thirdly, countries for which we have partial or late censual data that are insufficient, but which, supplemented with other additional data, allow us to draw reasonable conclusions. We shall not consider countries with no literacy census data at all (such as the Netherlands or Norway).

We intend to study the development of the literacy process and when high literacy, indicated by the 75% of the population (over a certain age) threshold[5], was reached in each country.

Literacy is an abstract and general tool for the acquisition and communication of knowledge. Consequently, it expands the individual's decision-making capacity and scope of freedom. She who teaches literacy knows that the skills she is imparting can be used against her own ideas and expectations. Not for nothing did the President of the Royal Society in 1807 oppose (successfully) in the House of Lords a bill to provide elementary schools in England[6]: "... the project... of giving education to the labouring classes of the poor... would enable them to read seditious pamphlets, vicious books, and publications against Christianity; it would render them insolent to their superiors."

not entail, in principle, a general ability to read or write, although there can be statistical correlations (see Furet and Sachs (1974)).

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

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

[4] This is not the place to answer some questions that naturally arise. Why did the UK government choose to provide us with literacy data for Ireland, but omitted doing the same for England, Wales, and Scotland? Why did the Swedish government wait until 1930 to include literacy questions in the census?

[5] The selection of a threshold of high literacy has a certain degree of arbitrariness. A lower bound might be the 70% set by seminal Bowman and Anderson (1963) for the higher threshold of literacy as regards economic development. On another note, literacy benefits social and personal aspects beyond economic development. At any rate, the choice of 70% instead of 75% would not lead to any change in our resulting classification.

[6] Quoted in Cipolla (1969), pp. 65–66.

It must be considered when the data refer to literacy (ability to read and write) or semi-literacy (ability to read). In this paper, we shall refer to literacy[7]. At any rate, certain forms of "restricted semi-literacy", in which the ability to "read" exclusively a limited set of *known* texts is acquired, although may be useful knowledge (or a mechanism of ideological control), fall short of the concept of literacy or semi-literacy as a general tool.

We conclude that there are four groups according to when the 75% literacy threshold is reached. The first group corresponds to countries where the threshold had been already reached at the beginning of the Second Industrial Revolution, in the years 1871–1880. It comprises the Austro-German bloc and some neighbouring areas strongly influenced by it. The second group surrounds the first group to the west and north and includes the countries where the 75% level was reached before the First World War (the axis France-Great Britain belongs here). The third group corresponds to areas peripheral respect to the first group, where the 75% threshold was reached by the Second World War. The fourth group includes only the outermost Portugal.

The article is organized as follows. Section 2 introduces the four literacy groups. The composition of these groups is justified in Sections 3, 4, and 5. Section 6 provides some final remarks.

2 Literacy groups

We consider the following classification of countries in Western Europe[8] as for the development of the literacy process from the mid-nineteenth century to the Second World War (see Figure 1):

GROUP I. *Countries where the literacy rate reached 75% already by the period 1871–1880*: Germany, Austria, Czechia and Denmark.

GROUP II. *Countries where the 75% level was reached before the First World War*: Ireland, Belgium, France, Slovenia, Great Britain and Sweden.

GROUP III. *Countries where the 75% threshold was reached by the Second World War*: Spain, Finland, Italy.

GROUP IV. *Countries where the 75% threshold was reached after the Second World War*: Portugal.

We shall justify this classification in the following three sections, which correspond to three levels of the quality of available data.

3 Literacy from complete censual data

3.1 Literacy from literacy censual data

Table 1 shows the literacy data of Western European countries for which "modern census" data exist in the nineteenth century. For each country and census[9], three percentages of literacy are stated: for men and women, separated by a hyphen, and the overall percentage in the bottom row. The

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

[8] Present-day countries will be considered, even when inaccuracies are inevitable because of alterations in borders and displacements of populations. For practical reasons, the British Islands will be divided into (the whole of) Ireland and Great Britain. Transleithania (see below) and Poland have not been examined, as the drastic changes of borders make data reconstruction very difficult.

[9] After the *Compromise* (*"Ausgleich"*) *of 1867*, the Austrian Empire was transformed into the dual monarchy of Austria-Hungary, constituted by two parts, with their respective parliaments and governments: *Cisleithania* (the Austrian part) and *Transleithania* (lands of the "Archiregnum Hungaricum"). Cisleithania was divided into 16 *crown lands* (*"Kronländer"*), each one with its own land parliament ("Landtag"). The data of present-day Austria, Czechia and Slovenia have been extracted

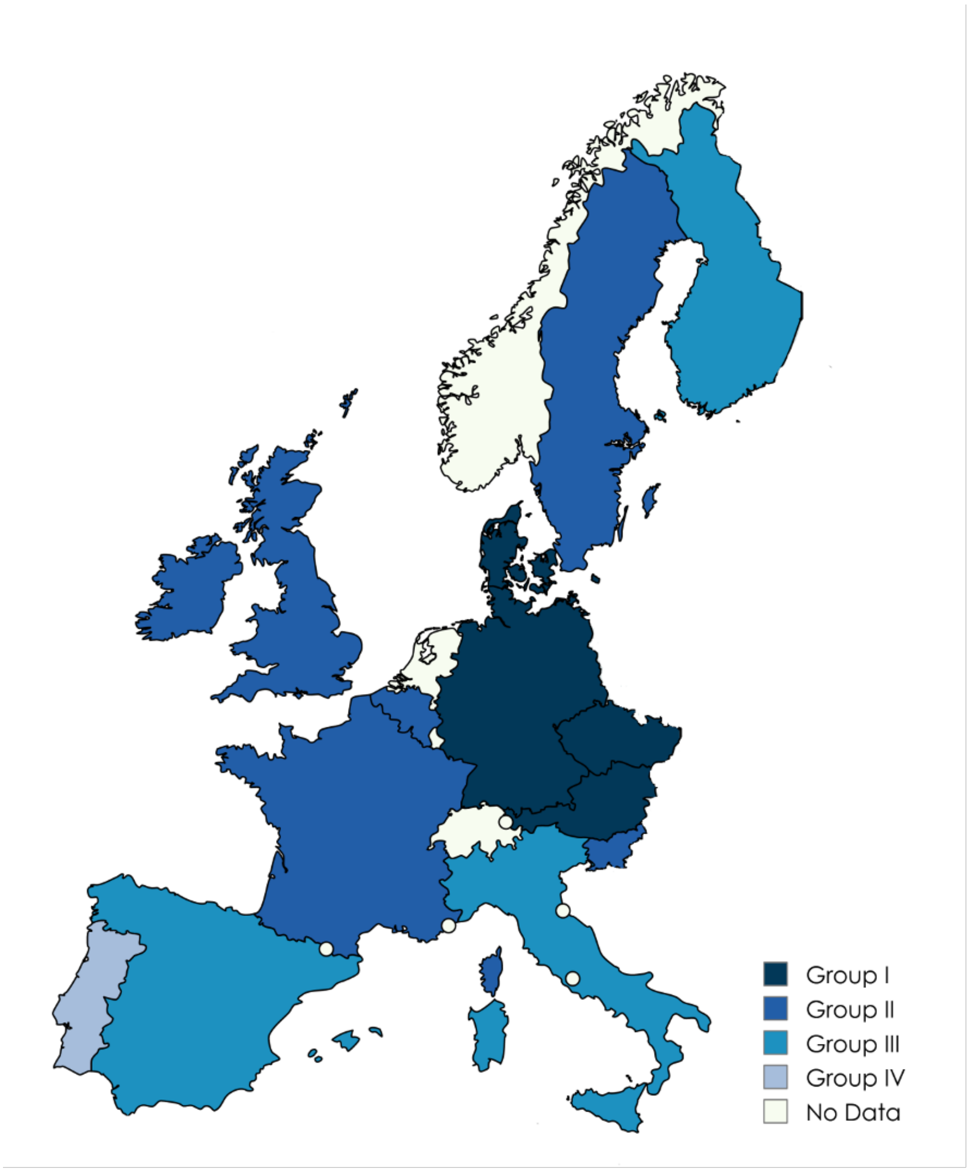


Figure 1: Literacy process in Western Europe (census-like data).

minimum age limits are indicated in brackets under the name of every country; further clarifications are in the Appendix. Portugal is not included in this table, as only questions on semi-literacy (i.e., ability to read) were posed in the censuses.

from the census of Cisleithania, following the historical divisions of the time (there is no full correspondence between old and new borders).

Year	Ireland	Spain	Italy	Belgium	France	Prussia	Austria	Czechia	Slovenia	Finland
	(≥ 5)	(≥ 10)	(≥ 12, ≥ 10)	(≥ 15)	(≥ 6, ≥ 10)	(≥ 10)	(≥ 6, ≥ 11)	(≥ 6, ≥ 11)	(≥ 6, ≥ 11)	(≥ 10, ≥ 15)
1841	37-18									
	28									
1851	41-25									
	33									
1860/61	49-34	38.9-11.2	30.4-14.0							
	41.3	24.8	22.2							
1866				59.1-51.6	61.5-49.9					
				55.4	55.7					
1871/72	54.7-44.3				63.4-53.2	89.2-83.6				
	49.4				58.3	86.3				
1877		43.5-17.9								
		30.3								
1880/81	62.6-56.1		45.6-27.5	71.5-64.1			82.7-77.0	88.0-79.6	39.9-28.8	16.2-10.2
	59.3		36.5	67.8			79.8	83.6.	34.1	13.1
1887		48.2-22.8								
		35.1								
1890/91	72.4-68.9			76.4-69.9			89.7-85.2	94.1-88.1	55.7-44.4	25.9-19.3
	70.6			73.1			87.4	91.0	49.8	22.5
1900/01	80.3-78.5	52.7-30.5		82.9-78.0	86.5-80.6		93.1-90.3	96.4-92.9	72.7-65.0	41.1-36.5
	79.4	41.2		80.4	83.5		91.7	94.6	68.7	38.8
1910/11	84.7-83.7	57.6-38.6		88.3-84.9	90.3-85.9		95.9-94.5	97.7-95.7	84.2-79.1	57.4-53.3
	84.2	47.7		86.6	88.1		95.2	96.7	88.5	55.3

Table 1: Literacy rates in Western Europe before WW1. Sources and notes in the Appendix.

Country	1920	1930	1940
Finland	71.0-68.8	84.9-83.4	
(≥ 15)	69.9	84.1	
Spain	63.9-48.1	74.8-59.4	82.7-71.5
(≥ 10)	55.7	66.8	76.8

Table 2: Literacy rates in Finland and Spain, 1920-1940. Sources and notes in the Appendix.

From the data in Table 1, the German-speaking countries (Prussia and Austria) and neighbouring Czechia belong to Group I. Ireland, Belgium, France, and Slovenia[10] are in Group II. From Table 1 and Table 2, Finland[11] and Spain are in Group III.

Border alterations and (mainly ethnically based) population displacements are to be considered in interpreting Table 1. Thus, on the one hand, the extent to which the data for "Prussia" can be extrapolated to "Germany" is discussed in Section 4. On the other hand, it is relevant to analyse the literacy rates of ethnic minorities in Austria, the Czech Republic, and Prussia.

As for the crown lands of present-day Austria, while in the exclusively German-speaking crown lands, except in Salzburg, the literacy rates (1880 census, population aged 6 or over) were above the Prussian average (1871 census, population aged 10 or over), in Styria or Carinthia, with significant Slovene-speaking minorities (Styria 32.7%, Carinthia 29.7%), the rates were considerably lower (Styria 62.6%, Carinthia 52.9%). In Tyrol (which at that time included Trentino), with an impor-

[10] The corresponding column in Table 1 gathers the data of Carniola and Gorizia & Gradisca, the two crown lands of Cisleithania with a Slovene-speaking majority. Literacy grew widespread there only as Slovene-speaking schools became available. In Carniola a small German-speaking minority (among an overwhelming Slovene-speaking majority) had controlled local politics, while in Gorizia & Gradisca the Italian-speaking minority (about one third of the population) had historically prevailed over the Slovene-speaking majority. In general, the creation of Slav-language schools in Cisleithania progressed throughout the 19th century (starting with the Czech ones), although it was only at the end of the Habsburg monarchy that sufficient levels were achieved; it should be noted that some parents with Slav mother tongue (especially Jewish parents) preferred their children to study in German-language schools. See Urbanitsch (2021).

[11] Finland was part of the Kingdom of Sweden until 1809 and then became part of the Russian Empire until 1917 as the autonomous Grand Duchy of Finland. Note that Finland followed the Swedish model, based on home instruction of the ability to read *known texts*, with high *restricted semi-literacy* and low literacy until modern school systems were introduced (see below in the subsection on Sweden).

tant Italian-speaking minority (45.4%), the overall rate was 81.3%, whereas the rate for the German-speaking districts was 87.0%.

In contrast with Austria, in Czechia there was no significant difference between the literacy levels of Germanophones and Czech speakers[12], and the detailed censual data available show that this was the case at least since the mid-nineteenth century[13].

Politics banned any ethnic information in the 1871 Prussian census[14]. At any rate, some words must be said in favour of the perspicacity of the officials of the Prussian Statistical Office in the 1870s, whose head was none other than Ernst Engel. The following comment in an article presenting the results of the 1871 census for literacy and confession could be read in the official journal of the Prussian Statistical Office[15]: "The table, which contains all these data for each department and the like, shows that although the Catholics in most parts of the country have less favourable figures than the Protestants, the size of the difference is mainly due to the greater proportion of Catholics in the Polish-speaking population"[16]. At that time, the Prussian Statistical Office had no reliable data on the mother tongue (or on the usual language) of the population. Language questions were introduced in a general Prussian census for the first time in 1861, but the results were unreliable, as the number of non-German speakers included only those who did not know German. The Ministry of Interior prohibited the posing of any question on language until the 1890 census[17] (the year of Bismarck's dismissal), despite the attempts of senior statisticians like Richard Böckh (see Labbé (2007)). In fact, a swathe of land along the far east of the country (Prussia proper, Posen, and Upper Silesia) had literacy rates below 75%, with a minimum of 57.1% in Bromberg. This area corresponded to the districts with a sizeable Polish-speaking minority (Prussia proper) or a Polish-speaking majority (Posen and Upper Silesia). The rest of the country had literacy rates above approximately 90%, except for part of Pomerania (83.3% in Köslin and 84.1% in Stralsund)[18].

[12] In Czechia (data of 1880), 35.9% of the population was Germanophone (37.2% in Bohemia, 29.4% in Moravia, and 48.9% in Austrian Silesia).

[13] The 1910 Austrian (Cisleithanian) census provides literacy rates by language group, further disaggregated by age interval (which allows for backward projection of results). The literacy rate for males aged 61-70 was 91.9% for German-speakers and 93.8% for Czech-speakers, and for those over 70 the figures were 89.1% and 91.8%, respectively. For females aged 61-70 the rates were 87.3% for German-speakers and 85.1% for Czech-speakers, and for those over 70 the rates were 82.2% and 78.6%, respectively. In that census, the overall male rates (population aged 11 or over) were 97.0% for German-speakers and 97.9% for Czech-speakers, and the female rate was 95.5% for both language groups.

[14] In 1890, 10.1% of the Prussian population were Polish-speaking. About the discriminatory policies of the Prussian government against the Polish ethnic minority, see Deutscher Bundestag (2019) and Kerstin et al. (2020).

[15] Engel (1874), p. 150. The journal was edited ("redigirt" (sic)) by the head of the Statistical Office. Besides, this article is signed with the initials "K. B.".

[16] "Aus der Tabelle, welche all diese Daten für jeden Regierungsbezirk und dgl. enthält, geht hervor, dass zwar die Katholiken in den meisten Landestheilen ungünstigere Zahlen aufweisen, als die Protestanten, dass aber die Grösse der Differenz vorzugsweise durch den stärkeren Antheil der Katholiken an der polnisch redenden Bevölkerung veranlasst wird."

[17] The language census of 1890 is methodologically rigorous, despite some flaws (see Belzyt (1998) for a critical analysis, with proposed corrections disaggregated by department (Polish) or district (Danish)).

[18] The correlation coefficient between the literacy rate (1871 census) and the proportion of the population having Polish as mother tongue (1890 census) is $\rho = -0.8622$ (we consider disaggregation by department and include Masurian and Kashubian speakers within Polish speakers). Using the 1890 and other linguistic data, Kerstin et al. (2020) shows that Prussian literacy in 1871 is to a large extent explained by having Polish as mother tongue or not, whereas whether the individuals are Protestant or Catholic is not significant. A parsimonious model of literacy in Prussia where only ethnic (linguistic) and religious regressors are considered, with disaggregation by department and data of the censuses of 1871 and 1890, leads us to the same conclusion as Kerstin et al. (2020).

3.2 Literacy from semi-literacy censual data

The Portuguese censuses did not provide data on literacy, but only on semi-literacy. The same occurs with the Italian censuses, except in 1861 and 1881, in which both literacy and semi-literacy figures were given. The data appear in Table 3.

In contrast with France, the percentage of semi-illiterates (people who can read but not write) was small in Italy already in the 1860s[19]. The reader can adjust the Italian data of Table 3 slightly downwards to obtain an estimation of literacy[20]. A different question is to gauge the effect of some heterodox instructions given to the enumerators in the censuses of 1921 and 1931, at loggerheads with the principle that censuses are to reflect what it is, and not what it should be[21]. The Italian census of 1941 was never carried out, but we may place Italy in Group III.

In the case of Portugal, we have no censual reference to estimate the percentage of semi-illiterates. In any case, as the literacy rate is less than or equal to the semi-literacy rate, Portugal belongs to Group IV[22].

Country	**1900/01**	**1911**	**1920/21**	**1930/31**	**1940**
Italy	58.3-45.4	68.4-57.5	76.7-70.0	82.2-74.8	
(≥ 10)	51.8	62.8	73.2	78.4	
Portugal	36.1-18.2	40.4-23.0	43.6-27.2	49.6-31.1	58.5-41.5
(≥ 10)	26.6	31.1	34.8	39.8	49.6

Table 3: Semi-literacy rates in Italy and Portugal, 1900-1940. Sources and notes in the Appendix.

[19] In Italy the percentage of semi-illiterates in 1861 was 3.9% for men, 5.5% for women and 4.7% overall. In France the figures in 1866 were 9.7% for men, 13.2% for women and 11.5% overall. See Diebolt et al. (2005) on the rise of mass schooling in France.

[20] The Italian percentage of semi-illiterates in 1881 was 1.2% for men, 3.4% for women and 2.3% overall. These data are approximately one percentage point inferior to the Spanish ones of 1887 (2.2% for men, 4.5% for women and 3.4% overall). As guidance for the adjustment of the Italian figures of Table 3, the Spanish data of semi-illiteracy are available until 1930: in 1900 they were 1.6% for men, 3.4% for women and 2.6% overall; in 1910 they were 1.0% for men, 2.3% for women and 1.7% overall; in 1920 they were 0.5% for men, 1.1% for women and 0.8% overall; in 1930 they were 0.4% for men, 1.1% for women and 0.8% overall (Vilanova Ribas and Moreno Julià (1992)).

[21] All children registered in a school were to be automatically considered literate, even those six years old, with the argument (see the Italian census of 1931, Vol. IV, p. *95) that "at the date of the census, i.e. at the end of April, those enrolled in the first elementary class should, on the basis of the school programmes, already know the entire alphabet and therefore be able to read a printed text" ("gli iscritti alla prima classe elementare, alla data del censimento, cioè alla fine di aprile, dovevano, in base ai programmi scolastici, conoscere già tutto l'alfabeto ed essere in grado, quindi, di leggere uno stampato"). This criterion was implemented in the census of 1931, and to some extent (perhaps) in that of 1921 (see ibid.): "in 1921, during the counting, the cited conventional norm of considering *all* schoolchildren literate could not be rigorously applied" ("nel 1921, durante gli spogli, non potè essere applicata rigorosamente la citata norma convenzionale di considerare alfabeti *tutti* gli scolari").

[22] The Portuguese rate was 74% in the 1970 census and 79% in the following census, in 1981 (see Candeias (2004)).

4 Literacy from partial censual data

4.1 German literacy from Prussian censual data

The literacy data of the 1871 Prussian census were a sensation among the European élites[23]. Many found them scary also: that very same year the German Reich had been founded, after the military defeat of France.

In fact, 62.4% of the population of the new Germany (1871 census, excluding the annexed Alsace-Lorraine) corresponded to the Kingdom of Prussia. Besides, the Prussian literacy data of 1871 were not particularly Prussian, but German, as we shall argue now, and therefore "Prussia" can be replaced by "Germany" in Group I.

In 1866, after the Austro-Prussian War, Prussia incorporated territories of several German states, increasing its population by 21.8%[24]. The expansion of the Kingdom of Prussia permits obtaining, in its 1871 census, literacy data of territories in which the level of literacy achieved was not attributable to the action of the Prussian authorities. After 1850 we have the following five groups of annexations:

(1) *Hohenzollern-Sigmaringen and Hohenzollern-Hechingen*. In 1850, after the abdication of their respective princes, the principalities of Hohenzollern-Sigmaringen and Hohenzollern-Hechingen were incorporated into Prussia, becoming the department (*Regierungsbezirk*[25]) of Sigmaringen[26].

(2) *Schleswig and Holstein*. After the Second Schleswig War (1864), the Duchies of Schleswig, Holstein and Lauenburg (until then held by the Danish monarch in personal union) were ceded to Prussia and Austria. In 1866 Prussia assumed control of the three territories, which were eventually integrated into the department of Schleswig[27].

(3) *Hanover, Hesse-Kassel, Nassau and Frankfurt*. Annexed in 1866. The Kingdom of Hanover became the province of Hannover, divided into six departments. The Electorate of Hesse (Hesse-Kassel) turned into the department of Kassel, the Duchy of Nassau and the Free City of Frankfurt became the department of Wiesbaden.

(4) *Some territories of Hesse-Darmstadt*. The Duchy of Hesse (Hesse-Darmstadt) had to cede the districts (*Kreise*[28]) of Biedenkopf and Vöhl to Prussia in 1866, which were incorporated into the departments of Wiesbaden and Kassel, respectively[29].

[23] The 1871 census was the first census of the newly unified German Reich. Member states had to provide information on certain core variables to the *Statistisches Reichsamt*, but they were free to collect some additional statistics. Taking advantage of this possibility, literacy data were recorded in Prussia, for the first and last time. See Michel (1985) and Gehrmann (2012) on modern censuses in the German states prior to 1871.

[24] Data as of 1867 (see the Prussian census of 1871, p. 6–7).

[25] We translate *"Regierungsbezirk"* by "department". There were 36 *Regierungsbezirke* in Prussia in 1871; they were comparable in size to French *"départements"* (admittedly, the population of a *Regierungsbezirk* was on average greater by approximately one third than that of a *département*).

[26] Despite the twenty years elapsed until the census, the very high literacy rate (97.2%, the highest of all departments, except for the capital Berlin) implies that literacy was widespread at all ages in 1871 and thus that literacy levels were already high before the incorporation into Prussia.

[27] The Duchy of Lauenburg was not formally incorporated into the Kingdom of Prussia until 1876, and thus its results were not included in the 1871 Prussian census. Lauenburg was small (47,347 inhabitants in 1871).

[28] We translate *"Kreis"* by "district". Districts could consist of a large enough municipality (*"Stadtkreise"*) or of several municipalities. During industrialisation and the concomitant process of urban growth, the number of *Stadtkreise* grew steadily.

[29] Besides, in 1866 the Grand Duchy of Hesse inherited the Landgraviate of Hesse-Homburg (27,563 inhabitants in 1865) but had to cede its territory to Prussia later that year, and then it was divided between the departments of Koblenz and Wiesbaden.

(5) *Some territories of Bavaria*. Apart from a tiny exclave, the Kingdom of Bavaria ceded to Prussia in 1866 the districts of Gersfeld and Orb[30]. Both were included in the department of Kassel.

The North German Confederation was created in August 1866. The new Prussia enlarged by the annexations made up roughly 80% of the population of the Confederation. Against the background of the Prussian global literacy rate, 86.3% in the 1871 census, it is worth considering the literacy rates in that census of the nine new departments formed with the annexations of Prussia in 1866, corresponding to territories of its enemies in the Austro-Prussian War: Schleswig, 95.1%; Hannover, 94.3%; Hildesheim, 89.3%; Lüneburg, 93.5%; Stade, 92.2%; Osnabrück, 94.8%; Aurich, 91.1%; Kassel, 92.5%; Wiesbaden, 97.2%.

The south German states of Hesse-Darmstadt[31], Bavaria, Württemberg and Baden did not enter the North German Confederation. From the Prussian census of 1871, the literacy of the (reputedly poor) territories (Biedenkopf and Vöhl) just annexed from Hesse-Darmstadt can be obtained: 92.9%; the literacy rate of the territories (Gersfeld and Orb) annexed from Bavaria was 94.4%. As for Württemberg and Baden, the Prussian data of the annexed, rural Sigmaringen, a narrow strip of land sandwiched between them, are to be considered: the literacy rate was 97.2%.

4.2 Danish literacy from Prussian censual data

In the section on international comparisons of the Italian census of 1881, a letter from the Italian Statistical Office to its Danish counterpart is mentioned, as is the reply from its director:

"The Director of Statistics of Denmark has written to us that, since education is compulsory for children from 7 to 14 years of age, at this last age everyone must produce the certificate of knowing how to read and write. However, this does not prevent there being illiterates, since there are those who have forgotten what they learned in compulsory schools, which are very few, after all, and the idiots (sic) who have not been able to obtain the aforementioned certificate".[32]

There is support for this qualitative statement through census data, albeit only for a part of the population. As mentioned above, after the Second Schleswig War (1864), Prussia acquired in 1866 (after a brief period of shared sovereignty with Austria) the Duchies of Schleswig, Holstein, and Saxe-Lauenburg, formerly under the sovereignty of the King of Denmark. Thereby the total population of the Danish monarchy decreased by 38.5%. The data of Schleswig-Holstein appeared in the Prussian census of 1871, and its literacy level was among the highest in Prussia: 95.1%.

The Duchy of Holstein was German-speaking and part of the Holy Roman Empire, although since 1773 its sovereign was the King of Denmark. The Duchy of Schleswig had been under the sole sovereignty of the King of Denmark since 1713, and contained German-speaking and Danish-speaking areas. Danish-language speakers were concentrated in northern Schleswig, in the districts of Hadersleben, Sonderburg and Apenrade, and in part of the district of Tondern.[33] All these areas showed high literacy figures in the Prussian census of 1871: Hadersleben, 94.0; Sonderburg, 97.0; Apenrade, 96.4; Tondern, 96.1.

[30] Gersfeld was a "*Bezirksamt*", the Bavarian equivalent of a Prussian "*Kreis*", and Orb was part of the *Bezirksamt* of Gemünden.

[31] The part of Hesse-Darmstadt north of the river Main was forced into the North German Confederation from the start.

[32] "Il direttore della statistica della Danimarca ci ha scritto che, essendovi colà l'obbligo della istruzione pei fanciulli da 7 a 14 anni, a tale ultima età ognuno deve produrre il certificato di saper leggere e scrivere. Ciò però non impedisce che vi siano degli analfabeti, poichè vi sono quelli che hanno dimenticato ciò che appresero nelle scuole obbligatorie, i quali del resto sono pochissimi, e gli idioti i quali non hanno potuto procurarsi il suddetto certificato". (Italian census of 1881, *Relazione generale*, p. 136).

[33] In the 1890 Prussian census the proportion of Danish speakers was between 80% and 90% in the first three districts and around 50% in Tondern.

In contrast to the Scandinavian countries, Denmark had established a school network broadly covering the country by the middle of the nineteenth century.[34] The laws of 1814 prescribed compulsory schooling for seven years, including reading, writing, and arithmetic (see Larsen (2017)). In this sense, a statistical study of soldiers in 1859 indicated that 88.3% of them could read and write.[35] All in all, it is reasonable to assign Denmark to Group I.

5 Literacy from insufficient censual data

In the last two remaining subsections, on Great Britain and Sweden, censual data are inadequate, and additional data must be assessed taking censual data as a yardstick.

5.1 Great Britain

It is remarkable that, in contrast with Ireland, there are no censual literacy data for Great Britain (England and Wales, and Scotland)[36]. Apart from the indirect information that can be obtained from Irish censuses, in the case of Great Britain we only have literacy data (mainly about signatures in documents) typical of the age prior to the development of modern official statistics.

The Irish 1871 census provides some hints on the literacy situation in contemporary Great Britain, especially if attention is paid to the Protestant minority[37] (which interacted significantly with England or Scotland). Almost all Protestants were either Episcopalians (12.34% of the population) or Presbyterians (9.19%). Presbyterians were in general of Scottish descent and aware of their roots; 96% of them lived in the province of Ulster, with Scotland just across the North Channel. Their literacy rates were 73.9% for men, 63.4% for women and 68.5% overall[38]. Episcopalians constituted the established church in both England ("Church of England") and Ireland ("Church of Ireland"), but their small number in Ireland made a difference[39]. The core of the Episcopalians in Ireland exerted "the Ascendancy", i.e., the domination of the economy and the social and political life of the country[40]. The literacy rates for Irish Episcopalians were 72.8% for men, 65.3% for women and 69.0% overall. We may consider 69.0% as a tentative upper bound for the literacy rate of England.

[34] "The main reason for the Danish decision to introduce elementary education through a compulsory school system may have been the close cultural relations to Germany." (Tveit (1991)).

[35] See Markussen (1990). The Danish constitution of 1849 stated that all men had to report for military service.

[36] Following the merger of the parliaments of Scotland (1707) and Ireland (1801) with the Parliament of Westminster, the British Isles were under the authority of a single parliament. This situation was maintained until the Government of Ireland Act (1920) and the Anglo-Irish Treaty (1921), leading to the independence of Ireland (except Ulster).

[37] In 1871, 76.69% of the Irish population were Catholics, dispossessed of their land and economically, socially and politically oppressed; their literacy rates were 48.8% for men, 37.9% for women and 43.2% overall.

[38] In contrast with Ireland, the Presbyterian Church was the established church in Scotland, and it had there its own school system funded by a tax charged on landed property (beginning in 1696); industrialization and a more diverse society (especially after the 1843 Disruption) made this system increasingly inadequate (see Anderson (1983)). The creation of a Scottish national system of elementary education would only take place with the Education (Scotland) Act 1872. On the other hand, in 1831 a national school system was created in Ireland which financed schools that were, in practice, denominational; in 1839 the Irish Presbyterian Church entered the system on very favourable terms.

[39] The Church of Ireland maintained its own network of schools and was the established church in Ireland until 1869 (funded by taxation on the entire population and contributions of wealthy members). From 1869 most of its schools were progressively integrated into the national system introduced in 1831. The funding of the Irish Episcopalian elementary education (before 1869) had parallels with that existing in England for Anglicans prior to the 1870 Act, although the character of small privileged group of the Church of Ireland is not exactly applicable to the Church of England.

[40] Although more than half of Episcopalians lived in Ulster (58.87%), they were more spread about Ireland than Presbyterians.

From 1839 for England and Wales and from 1855 for Scotland, aggregate statistics on whether spouses signed with their names or just made a mark on the marriage document were published. These data have the limitation (see above) of any signature-based evidence for assessing literacy. Certainly, the data come in the case of marriages from a large population group in which both sexes are equally represented, although there is a strong age bias. On the other hand, in France both census literacy data (in 1866 and 1872, and then from 1900) and marriage signatures data (from 1854) are available.

The French marriage signatures data are not far apart from those in England and Wales: for men the percentage of marriages with (proper) signature is always higher for England and Wales, but the difference is less than or equal to 4 points in all years of the period 1854-1880 (with only two exceptions), and the average difference is 3.3 percentage points; for women the percentage of marriages with signature is also always higher for England and Wales, and the difference is rather stable and not very large in the period 1854-1880 (between 7 and 9 points in all years, with only five exceptions), with an average difference of 8.1 percentage points[41]; therefore the average overall difference between England & Wales and France in the period 1854-1880 is 5.7 points. Looking at the parallelism and relative proximity of both series on marriage signatures, it suggests itself to take advantage of the relationship between census literacy data and marriage signatures data in France to estimate what the census data in England and Wales would have been like. In 1872 the percentage of spouses able to write a signature at marriage in France is 71% (77% for bridegrooms, 65% for brides), while the overall literacy rate in the census is 58.3%. The reduction coefficient is $0.82 = 58.3/71$, attributable to the age bias of marriage data and to the general fact that signature-based figures overstate literacy. Since in 1872 the percentage of spouses able to write a signature at marriage in England & Wales is 77.5% (81% for bridegrooms, 74% for brides), a crude estimate of censual literacy is 63.5% ($63.5 = 0.82 \cdot 77.5$). This figure is indeed below the aforesaid 69.0% tentative upper bound for the literacy rate of England.

As for Scotland, in 1872 the percentage of spouses able to write a signature at marriage is 84.5% (90% for bridegrooms, 79% for brides), and a parallel (now riskier) estimate of censual literacy is 69.3% ($69.3 = 0.82 \cdot 84.5$); this rate is in line with the literacy rate of Irish Presbyterians mentioned above. Considering the relative weights of the populations of England & Wales and Scotland[42], the resulting estimate of the censual literacy rate of Great Britain in 1872 is 64.2%.

All in all, we can conclude that the male literacy rates of France and Great Britain were similar in the 1870s, although the gender gap was larger in France, with the result of a higher overall literacy in Great Britain, but moderately so (around 6 percentage points in 1872). On the other hand, literacy in Great Britain at the time was much lower than in Germany (the difference might be between 22 and 24 percentage points in 1872).

The marriage signatures data for 1900 allow us to assume that literacy in both England & Wales and Scotland was slightly higher than French literacy. As the latter was then already above the 75% threshold, we may place Great Britain in Group II.

5.2 Sweden

Before the implementation of the School Act of 1842, the Swedish elementary education model was based on home instruction of the ability to read *known texts*: a set of selected religious texts, where submission to authority was emphasized (see Tveit (1991) and Nilsson and Pettersson (2008)). This

[41] See Flora et al. (1983), p. 81–83. The years 1870 and 1871 have not been considered (Franco-Prussian War). In the period 1881-1900 the differences between England & Wales and France were smaller.

[42] In the 1871 census the population of Great Britain is 26,072,284, including 3,360,018 of Scotland (12.89%).

model was within the *Weltanschauung* of "the world of the *Hustavla* "[43], in the words of Johansson (1977). Practically all Swedes could "read" in this very restricted way already by the end of the 18th century. The result was high *restricted semi-literacy* and low literacy.

The only Swedish census with literacy data was too late: in 1930 (showing less than 1% illiteracy). On the other hand, there are literacy data of conscripts only from 1875 (every five years). As a compensation for this paucity of statistical data, there is a remarkable sample of individual literacy assessments carried out by the parish pastors in the diocese of Lund, recorded from 1813 to the middle of the 1840s (see Nielsen and Svärd (1994))[44]; the resulting literacy rate (population aged 15 or over) is 10% (18.7% for men and 1.4% for women)[45]. Considering the imperfections of the Lund data, Nielsen and Svärd (1994) presents their estimate for the literacy rate as an interval, whose upper bound is 20%.

The 1842 Act imposed the creation of schools in all parishes, where the teaching of proper reading, writing and elementary arithmetic was made mandatory. The new school system was established rather fast, and literacy increased sharply among the new generations[46]. However, the older generations would not decide to die quickly just to improve the literacy rates of the country, and they would last as many years as apportioned to them. We can estimate, despite the deficiencies in data, that the low male literacy and very low female literacy before the implementation of the 1842 Act did not allow Sweden to go beyond the 75% threshold during the 1870s. In order to assess Swedish literacy in 1880, the age structure of the population is to be considered (see Statistiska Centralbyrån (1969)). In Table 4, the literacy rate for those born until 1830 is estimated through the mentioned sample in the diocese of Lund, and for those born in the intervals 1851–1860 and 1861–1865 is assessed by the literacy rates of conscripts (male by definition)[47]. The estimates for the intervals 1831–1840 and 1841–1850 are obtained by interpolation[48]. The estimate of the overall literacy rate (population aged 15 or over) in 1880 is 54.9%. If the upper bound of 20% literacy for the Lund data were applied to the first interval (20% instead of 10%), and linear interpolation for the two following intervals were also implemented, the resulting rate would be 59.4%, still well below the 75% threshold[49]. Following an

[43] "The *Hustavla* (a religious plaque which was hung on the wall), was a supplement to Luther's *Small Catechism*. It consisted of specific Bible verses arranged according to the traditional, Lutheran doctrine of a three-stage, social hierarchy – *ecclesia* (church), *politia* (state), and *oeconomia* (home or household). These selections of Scripture outlined the Christian duties and obligations which each stage in this hierarchy owed to the others") Johansson (1977)). "*Hustavla*" is the Swedish translation of the German term used by Luther ("*Haustafel*", meaning "house board").

[44] It is to be considered, on the one hand, that the Lund diocese had a much higher density of schools than the rest of the country: almost half of the permanent schools in Sweden were in Lund, where only 9% of the parishes lacked schools already in 1839 (see Westberg (2019)). On the other hand, the pastors were perhaps demanding (in order to mark a person down as literate) more than the literacy level resulting from a censual declaration.

[45] The writing ability of Swedish women before the application of the 1842 Act was very low. At any rate, Nielsen and Svärd (1994) suspects that women's writing ability was underreported in the Lund research, and guesses (based on limited school data) that the women's writing rate was one-fifth of the men's writing rate, i.e., 3.7%. Thus, the overall rate would change slightly to 11.1%.

[46] The number of teachers grew from approximately 1,500 in 1839 to 2,785 in 1847 and 3,458 in 1850 (see Westberg (2019)).

[47] See Flora et al. (1983), p. 81–82. We use the 1875 rate for those born in 1851–1860 and the average of the rates of 1880 and 1885 for the interval 1861–1865.

[48] The literacy rates 10.0 and 89.0 are assigned to the years 1825 and 1855, respectively. Then the values for the years 1835 (representing the interval 1831–1840) and 1845 (representing the interval 1841–1850) are calculated by linear interpolation.

[49] Note to what extent these estimates rely on two hypotheses: (1) the percentage of those who learned to read and write after the age of 15 (or 20 from 1851) is low; (2) the abysmal gender gap before the 1842 Act closed very fast and existed no longer in the 1851–1860 interval.

analogous procedure, the estimated literacy rate for 1910 is approximately 90%[50]. With the statistical evidence available, it seems appropriate to assign Sweden to Group II.

	Born until 1830	**Born 1831–1840**	**Born 1841–1850**	**Born 1851–1860**	**Born 1861–1865**	**All**
Percentage of the population ($\geq$15)	27.9	15.9	18.2	23.3	14.7	100.0
Literacy rate	10.0	36.3	62.7	89.0	96.5	54.9

Table 4: Cohorts and literacy in Sweden, 1880.

6 Final remarks

It is not the purpose of this article to establish an explanatory model, but to facilitate comparability of data. At any rate, any analysis must take idiosyncratic factors into account, sometimes along lines different from those of material resources and incentives, as illustrated by the top literacy cluster. Group I does not correspond to the most economically advanced countries on the eve of the second industrial revolution. After the first industrial revolution, only Great Britain and Belgium were industrialised countries. The importance of cultural factors must be considered. In the German cultural sphere, ideas and experiences, going beyond denominational divisions[51], were shared, particularly, but not only, among the élites. In this sense, religious movements such as Pietism[52] or reform Catholicism[53] favoured universal literacy. Some features of the process leading to mass literacy in the German cultural sphere are the following: (1) an early start, taking place already at the end of the eighteenth century; (2) state laws made elementary education compulsory[54] (as in Prussia[55]or

[50] The exact value depends on the literacy estimate for those born until 1830. The literacy rates of conscripts are used to assess the literacy rates of those born after 1850. Indeed, the size of the cohorts not having benefited from the full implementation of the 1842 Act had tapered off substantially by 1910: among the population aged 15 or over, 7.89% were born until 1840 and 9.61% in 1841–1850 (see Statistiska Centralbyrån (1969)).

[51] In the corresponding territory of the Holy Roman Empire of the German Nation (Germany and part of Cisleithania) there were 28,674,355 Protestants and 29,639,008 Catholics in 1880. In Denmark there were 1,958,678 Protestants and 2,985 Catholics in the same year. On the other hand, within the Protestant camp, the effect of the division between Lutheranism and Calvinism was limited. From 1817, a series of decrees by King Frederick William III disposed the unification of the Reformed and Lutheran congregations into one "united" church in Prussia. The king acted in his capacity as *summus episcopus* of all these churches, sometimes rather heavy-handedly: the Lutheran Church-Missouri Synod was formed by dissenting Lutherans emigrated to the United States. The Protestant churches of the territories annexed in 1866 were allowed to remain independent, but the king of Prussia replaced the former princes as *summus episcopus*, following the "*landesherrliche Kirchenregiment*" (i.e., the governing power of the holder of territorial power over Protestant churches), characteristic (not only) of German Protestantism.

[52] See Gawthrop and Strauss (1984). As for Denmark, see Tveit (1991).

[53] "Reform Catholicism ... occupied a position within Austrian Catholicism closely analogous to that of the Pietist movement in the Lutheran Church ... In their advocacy of lay Bible reading, both prepared the ground for the promotion of compulsory schooling in their respective states." (Melton (1988)). Certainly, reform Catholicism was not particularly Austrian, and its main reference was Pope Benedict XIV.

[54] "... the German cultural sphere, of which Austria was a part, relied on state force in education from the start whereas other West-European countries made schooling compulsory by and large only at later stages. " Cvrček (2020)).

[55] The Prussian *Generallandschulreglements* for Protestant schools (1763) and for Catholic schools (1765) were advanced for their time, but poorly enforced. Mass literacy was reached in Prussia under the *Allgemeines Landrecht* (1794), which established compulsory education: "Every inhabitant who is unable or unwilling to provide the necessary education for his children in his home is obliged to send them to school after they have completed their fifth year." ("Jeder Einwohner, welcher den nöthigen Unterricht für seine Kinder in seinem Hause nicht besorgen kann oder will, ist schuldig, dieselben nach zurückgelegtem fünften Jahre zur Schule zu schicken".)

Austria[56]); (3) the control of the schools was to a large extent in the hands of the churches (with the supervision of the state), whose personal (and, to some degree, financial) resources were used; (4) elementary education was fostered equally for boys and girls from the beginning (and consequently a low gender gap resulted).

Lagging behind Group I, laws establishing national elementary education systems were passed in the other countries, as the Falloux Law (1850) of France[57], the Moyano Law (1857) of Spain and the Casati Law of Italy (1859). In these three laws, municipalities were obliged to establish elementary schools[58]. The greatest delay was in the case of Great Britain, where economic leadership did not translate into leadership in literacy, and a public network was only created in 1870 in England & Wales and in 1872 in Scotland. As for compulsory education, it was prescribed (at least in theory) by the Moyano and Casati laws, and later in 1872 in Scotland, in 1880 in England & Wales and in 1882 in France.

Acknowledgements

The author is grateful to Gloria Quiroga (Universidad Complutense de Madrid) for her help with Figure 1.

Appendix

Sources for Table 1, Table 2, and Table 3

- **Ireland:** Censuses of 1841, 1851, 1861, 1871, 1881, 1891, 1901 and 1911.
- **Spain:** Censuses of 1860, 1877, 1887, 1900, 1910, 1920, 1930 and 1940, Gutiérrez and Quiroga (2024).
- **Italy:** Censuses of 1861 and 1881, United Nations Educational, Scientific and Cultural Organization (1953).
- **Belgium:** Censuses of 1866, 1880, 1890 and 1900, United Nations Educational, Scientific and Cultural Organization (1953).
- **France:** Censuses of 1866 and 1872, United Nations Educational, Scientific and Cultural Organization (1953).
- **Prussia:** Census of 1871.
- **Austria (Cisleithania):** Censuses of 1880, 1890, 1900 and 1910.
- **Finland:** Myllyntaus (1990).
- **Portugal:** United Nations Educational, Scientific and Cultural Organization (1953).

[56] The *Allgemeine Schulordnung* (1774) of Empress Maria Theresa (drawn up by the Silesian abbot Johann Ignaz von Felbiger) established compulsory elementary education (there was a different regulation for the lands of the Archiregnum Hungaricum, the *Ratio educationis* (1777), where this compulsory character was watered down). Mass literacy was reached in Austria (proper) and Czechia under the *Allgemeine Schulordnung* and its modified version, the *Politische Schulverfassung* (1805). Beyond that, the new Cisleithanian parliament created after the *Ausgleich* of 1867 passed in 1869 the elementary education law that was to be in force until the end of the Habsburg monarchy.

[57] The earlier Guizot Law (1833) established a public network of elementary schools, but only for boys.

[58] In Spain, the *desamortización* of 1855 had confiscated and then privatised most of the land belonging to the municipalities. The funding of public primary education was assigned to these impoverished municipalities, in a context where the assets of educational charities (mostly Catholic) had been previously confiscated.

Notes to Table 1, Table 2, and Table 3

In the data of Spain (1860, 1877, 1887, 1900, 1910, 1920 and 1930) and Prussia (1871), individuals who do not state their level of literacy are considered illiterate. These are the only countries where the number of persons unspecified for literacy is provided in all the censuses before the Second World War.

The total literacy rate for the population aged 9 or over in Ireland is 87.6 in 1911.

The adjustment of the data of the 1860 and 1877 Spanish censuses to the population aged 10 or over has been estimated in Gutiérrez and Quiroga (2024).

The figures from the 1861 census in Italy are for the population aged 12 or over, and those from the 1881 census for the population aged 10 or over. In the censuses of 1871, 1901, 1911, 1921 and 1931 there are only data on semi-literacy (in 1891 no census was carried out). The Italian data for 1861 do not include Lazio and Veneto, but from the 1871 census results it can be supposed that the inclusion of these two regions would do little to alter the global data of 1861.

The data from the 1866 and 1872 censuses in France are for the population aged 6 or over, and those from the 1901 and 1911 censuses for the population aged 10 or over.

The 1880 census in Austria (Cisleithania) provided literacy data without considering ages. As the literate population under the age of six is small, and the population aged 6 or over is known, the raw literacy data obtained dividing the literate population by the population aged 6 or over have been taken here to approximate literacy rates for the population aged 6 or over, as was done retrospectively in the "Introduction" of the 1890 census (Heft 1, pp. XXI–XXVI; there is a minor mistake in the calculation of the female literacy rate of Styria). The figures from the 1890, 1900 and 1910 censuses are for the population aged 11 or over.

The global literacy rates for Cisleithania in 1880 are 61.9% for men, 55.1% for women and 58.4% overall; in 1890 they are 69.4% for men, 63.0% for women and 66.1% overall; in 1900 they are 76.6% for men, 70.7% for women and 73.6% overall; in 1910 they are 83.9% for men, 78.8% for women and 81.3% overall.

Data from the 1880 and 1890 censuses in Finland are for the population aged 10 or over, and those from the 1900–1930 censuses for the population aged 15 or over.

Appendix: Censuses and Other Statistical Sources

Censuses of Ireland

Report of the Commissioners Appointed to Take the Census of Ireland, for the Year 1841. Her Majesty's Stationery Office. Dublin (1843). (See p. 438–439).

The Census of Ireland for the Year 1851. Part IV. Report on Ages and Education. Her Majesty's Stationery Office. Dublin (1855). (See p. 184–185).

The Census of Ireland for the Year 1861. Part II. Report and Tables on Ages and Education. Vol. I and Vol. II. Her Majesty's Stationery Office. Dublin (1863). (See p. 27 Vol. I, p. 984–985 Vol. II).

Census of Ireland, 1871. Part I. Area, Houses and Population: Also the Ages, Civil Condition, Occupations, Birthplaces, Religion, and Education of the People. Summary Tables for Ireland. Her Majesty's Stationery Office. Dublin (1875). (See p. 83–84).

Census of Ireland, 1871. Part III. General Report, with Illustrative Maps and Diagrams, Summary Tables, and Appendix. Her Majesty's Stationery Office. Dublin (1876). (See p. 432–433).

Census of Ireland, 1881. Part II. General Report, with Illustrative Maps and Diagrams, Tables, and Appendix. Her Majesty's Stationery Office. Dublin (1882). (See p. 140–141, 236–237, 383).

Census of Ireland, 1891. Part II. General Report, with Illustrative Maps and Diagrams, Tables, and Appendix. Her Majesty's Stationery Office. Dublin (1892). (See p. 144–145, 350–351, 533).

Census of Ireland, 1901. Part II. General Report, with Illustrative Maps and Diagrams, Tables, and Appendix. His Majesty's Stationery Office. Dublin (1902). (See p. 146–147, 392–393, 524).

Census of Ireland, 1911. General Report, with Tables and Appendix. His Majesty's Stationery Office. London (1913). (See p. 42–43, 44, 98–99, 238).

Censuses of Spain

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

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

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

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

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

Dirección General de Estadística: *Censo de la población de España según el empadronamiento hecho en la península e islas adyacentes el 31 de diciembre de 1920.* Imprenta de los hijos de M. G. Hernández. Madrid (Tomo I, 1922; Tomo II, 1924; Tomo III, 1926; Tomo IV, 1928; Tomo V, 1929; Tomo VI, 1929).

Dirección General del Instituto Geográfico, Catastral y de Estadística: *Censo de la población de España según el empadronamiento hecho en la península e islas adyacentes y posesiones del norte y costa* *occidental de África el 31 de diciembre de 1930.* Augusto Boué Alarcón. Madrid (Tomo III, Cuadernos I-XIII, 1935-1943).

Dirección General de Estadística: *Censo de la población de España según el empadronamiento hecho en la península e islas adyacentes y posesiones del norte y costa occidental de África el 31 de diciembre de 1940.* Barranco. Madrid (Tomos I-IV, 1943-1945).

Censuses of Italy

Statistica del Regno d'Italia: *Popolazione. Censimento generale (31 dicembre 1861) per cura del Ministro d'Agricoltura Industria e Commercio. Volume Secondo.* Tipografia Letteraria. Torino (1865). (See p. XXIII).

Ufficio Centrale di Statistica: *Popolazione classificata per età, sesso, stato civile ed istruzione elementare. Censimento 31 dicembre 1871. Volume II.* Tipografia Cenniniana. Roma (1875). (See p. 37–44 and p. X of "Introduzione").

Direzione Generale della Statistica: *Censimento della popolazione del Regno d'Italia al 31 dicembre 1881. Volume II.* Tipografia Bodoniana. Roma (1883). (See p. 584).

Direzione Generale della Statistica: *Censimento della popolazione del Regno d'Italia al 31 dicembre 1881. Relazione generale e confronti internazionali.* Tipografia Eredi Botta. Roma (1885). (See p. 136).

Istituto Centrale di Statistica del Regno d'Italia: *VII Censimento generale della populazione. Volume IV: Relazione generale.* Tipografia I. Failli. Roma (1935). (See p. *95).

Censuses of Belgium

Statistique de la Belgique: Population. *Recensement générale. (31 décembre 1866). Publié par le Ministre de l'Intérieur.* Bruxelles (1870). (See p. 300-301).

Statistique de la Belgique: *Population. Recensement général. (31 décembre 1880.) Publié par le Ministre de l'Intérieur.* Bruxelles (1884). (See p. 910-911).

Statistique de la Belgique: *Population. Recensement général du 31 décembre 1890 publié par le Ministre de l'Intérieur et de l'Instruction Publique. Tome II.* Imprimerie A. Lesigne. Bruxelles (1893). (See p. 288-289).

Statistique de la Belgique: *Population. Recensement général du 31 décembre 1900 publié par le Ministre de l'Intérieur et de l'Instruction Publique. Tome II.* Typographie-Lithographie A. Lesigne. Bruxelles (1903). (See p. 358-359).

Censuses of France

Statistique de la France: *Résultats généraux du dénombrement de 1866. Statistique de la France, Deuxième série, Tome XVII.* Imprimerie Administrative de Veuve Berger-Levrault. Strasbourg (1869). (See p.

XXVIII-XXIX, 6).

Statistique de la France: *Résultats généraux du dénombrement de 1872. Statistique de la France, Deuxième série, Tome XXI.* Imprimerie Nationale. Paris (1873). (See p. XXVI-XXVII, 7).

Censuses of Prussia

Königliches Statistisches Bureau in Berlin: *Die Ergebnisse der Volkszählung und Volksbeschreibung im Preussischen Staate vom 1. Dezember 1871. Preussische Statistik XXX.* Verlag des Königlichen und Statistischen Bureaus. Berlin (1875). (See p. 6-7, 113-125).

Blenck, E.: *"Die Volkszählung vom 1. Dezember 1890 in Preußen und deren endgültige Ergebnisse". Zeitschrift des Königlich Preussischen Statistischen Bureaus, 32, p. 177-264* (1892).

Censuses of Austria (Cisleithania)

K. K. Statistische Central-Commission: *Die Ergebnisse der Volkszählung und der mit derselben verbundenen Zählung der häuslichen Nutzthiere vom 31. December 1880 in den im Reichsrathe vertretenen Königreichen und Ländern. 2. Heft: Die Bevölkerung der im Reichsrathe vertretenen Königreiche und Länder nach Religion, Bildungsgrad, Umgangssprache und nach ihren Gebrechen.* Oesterreichische Statistik, I. Band, 2. Heft. Kaiserlich-Könichliche Hof- und Staatsdruckerei. Wien (1882). (See p. 40-48, 118-119).

K. K. Statistische Central-Commission: *Die Ergebnisse der Volkszählung und der mit derselben verbundenen Zählung der häuslichen Nutzthiere vom 31. December 1880 in den im Reichsrathe vertretenen Königreichen und Ländern. 4. Heft: Die Bevölkerung der im Reichsrathe vertretenen Königreiche und Länder nach Alter und Stand.* Oesterreichische Statistik, II. Band, 1. Heft. Kaiserlich-Könichliche Hof- und Staatsdruckerei. Wien (1882). (See p. 154-171, 176-193, 530-565).

K. K. Statistische Central-Commission: *Die Ergebnisse der Volkszählung vom 31. December 1890 in den im Reichsrathe vertretenen Königreichen und Ländern. 1. Heft: Die summarischen Ergebnisse der Volkszählung.* Oesterreichische Statistik, XXXII. Band, 1. Heft. Kaiserlich-Könichliche Hofund Staatsdruckerei. Wien (1892). (See p. XXI-XXVI).

K. K. Statistische Central-Commission: *Die Ergebnisse der Volkszählung vom 31. December 1890 in den im Reichsrathe vertretenen Königreichen und Ländern. 3. Heft: Die Bevölkerung nach Grössenkategorien der Ortschaften, Stellung zum Wohnungsinhaber, Geschlecht, Alter und Familienstand, Confession, Umgangssprache, Bildungsgrad, Gebrechen.* Oesterreichische Statistik, XXXII. Band, 3. Heft. Kaiserlich-Könichliche Hof- und Staatsdruckerei. Wien (1893). (See p. 174-185).

K. K. Statistische Zentral-Kommission: *Die Ergebnisse der Volkszählung vom 31. December 1900 in den im Reichsrathe vertretenen Königreichen und Ländern. 3. Heft: Die Alters- und Familienstandsgliederung, die Bevölkerung nach Altersklassen und der Aufenthaltsdauer innerhalb der Grössenkategorien der Ortschaften, die Umgangssprache in Verbindung mit der sozialen Gliederung der Wohnparteien, mit der Alters- und Familienstandsgliederung, mit dem Bildungsgrade nach Altersklassen, mit der Konfession.* Oesterreichische Statistik, LXIII. Band, 3. Heft. Kaiserlich-Könichliche Hof- und Staatsdruckerei.

Wien (1903). (See p. 92-132).

K. K. Statistische Zentral-Kommission: *Die Ergebnisse der Volkszählung vom 31. December 1910 in den im Reichsrathe vertretenen Königreichen und Ländern. 2. Heft des ersten Bandes: Die Bevölkerung nach der Gebürtigkeit, Religion und Umgangssprache in Verbindung mit dem Geschlechte, nach dem Bildungsgrade und Familienstande; die körperlichen Gebrechen; die soziale Gliederung der Haushaltungen.* Oesterreichische Statistik, Neue Folge, 1. Band, 2. Heft. Kaiserlich-Könichliche Hof- und Staatsdruckerei. Wien (1914). (See p. 40*-42*, 70-71).

K. K. Statistische Zentral-Kommission: *Die Ergebnisse der Volkszählung vom 31. December 1910 in den im Reichsrathe vertretenen Königreichen und Ländern. 3. Heft des ersten Bandes: Die Alters- und Familienstandsgliederung und Aufenthaltsdauer.* Oesterreichische Statistik, Neue Folge, 1. Band, 3. Heft. Kaiserlich-Könichliche Hof- und Staatsdruckerei. Wien (1914). (See p. 19*-22*).

Other statistical sources

Statistiska Centralbyrån (Statistics Sweden): *Historisk statistic för Sverige. Del 1. Befolkning. Andra upplagan. 1720-1967.* KL Beckmans Tryckerier AB. Stockholm (1969). (See p. 68).

References

Anderson, R.D. (1983). Education and the state in nineteenth-century Scotland. *The Economic History Review 36*(4), 518–534.

Baffour, B., T. King, and P. Valente (2013). The modern census: Evolution, examples and evaluation. *International Statistical Review 81*(3), 407–425.

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

Bowman, M.J. and C.A. Anderson (1963). Concerning the role of education in development. In C. Geertz (Ed.), *Old Societies and New States*, pp. 247–279. Glencoe, Ill: The Free Press. Reprinted by UNESCO in *Readings in the economics of education*, Paris (1968).

Candeias, A. (2004). Literacy, schooling and modernity in twentieth-century Portugal: What population censuses can tell us. *Paedagogica Historica 40*(4), 511–530.

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

Cvrček, T. (2020). *Schooling under control. The origins of public education in Imperial Austria 1769–1869.* Tübingen: Mohr Siebeck.

Deutscher Bundestag (2019). Staatliche Maßnahmen gegenüber der polnischen Minderheit und den Bevölkerungen in den überseeischen Gebieten des Deutschen Reichs 1871 bis 1918. Available at: https://www.bundestag.de/resource/blob/594340/1e751eb9b02b5b4ba65a928ce997dbee/WD-1-040-18-pdf-data.pdf (Accessed: 8 August 2023).

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

Engel, E. (1874). Religionsbekenntniss und Schulbildung der Bevölkerung des preußischen Staats. *Zeitschrift des Königl. Preußischen Statistischen Bureaus II. u. III. Heft, XIV. Jahrgang*, 143–152.

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

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

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

Gawthrop, R. and G. Strauss (1984). Protestantism and Literacy in Early Modern Germany. *Past & Present 104*, 31–55.

Gehrmann, R. (2012). German census-taking before 1871. Max Planck Institute for Demographic Research, wp2012-001, Mosaic Working Paper.

Gutiérrez, J.M. and G. Quiroga (2024). Gender gap and spatial disparities in the evolution of literacy in spain, 1860-1910. Borda Working Papers 2401, Available at: http://hdl.handle.net/10366/160662.

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

Kerstin, F., I. Wohnsiedler, and N. Wolf (2020). Weber Revisited: The Protestant Ethic and the Spirit of Nationalism. *Journal of Economic History 80*(3), 710–745.

Labbé, M. (2007). Institutionalizing the Statistics of Nationality in Prussia in the 19th Century (from local bureaucracy to state-level census of population). *Centaurus 49*, 289–306.

Larsen, C. (2017). A Diversity of Schools: The Danish School Acts of 1814 and the Emergence of Mass Schooling in Denmark. *Nordic Journal of Educational History 4*, 3–28.

Markussen, I. (1990). The development of writing ability in the Nordic countries in the eighteenth and nineteenth centuries. *Scandinavian Journal of History 15*, 37–63.

Melton, J. Van Horn (1988). *Absolutism and the eighteenth-century origins of compulsory schooling in Prussia and Austria*. Cambridge: Cambridge University Press.

Michel, H. (1985). Volkszählungen in Deutschland: Die Erfassung des Bevölkerungsstandes von 1816 bis 1933. *Jahrbuch für Wirtschaftsgeschichte/Economic History Yearbook 26*(2), 79–92.

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

Nielsen, A. and B. Svärd (1994). Writing Ability and Agrarian Change in Early 19th-Century Scania. *Scandinavian Journal of History 19*(3), 251–274.

Nilsson, A. and L. Pettersson (2008). The state of the people? Government policies and population movements in education and training in 19th century Swedish agriculture. In N. Vivier (Ed.), *The State and Rural Societies: Policy and Education in Europe 1750-2000*, pp. 215–230. Turnhout: Brepols.

Tveit, K. (1991). The development of popular literacy in the nordic countries: A comparative historical study. *Scandinavian Journal of Educational Research 35*(4), 241–252.

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

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

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

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

Westberg, J. (2019). Basic Schools in Each and Every Parish: The School Act of 1842 and the Rise of Mass Schooling in Sweden. In J. Westberg, L. Boser, and I. Brühwiler (Eds.), *School Acts and the Rise of Mass Schooling: Education Policy in the Long Nineteenth Century*, pp. 195–221. Cham, Switzerland: Palgrave Macmillan.

5 Acknowledgement to Reviewers

The Editors of Spanish Journal of Statistics gratefully acknowledge the assistance of the following people, who reviewed manuscripts:

José Manuel Alonso, Department of Economics, University of Cantabria, Spain

Nancy Dávila, Department of Quantitative Methods in Economics and TIDES Institute, University of Las Palmas de Gran Canaria, Spain.

Emilio Gómez-Déniz, Department of Quantitative Methods in Economics and TIDES Institute, University of Las Palmas de Gran Canaria, Spain.

General Information

The Spanish Journal of Statistics (SJS) is the official journal of the National Statistics Institute of Spain (INE). The journal replaces Estadística Española, edited and published in Spanish by the INE for more than 60 years, which has long been highly influential in the Spanish scientific community. The journal seeks papers containing original theoretical contributions of direct or potential value in applications, but the practical implications of methodological aspects are also welcome. The levels of innovation and impact are crucial in the papers published in SJS.

SJS aims to publish original sound papers on either well-established or emerging areas in the scope of the journal. The objective of papers should be to contribute to the understanding of official statistics and statistical methodology and/or to develop and improve statistical methods; any mathematical theory should be directed towards these aims. Within these parameters, the kinds of contribution considered include:

- Official Statistics.
- Theory and methods.
- Computation and simulation studies that develop an original methodology.
- Critical evaluations and new applications
- Development, evaluation, review, and validation of statistical software and algorithms.
- Reviews of methodological techniques.
- Letters to the editor.

One volume is published annually in two issues, but special issues covering up-to-date challenging topics may occasionally be published.

Author guidelines

The Spanish Journal of Statistics publishes original papers in the theory and applications of statistics. A PDF electronic version of the manuscript should be submitted to José María Sarabia, Editor in chief of SJS via email to sjs@ine.es. Submissions will only be considered in English.

Manuscripts must be original contributions which are not under consideration for publication anywhere else. Its contents have been approved by all authors and. A single-blind refereeing system is used, so the identity of the referees is not communicated to the authors. Manuscripts that exceed 30 journal pages are unlikely to be considered for publication. More detailed information can be found at https://www.ine.es/sjs.

Published by the Spanish National Statistical Institute