Uncovering Companies Missing from the *SABI* Database: A Web Scraping Approach

Xin-Hui Huang; Josep Domenech

Recommended citation:

Huang, Xin-Hui; Domenech, Josep (2025). "Uncovering Companies Missing from the SABI Database: A Web Scraping Approach". *Profesional de la información*, v. 34, n. 2, e34202.

https://doi.org/10.3145/epi.2025.ene.34202

Manuscript received on 10th January 2025
Accepted on 12th March 2025

Xin-Hui Huang
https://orcid.org/0009-0002-8167-7285
Universitat Politècnica de València
Dept. Economics and Social Sciences
46022 València (Spain)
xhuang@etsinf.upv.es

Josep Domenech ✉
https://orcid.org/0000-0002-7302-5810
Universitat Politècnica de València
Dept. Economics and Social Sciences
46022 València (Spain)
jdomenech@upvnet.upv.es

Abstract

This study evaluates the completeness and representativeness of the *SABI* database, a widely used commercial source for firm-level data in Spain and Portugal, by comparing it to *BORME*, the official Spanish business register. Using web scraping techniques, we collected and processed approximately 100,000 *BORME* publications in PDF format, covering the period from 2010 to 2023. These were transformed into a structured dataset comprising over 1.2 million companies, which we then matched against *SABI* records from the same period. Our analysis reveals that *SABI* covers only 38.3% of newly established companies, with significant underrepresentation of younger firms, small enterprises, specific sectors, and certain regions. Furthermore, we find clear evidence of survivorship bias: the longer a company has been dissolved, the less likely it is to appear in *SABI*. Sectoral and geographic disparities are also substantial, and the coverage is skewed toward firms with higher initial capital and specific legal forms. These findings suggest that *SABI* represents a non-random subset of the Spanish business population, and caution should be exercised when using it for empirical research. Adjustments for sample bias are recommended to improve the reliability of analyses based on this database.

Keywords

SABI Database, Data Quality, Web Scraping, Database Reliability, Commercial Information, Directories, Firm Information, Data Reliability, Bias.

1. Introduction

The *SABI* database, developed by *Bureau van Dijk*, is a financial database and analysis tool that provides information on companies in Spain and Portugal. It is widely used by businesses, researchers, financial analysts, and professionals to access comprehensive information about companies. The variety of research using *SABI* illustrates its broad applicability across many different topics and studies (**Martínez-Matute; Urtasun**, 2022; **Rizov *et al.***, 2022; **Sánchez-Infante Hernández *et al.***, 2020). However, despite its popularity, concerns regarding its completeness and representativeness still persist. These limitations can significantly impact the reliability and interpretation of research findings derived from *SABI* data, creating potential biases and affecting policy and business decisions.

Similar to *SABI*, other databases cover different geographic areas. *Bureau van Dijk* also offers *FAME* (UK and Ireland), *AIDA* (Italy), *DAFNE* (Germany) and *ORBIS* (Global), among others. These databases present a comprehensive reach and frequently serve as a proxy for the total firm population in research (**Garcés-Galdeano *et al.***, 2024; **Martinez-Sanchez; Lahoz-Leo**, 2018; **Opazo-Basáez *et al.***, 2024). However, these databases do not exhaustively represent the corporate landscape, as they offer limited coverage, especially for small and micro firms (**Almunia *et al.***, 2018; **Bajgar *et al.***, 2020; **Pinto Ribeiro *et al.***, 2010). Therefore, the practice of considering companies listed there as the population may overlook the fact that they constitute a non-random sample rather than a complete census. This distinction is crucial for accurately interpreting findings derived from its data, highlighting the need for awareness regarding its scope and limitations in research.

In contrast, the *BORME* is the official gazette for business registrations and updates in Spain, providing a complete legal record of new companies, modifications, and terminations. As a primary source of official business information, *BORME* plays a crucial role in maintaining transparency and up-to-date records of the business landscape in Spain. As such, *BORME* represents the full population of registered companies in Spain. Although the information available for each firm in *BORME* is limited, its comprehensive nature offers a reliable benchmark for assessing the coverage of business databases like *SABI*. By comparing these sources, it becomes apparent that while *BORME* encompasses the entire population, *SABI* includes only a non-random sample of firms. This difference has implications for the interpretation of research findings based on *SABI* data. The challenge lies in *BORME*'s format —a huge collection of web-based PDFs without tabular data— which complicates direct comparisons with *SABI*'s structured database. Overcoming this barrier requires innovative data extraction and analysis methods, emphasizing the importance of advanced technological solutions in bridging the information gap between these two resources.

Web scraping, the automated extraction of information from websites, has become a widely used method for data collection in research (**Trezza**, 2023). It enables researchers to gather large volumes of data directly from online sources, complementing or replacing traditional datasets. **Edelman** (2012) defines "scraping the Internet for data" as collecting information (e.g. prices, quantities, text) that is already available on websites but not yet organized in a form useful for analysis. In essence, web scraping leverages the vast, real-time information on the web to create custom datasets for specific research questions. This approach offers efficiency gains: instead of manual copy-pasting (which is time-consuming and error-prone), automated scrapers can retrieve data quickly and with fewer transcription errors (**Dogucu; Çetinkaya-Rundel**, 2021). As a result, web scraping offers "exceptional possibilities" for researchers by increasing the amount of information available and lowering the costs of data collection (**Edelman**, 2012). However, this method also poses challenges, including legal issues related to copyright and compliance with terms of service, data security concerns, and technical complexities that researchers must carefully consider.

Our study aims to identify and describe companies that appear in *BORME* but are not found in the *SABI* database. While the representativeness of databases like *ORBIS*, *Bloomberg SPLC*, and *Compustat* has been examined in prior research (**Bajgar *et al.***, 2020; **Liu**, 2020; **Pinto Ribeiro *et al.***, 2010), studies specifically focused on *SABI* are lacking. The research addresses this gap by systematically investigating the completeness and potential biases of the *SABI* database. We investigate the differences between *BORME* and *SABI* to reveal key characteristics of omitted companies, such as their year of establishment and year of dissolution. Our goal is to understand how these characteristics affect the completeness and reliability of the *SABI* database. By identifying potential biases or omissions, our results provide valuable insights for improving the quality of economic analyses, policymaking, and business strategy development that rely on such databases.

2. Literature Review

Concerns regarding data quality in business databases are a recurrent topic in the literature due to the implications for firm-level research, although such concerns are less frequently discussed for *SABI* database. This section reviews some studies on the prevalent issues, ranging from missing values to data errors and biases that affect the reliability and representativeness of datasets. Given that both *SABI* and *Orbis* are provided by *Bureau van Dijk*, it is reasonable to expect similar issues in both databases.

2.1. Representativeness and Selection Bias

A particular concern with representativeness arises when using widely-used databases like *Orbis*. These databases often fail to provide a nationally representative sample, which can significantly skew interpretations and policy implications. Studies, such as those by **Kalemli-Özcan *et al.*** (2024), underscore the risks of relying on a single release, or "vintage," of the Orbis database. Here, a "vintage" refers to data collected at a specific point in time, which may not accurately reflect conditions for smaller or medium-sized enterprises that have more volatile reporting practices or shorter lifespans in the database. Selection bias, including survivorship bias, is common in business databases. Small firms and those with poor performance or that do not report are more likely to be excluded from these databases. This bias can significantly distort findings, leading to overestimations of success rates, financial health, and the overall resilience of businesses.

In terms of specific databases, *Orbis* tends to exclude companies that do not report after a certain period and it under-represents smaller firms (**Bajgar *et al.***, 2020; **Gal**, 2013; **Kalemli-Özcan *et al.***, 2024; **Pinto Ribeiro *et al.***, 2010). Similar patterns of low coverage of small firms are reported in the *SABI* database (**Almunia *et al.***, 2018; **Casillas *et al.***, 2024). Additionally, smaller and underperforming firms are more likely to be excluded from *Thomson Datastream* (**Andrikopoulos *et al.***, 2007; **Ince; Porter**, 2006), further skewing research outcomes. Additionally, the regional and country-specific representation in *Orbis* is inconsistent (**Bajgar *et al.***, 2020; **Gal**, 2013). These discrepancies can lead to significant biases in understanding regional economic dynamics, particularly when extrapolating findings from the dataset

to broader economic contexts. Similarly, sectoral representation issues are pronounced (**Bajgar *et al.***, 2020), with the services sector frequently underrepresented (**Gal**, 2013). This underrepresentation can lead to incomplete or skewed analyses, especially in economies where services play a significant role.

The representation of younger firms is another critical area where *Orbis* data shows significant gaps. **Bajgar *et al.*** (2020) and **Gal** (2013) both note that younger firms are underrepresented, which is problematic since these firms often drive innovation and economic dynamism. This, together with the temporal variability in coverage, complicate longitudinal analyses necessary for assessing the entry and exit dynamics of firms (**Kalemli-Özcan *et al.***, 2024). Such variability can obscure the true nature of economic changes over time, affecting the reliability of research findings that aim to inform policy and business strategy. Even within their size and region class, more productive firms are disproportionately represented (**Bajgar *et al.***, 2020; **Gal**, 2013), leading to an insufficient variability that can distort analyses of productivity and economic performance across the firm spectrum.

Missing values are one of the most prevalent data quality problems among companies present in the database. Researchers may use special procedures or filters to exclude companies with missing values from the sample. However, missing values are not random, and this practice may inevitably create omission bias or selection biases (**Elton *et al.***, 2001; **Liu**, 2020; **Weiß; Mühlnickel**, 2014). Dropping all observations that contain missing values is a naïve strategy and can noticeably affect on the statistical power of the tests (**Hribar**, 2016), potentially leading to misleading results. **Kalemli-Özcan *et al.*** (2024) note that missing values could also occur due to the cap on the amount of data allowed to be downloaded and highlight that firm-level data from *Orbis* are not nationally representative. A high number of missing values may render a database unusable for specific research (**Francis *et al.***, 2018).

2.2. Data Quality Issues

Data quality remains a central challenge in empirical research, particularly when utilizing large-scale datasets. Several studies have highlighted concerns regarding the accuracy and completeness of data in *Orbis*. **Bajgar *et al.*** (2020) noted that data on certain economic variables are often rounded to the nearest hundred or thousand in some countries. This rounding can obscure fine-grained variations in data, potentially leading to biased estimates and conclusions. Similarly, issues with missing data are prevalent, as observed by **Kalemli-Özcan *et al.*** (2024) and **Gal** (2013), who pointed out gaps in key economic indicators like value added.

Duplicate records are a significant issue in *Orbis* (**Kalemli-Özcan *et al.***, 2024). Such duplicates can artificially inflate the size of the dataset and distort findings from statistical and econometric analyses. Resolving duplicates requires robust data cleaning processes, which can be resource-intensive and may not always be feasible. Similarly, the problem of entity ambiguity further complicates data quality. **Pinto Ribeiro *et al.*** (2010) raised questions about whether records in *Orbis* represent individual enterprises or establishments, a distinction that can affect the interpretation of economic dynamics at different scales. Moreover, **Arndt** (2023) discussed the challenges in accurately identifying the ultimate owners in firm-level datasets. This ambiguity can lead to incorrect assumptions about control and ownership structures, affecting analyses related to corporate governance and economic influence.

2.3. Methodological Challenges

The analysis of firm-level financial data presents several methodological challenges that can significantly impact research outcomes. First, it must be taken into account that *Orbis* include both consolidated and unconsolidated financial statements. The inclusion of both types of statements necessitates careful management to avoid double counting (**Bajgar *et al.***, 2020; **Kalemli-Özcan *et al.***, 2024). Although consolidated statements aggregate the financial data of a parent company and its subsidiaries, its presentation in *Orbis* is not always consistent. Variable meanings within financial datasets can evolve due to changes in political, legal, or administrative contexts. Such changes may not only affect the comparability of data year over year but also the reliability of longitudinal studies that fail to account for these shifts in variable definitions and contexts (**Pinto Ribeiro *et al.***, 2010). The issue of static header data, or the vintage problem, arises when databases only provide the most current information available and lack historical records for time-series analysis. This issue predominantly affects data fields such as company name, address, and industry code. This has been observed across various databases including *Compustat*, *Orbis*, *Worldscope*, *Datastream*, and *DAFNE* (**Beuselinck *et al.***, 2023; **Kalemli-Özcan *et al.***, 2024; **Liu**, 2020).

3. Methodology

This section describes the data sources for companies established between 2010 and 2023 and the procedures employed to build the dataset and analyze it. Specifically, we explain how the information from the *BORME* and *SABI* databases was processed and merged to assess the coverage and representativeness of *SABI*. We also detail the statistical methods applied to compare distributions between both data sources.

3.1. Data from *SABI*

SABI is a commercial database that is regularly updated, including its coverage. As of January 9, 2024 (version 145.00, update 293), it contains 1,911,775 companies in Spain[1]. Among these, 502,123 firms were established between 2010 and 2023. *SABI*'s data comes already in a tabular format, containing most relevant variables like company name, address, incorporation date, sector classification, and financial information. However, as other research points out, it is not exhaustive, and some firms do not appear in the database.

3.2. Data from BORME: Crawling and Scraping Process

BORME is the official gazette in which business-related events (e.g., incorporations, capital changes, dissolutions) are published daily in PDF format. Figure 1 provides an overview of the entire data pipeline, from crawling to final dataset creation.

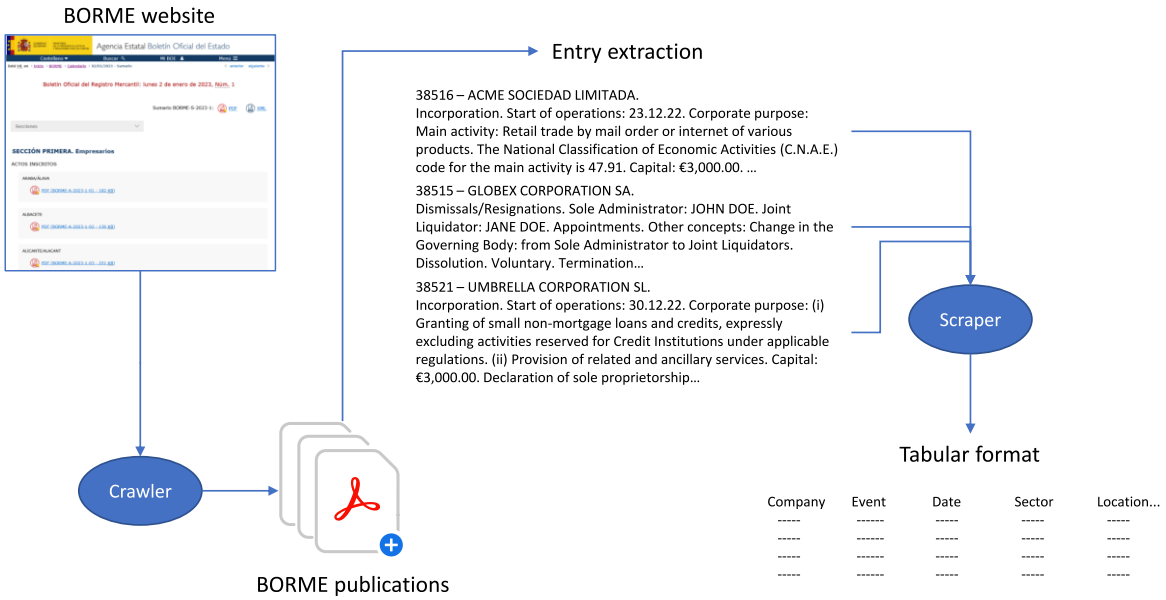


Figure 1: Data Pipeline for Extracting Company Information in Tabular Format after Crawling the *BORME* Website and Scraping its Publications.

Crawling: To collect the publications on *BORME* website, a common *Python*-based crawler was used. The crawler systematically iterated over the official website's URL structure, retrieving approximately 100,000 PDF documents corresponding to all daily publications within the 2010–2023 period. The approach included: i) identifying all valid links for *BORME* daily issues within the specified timeframe; ii) downloading each PDF file and storing it locally for further processing; and iii) recording minimal metadata for each publication (date, issue number, etc.) to enable organized data management and potential re-downloading of specific files if needed.

Scraping and text processing: Once downloaded, each PDF was converted to raw text. The goal was to transform each *BORME* publication from a single unstructured text file into a structured set of entries, where each entry corresponds to one company-related registry event, such as company establishments, capital increases, or bankruptcies. Although the entries follow a similar structure, there are slight variations among the 60 registries in Spain. To reliably segment the text, an initial analysis was performed to identify the 5,000 most frequent n-grams across the extracted text. These n-grams were manually reviewed and classified according to whether they served as keywords marking the start of a field in the *BORME* entries[2]. The resulting set of keywords was then used to divide each registry entry into distinct fields, thus converting the raw text into a structured, tabular format.

Finally, the different entries related to the same company were grouped. Firms may appear in multiple entries over time, especially in cases of capital change, board reshuffling, relocations, or name changes. Hence, all entries indicating a new name were managed to keep track of the historical changes in company identity. Since each *BORME* publication only identifies companies by name (no unique ID is provided), this step was important to ensure that events corresponding to the same firm were consistently grouped.

[1] The list of companies was downloaded after selecting "All companies" as the search strategy. Notice that the number of companies found using this procedure is significantly lower than the number shown in their commercial information page, where they claim to have information on 2.9 million companies.

[2] Some examples of these keywords are: *Constitución* (Incorporation), *Disolución* (Dissolution), *Objeto social* (Corporate purpose), *Actividad principal* (Main activity).

To reduce inconsistencies, normatively accepted abbreviations for legal forms were standardized (e.g., *Sociedad Limitada*, *S.L.*, and *SL* were all mapped to a unified label *SL*). These standardized company names also facilitated the matching with *SABI*, where similar variations in names can appear. Following these procedures, around 9,956,791 registry entries (relating to 3,051,505 companies) were processed. As a final filter, only companies established between 2010 and 2023 were retained, yielding 2,917,784 entries belonging to 1,298,056 companies.

To analyze coverage and identify which companies from *BORME* are missing in *SABI*, the cleansed *BORME* dataset was merged with the *SABI* list of companies. As *BORME* does not provide a numerical company ID, matching was carried out using the standardized company names. This procedure is feasible and reliable because, under Spanish law, company names must be unique. That is, it is not legally possible for two distinct companies to be registered with the same exact name. This legal constraint ensures that name-based matching between *BORME* and *SABI* is unambiguous and can be applied without the risk of duplicity. Ultimately, the 1,298,056 companies from *BORME* were successfully compared with 502,123 companies in *SABI*. This merged dataset forms the basis for the subsequent analysis, which explores differences in coverage across multiple dimensions, including year of establishment, legal form, geographic location, sector, and dissolution rates.

3.3. Statistical Methods

To assess whether *SABI* provides a representative sample of companies relative to *BORME*, we compare the distribution of firms along several dimensions. For categorical variables such as the legal form, sector, and province of incorporation, we employ the chi-squared test to determine whether their distributions in *SABI* differ significantly from those in *BORME*. For the continuous variables, such as the incorporation date and the initial capital, we use the Kolmogorov–Smirnov (K-S) test to compare its distribution across the two sources.

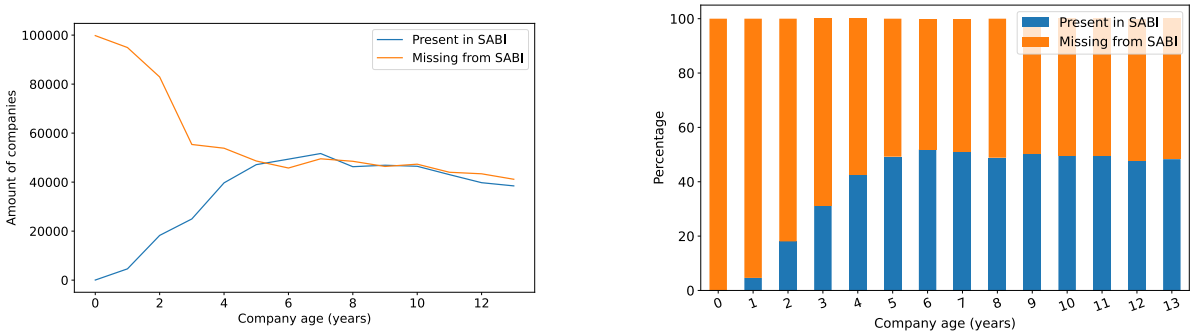
4. Results

The following subsections examine in detail how the coverage of the *SABI* database varies according to different company characteristics. Each dimension provides further insight into the potential biases present in the data, supporting a more precise assessment of its representativeness for empirical research.

4.1. Coverage by Year of Establishment

Overall, *SABI* offers data about approximately 38.3% of the companies established between 2010 and 2023. However, this coverage is not independent of the year in which the companies were established. Figure 2a compares the amount of companies present and missing from *SABI* across different age groups. As one can observe, the vast majority of firms established in the last few years are not listed in *SABI*. For companies established 5 or more years ago, the amount of covered companies is similar to those missing. This can also be observed in Figure 2b, where the coverage of *SABI*, expressed as a percentage, is represented. This stacked bar chart shows that nearly 100% of companies aged 0-1 years are missing from the database, with this percentage progressively decreasing as company age increases. By the age of 5 years, around half of the companies are present in the *SABI* database, a proportion that remains stable for older companies. The integration of these two figures highlights a clear trend: younger companies are predominantly absent from *SABI*, while the database's completeness improves significantly with the increasing age of companies, particularly beyond the 5-year mark. A one-sample K-S test confirms that the age distribution of companies in *SABI* differs significantly from that of all registered companies in *BORME* ($D = 0.2165$, p-value < 0.001), reinforcing the evidence that *SABI* disproportionately excludes younger firms.

This highlights the limited coverage and representativeness of samples drawn from *SABI*, particularly for recently created companies. Researchers and analysts should be aware of this bias when using *SABI* data, as it may impact studies involving younger firms or those seeking to understand the dynamics of newly established businesses.



a. Amount of companies by age

b. SABI coverage by the company age in years

Figure 2: The Proportion and Amount of Companies present in SABI, Overall and by their Age.

4.2. Dissolved Firms

To study the survivorship bias on the coverage of *SABI*, the duration for which data on dissolved companies are retained is examined. Figure 3 shows that, among the companies dissolved in the last four years (2020-2023), the proportion of companies present in *SABI* is around 40%, similar to *SABI*'s overall coverage. However, for companies dissolved more than four years ago, the proportion included in the database decreases by 3.7 percentage points per year. This trend evidences a diminishing likelihood of dissolved companies remaining in the database as time progresses, underscoring the influence of survivorship bias on the observed stability and longevity of businesses within *SABI*. The data point for the year 2010 does not follow this trend, likely due to the low number of companies that were dissolved that year. It is worth noting that the dataset only includes companies created from 2010 onwards; therefore, this particular bar represents the 125 companies that were both established and dissolved in 2010, with only 45 of them being present in *SABI*. A one-sample K-S test (D = 0.2063, p-value < 0.001) confirms that the distribution of ages for dissolved companies in *SABI* differs significantly from that in the broader population of firms, reinforcing the presence of survivorship bias.

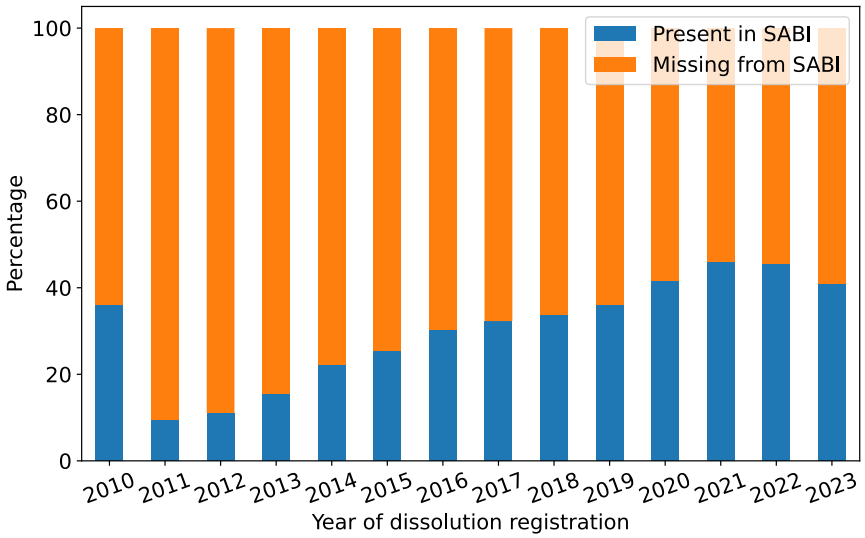


Figure 3: Distribution of Dissolved Firms' Presence in SABI by Year of Dissolution.

4.3. Geographical Coverage

The data has been analyzed to assess potential geographic bias within Spain. Figure 4 depicts the coverage by province of establishment, revealing significant variation across regions. Coverage ranges from 65% in Lugo (northwest) to 18% in Guadalajara (center). Despite this disparity, there is no apparent relationship between coverage and the characteristics of each province. Provinces at both extremes of the coverage spectrum have relatively low populations, while the most populated provinces (Madrid, Barcelona, and Valencia) occupy a relatively central position in terms of coverage. Furthermore, no clear differences are observed between coastal and non-coastal provinces. A Pearson's chi-squared test confirms that the distribution of companies across provinces in *SABI* differs significantly from that in *BORME* (x^2= 34,570, df = 51, p-value < 0.001), indicating that geographic coverage is not uniform. However, the lack of a clear pattern suggests that the variation is not systematically linked to regional characteristics.

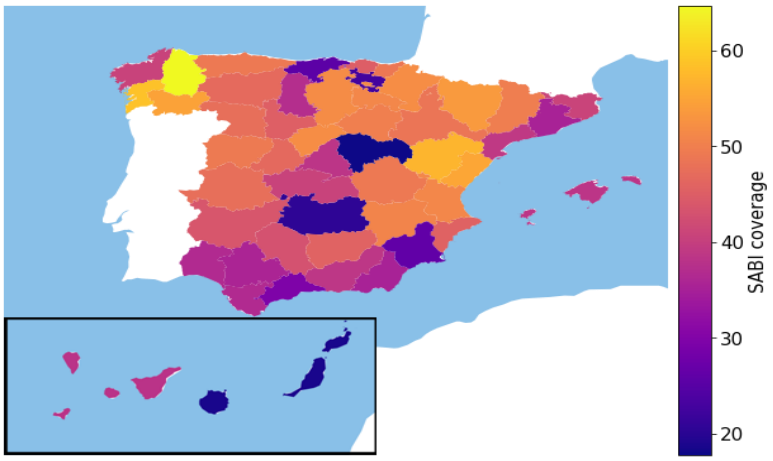


Figure 4: Proportion of companies present in *SABI* by province of establishment.

4.4. Sectoral Coverage

The sectoral coverage of companies in *SABI* is illustrated in Figure 5. The graph categorizes companies according to the NACE (Nomenclature of Economic Activities) classification, displaying the proportion of firms present in and missing from the *SABI* database. To perform this analysis, only those companies whose registry entries include their activity code have been considered.

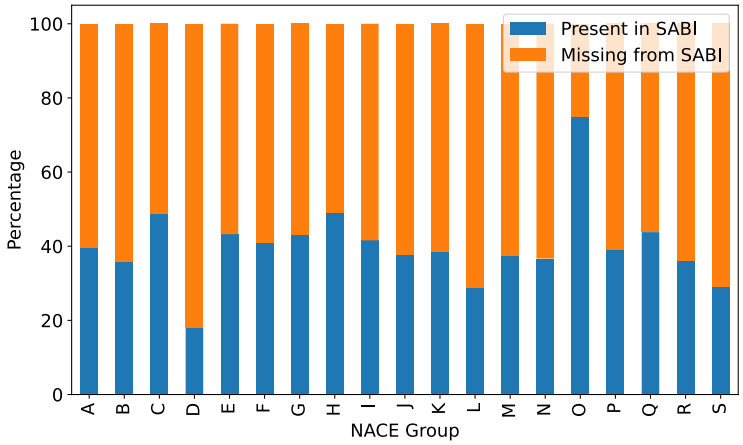


Figure 5: Proportion of Companies that are Present in *SABI* and those that are Missing in *SABI* for each NACE Group.

Figure 5 evidences that there is significant variation in the coverage across different sectors. Notably, Section O (Public administration and defense) stands out with the widest coverage, with over 70% of companies in this sector being present in the database. This is followed by Sections C (Manufacturing) and H (Transportation and storage), which have around 50% coverage each. Conversely, certain groups show significantly lower inclusion rates. Section D (Electricity, gas, steam, and air conditioning supply) exhibits the least coverage, with less than 20% of companies present in the database. Following this, Sections L (Real estate activities) and S (Other service activities) also have low coverage, with only around 30% of companies in these sectors being included in *SABI*. A Pearson's chi-squared test confirms that the distribution of sectoral coverage in *SABI* deviates significantly from that of the overall firm population (x^2= 144.79, df = 18, p-value < 0.001), reinforcing the presence of selection bias across industries. This bias can impact the representativeness of analyses based on *SABI* data, as certain sectors are systematically underrepresented.

4.5. Coverage by Legal form

Unlike *SABI*, *BORME* does not collect information on traditional ways of measuring company size, such as the number of employees or the revenue. Hence, it is not possible to know the actual size of companies not present in *SABI*. However, *BORME* includes some information that is correlated with the company size. One of these is the company legal form, which is mandatory in the company names in all registries. There are two main forms of company legal forms in Spain. *Sociedad Limitada* (*SL*) is the legal form used for limited liability companies, which is the predominant form used by generally smaller companies than *Sociedad Anónima* (*SA*), which is used for corporations or public limited companies.

Table 1 provides a comparison of the coverage of different legal forms in the *SABI* database. The table shows the absolute numbers and percentages of companies present and missing from *SABI*, categorized by their legal form.

Table 1: Coverage of Different Legal Forms in *SABI*.

Legal form	In *SABI*	Not in *SABI*	%In *SABI*	%Not in *SABI*
SL	493,774	831,318	37.26%	62.74%
SA	3,043	2,987	50.46%	49.54%
Other	2,951	8,941	24.82%	75.18%

The coverage of companies varies significantly by legal form. *SL*s, which represents the majority of companies, has a relatively low coverage, with only 37.26% of these companies present in *SABI*. SAs has a higher inclusion rate (50.46%), although far from a complete coverage. The "Other" category, which includes less common legal forms, has the lowest coverage at 24.82%. A chi-squared test (x^2= 1236.97, df = 2, p < 0.001) confirms that *SABI* is not a random sample of *BORME* and suggests a bias toward larger companies, as indicated by the higher representation of SAs.

4.6. Initial Capital

The initial capital of a company is also related to its size and can provide insights into its economic significance and potential scalability. Focusing on *SL* companies, Figure 6 illustrates the initial capital of companies over the years, categorized by their presence in the *SABI* database.

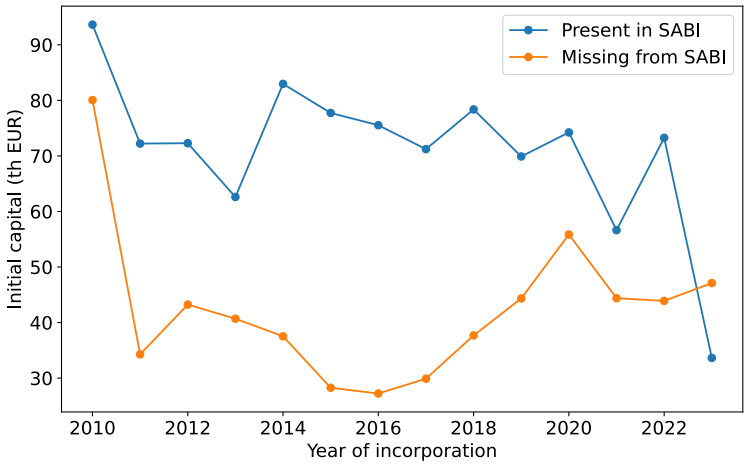


Figure 6: Average Initial Capital of SL Companies by Year of Incorporation, Categorized by their Presence in the *SABI* Database.

SL companies present in *SABI* generally have, on average, higher initial capital compared to those missing from the database. Over time, the initial capital for companies present in *SABI* shows a fluctuating but generally higher trend compared to those not in *SABI*. Although the variability across years is high, the disparity in initial capital between the two groups suggests that *SABI*'s coverage is biased towards larger companies with higher initial investments, even within the group of limited liability companies. This is further confirmed by a one-sample K-S test, which shows a statistically significant difference in the distribution of initial capital between *SABI* and the full *BORME* dataset (D = 0.51134, p-value < 0.001), reinforcing the conclusion that *SABI* disproportionately includes firms with higher initial capital. Hence, the generalizability of results based on samples drawn from *SABI* is affected.

5. Conclusions

This study investigated the limitations of the *SABI* database in capturing the Spanish business landscape by comparing it to the *BORME* official gazette. Our web scraping approach enabled the identification and analysis of companies that are listed in *BORME* but absent in *SABI*. The findings offer significant insights into the representativeness and reliability of the *SABI* database. Our research highlights that *SABI* covers approximately 38.3% of companies established between 2010 and 2023, with a notably declining representation of newly established firms. This finding is consistent with earlier studies that documented the underrepresentation of younger firms in *Orbis* (**Bajgar *et al.***, 2020; **Gal**, 2013), suggesting that such bias also extends to *SABI*. This trend poses a challenge for researchers relying on *SABI* for studying emerging business trends and the dynamics of new firm establishments. Additionally, the geographic variability in *SABI*'s coverage is pronounced, ranging significantly across different provinces in Spain. This heterogeneity indicates that *SABI*'s data is not uniformly representative of the entire country, which can lead to regional biases, underscoring the need for regional adjustments when utilizing *SABI* for economic research. Our analysis also revealed a significant survivorship bias. Companies dissolved more than four years ago are progressively less likely to be found in *SABI*, with the likelihood decreasing by 3.7 percentage points per year. This pattern aligns with prior findings on survivorship and selection bias in *Orbis* and other financial databases (**Kalemli-Özcan *et al.***, 2024), reinforcing concerns about overestimating firm longevity and success.

Furthermore, sectoral coverage analysis shows substantial variation. This bias affects the comprehensiveness of economic analyses derived from *SABI*, potentially skewing insights and policy recommendations for specific industries. In terms of company size, we found that the *SABI* database tends to favor larger entities, such as public limited companies (*SA*s), over smaller limited liability companies (*SL*s) and other legal forms. This bias towards larger firms is also reflected in the higher initial capital of companies present in *SABI* compared to those absent. Such biases suggest that *SABI* data may not fully capture the diversity of business sizes and types, particularly underrepresenting smaller enterprises with lower initial capital. The limited coverage of the *SABI* database can be attributed to several factors. One reason could be the non-fulfillment of obligations by companies to deposit their financial statements in the registry. This non-compliance with legal requirements results in many companies remaining unrecorded in *SABI*, which relies on these submissions for database updates. Another contributing factor is the potential inefficiency or gaps in *SABI*'s data collection processes. Inconsistencies in how *SABI* retrieves, processes, and updates data from the registry may lead to incomplete records. Additionally, *SABI*'s unclear policy for removing companies from its database could also affect coverage issues. If the criteria and procedures for company removal are not consistently applied, it could result in the inadvertent exclusion of active companies or the retention of outdated information.

In addition to these points, the non-random nature of the *SABI* dataset should be emphasized. The selection biases identified throughout this study indicate that *SABI* essentially represents a non-random sample of the full population

of registered companies. Consequently, statistical inferences based on *SABI* data must be approached with caution. Methodologies specifically designed for non-random samples, such as selection models, inverse probability weighting, and propensity score adjustments, can be applied to adjust for these biases (**Golini; Righi**, 2024; **Wu**, 2022). Future research should consider employing such techniques to correct for the selection effects inherent in *SABI*, thereby enhancing the reliability of empirical findings derived from its data. Several limitations must be considered when interpreting the results of this study. First, the web scraping techniques used, while robust, are not infallible and may have missed some companies due to technical constraints or errors in data extraction. Additionally, the analysis focuses on the period between 2010 and 2023, potentially missing longer-term trends and changes in database coverage. In conclusion, this study highlights significant gaps and biases in the *SABI* database. Practitioners should consider supplementary data sources or adjust their methodologies to account for the identified biases when using *SABI* for analysis.

6. Funding

This work was partially supported by Grant PID2023-152106OB-I00 funded by *MICIU/AEI/* 10.13039/501100011033 and by ERDF/EU; Grant CIAICO/2023/272 funded by *Generalitat Valenciana*; and by *Ministerio de Educación y Formación Profesional*.

References

Almunia, Miguel; Lopez Rodriguez, David; Moral-Benito, Enrique. (2018). "Evaluating the macro-representativeness of a firm-level database: an application for the Spanish economy". *Banco de Espana Ocassional Paper*, No. 1802. Banco de España, Madrid. *https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/18/Files/do1802e.pdf*

Andrikopoulos, Panagiotis; Daynes, Arief; Pagas, Paraskevas; Latimer, David. (2007). "UK Market, Financial Databases and Evidence of Bias". *Occasional Paper Series Paper*, No. 79. Leicester Business School, De Montfort University, Leicester.

Arndt, Lukas. (2023). "When Should We Believe Research Using ORBIS Firm Data?". *Available at SSRN 4584106*. *https://doi.org/10.2139/ssrn.4584106*

Bajgar, Matej; Berlingieri, Giuseppe; Calligaris, Sara; Criscuolo, Chiara; Timmis, Jonathan. (2020). "Coverage and representativeness of Orbis data". *OECD Science, Technology and Industry Working Papers*, No. 2020/06. OECD Publishing, Paris. *https://doi.org/10.1787/c7bdaa03-en*

Beuselinck, Christof; Elfers, Ferdinand; Gassen, Joachim; Pierk, Jochen. (2023). "Private firm accounting: the European reporting environment, data and research perspectives". *Accounting and Business Research,* v. 53, n. 1, pp. 38-82. *https://doi.org/10.1080/00014788.2021.1982670*

Casillas, José Carlos; Escribá-Esteve, Alejandro; Gómez-Miranda, María Elena; López-Fernández, María Concepción; Lorenzo-Gómez, Daniel; Requejo, Ignacio; Rojo-Ramírez, Alfonso A. (2024). "SAFER Methodology: A Proposal for the Identification of Family Firms in Spain Based on the SABI Database". *European Journal of Family Business,* v. 14, n. 1, pp. 85-97. *https://doi.org/10.24310/ejfb.14.1.2024.18965*

Dogucu, Mine; Çetinkaya-Rundel, Mine. (2021). "Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities". *Journal of Statistics and Data Science Education,* v. 29, n. sup1, pp. S112-S122. *https://doi.org/10.1080/10691898.2020.1787116*

Edelman, Benjamin. (2012). "Using Internet Data for Economic Research". *Journal of Economic Perspectives,* v. 26, n. 2, pp. 189-206. *https://doi.org/10.1257/jep.26.2.189*

Elton, Edwin J.; Gruber, Martin J.; Blake, Christopher R. (2001). "A First Look at the Accuracy of the CRSP Mutual Fund Database and a Comparison of the CRSP and Morningstar Mutual Fund Databases". *The Journal of Finance,* v. 56, n. 6, pp. 2415-2430. *https://doi.org/10.1111/0022-1082.00410*

Francis, Rick N.; Mubako, Grace; Olsen, Lori. (2018). "Archival research considerations for CRSP data". *Accounting Research Journal,* v. 31, n. 3, pp. 360-370. *https://doi.org/10.1108/ARJ-06-2016-0065*

Gal, Peter N. (2013). "Measuring Total Factor Productivity at the Firm Level using OECD-ORBIS". *OECD Economics Department Working Papers*, No. 1049. OECD Publishing, Paris. *https://doi.org/10.1787/5k46dsb25ls6-en*

Garcés-Galdeano, Lucía; Kotlar, Josip; Caicedo-Leitón, Ana Lucía; Larraza-Kintana, Martín; Frattini, Federico. (2024). "Absorptive capacity in family firms: exploring the role of the CEO". *International Journal of Entrepreneurial Behavior & Research,* v. 30, n. 6, pp. 1349-1371. *https://doi.org/10.1108/IJEBR-02-2023-0123*

Golini, Natalia; Righi, Paolo. (2024). "Integrating probability and big non-probability samples data to produce Official Statistics". *Statistical Methods & Applications,* v. 33, n. 2, pp. 555-580. *https://doi.org/10.1007/s10260-023-00740-y*

Hribar, Paul. (2016). "Commentary On: Do Compustat Financial Statement Data Articulate?". *Journal of Financial Reporting,* v. 1, n. 1, pp. 61-63. *https://doi.org/10.2308/jfir-51355*

Ince, Ozgur S.; Porter, R. Burt. (2006). "Individual Equity Return Data From Thomson Datastream: Handle with Care!". *Journal of Financial Research,* v. 29, n. 4, pp. 463-479. *https://doi.org/10.1111/j.1475-6803.2006.00189.x*

Kalemli-Özcan, Şebnem; Sørensen, Bent E.; Villegas-Sanchez, Carolina; Volosovych, Vadym; Yeşiltaş, Sevcan. (2024). "How to Construct Nationally Representative Firm-Level Data from the Orbis Global Database: New Facts on SMEs and Aggregate Implications for Industry Concentration". *American Economic Journal: Macroeconomics,* v. 16, n. 2, pp. 353-374. *https://doi.org/10.1257/mac.20220036*

Liu, Grace. (2020). "Data quality problems troubling business and financial researchers: A literature review and synthetic analysis". *Journal of Business & Finance Librarianship,* v. 25, n. 3-4, pp. 315-371. *https://doi.org/10.1080/08963568.2020.1847555*

Martínez-Matute, Marta; Urtasun, Alberto. (2022). "Uncertainty and firms' labour decisions. Evidence from European countries". *Journal of Applied Economics,* v. 25, n. 1, pp. 220-241. *https://doi.org/10.1080/15140326.2021.2007724*

Martinez-Sanchez, Angel; Lahoz-Leo, Fernando. (2018). "Supply chain agility: a mediator for absorptive capacity". *Baltic Journal of Management,* v. 13, n. 2, pp. 264-278. *https://doi.org/10.1108/BJM-10-2017-0304*

Opazo-Basáez, Marco; Monroy-Osorio, Juan Carlos; Marić, Josip. (2024). "Evaluating the effect of green technological innovations on organizational and environmental performance: A treble innovation approach". *Technovation,* v. 129, pp. 102885. *https://doi.org/10.1016/j.technovation.2023.102885*

Pinto Ribeiro, Samuel; Menghinello, Stefano; De Backer, Koen. (2010). "The OECD ORBIS Database: Responding to the Need for Firm-Level Micro-Data in the OECD". *OECD Statistics Working Papers*, No. 2010/01. OECD Publishing, Paris. *https://doi.org/10.1787/5kmhds8mzj8w-en*

Rizov, Marian; Vecchi, Michela; Domenech, Josep. (2022). "Going online: Forecasting the impact of websites on productivity and market structure". *Technological Forecasting and Social Change,* v. 184, pp. 121959. *https://doi.org/10.1016/j.techfore.2022.121959*

Sánchez-Infante Hernández, Juan Pablo; Yañez-Araque, Benito; Moreno-García, Juan. (2020). "Moderating effect of firm size on the influence of corporate social responsibility in the economic performance of micro-, small- and medium-sized enterprises". *Technological Forecasting and Social Change,* v. 151, pp. 119774. *https://doi.org/10.1016/j.techfore.2019.119774*

Trezza, Domenico. (2023). "To scrape or not to scrape, this is dilemma. The post-API scenario and implications on digital research". *Frontiers in Sociology,* v. 8, pp. 1145038. *https://doi.org/10.3389/fsoc.2023.1145038*

Weiß, Gregor N. F.; Mühlnickel, Janina. (2014). "Why do some insurers become systemically relevant?". *Journal of Financial Stability,* v. 13, pp. 95-117. *https://doi.org/10.1016/j.jfs.2014.05.001*

Wu, Changbao. (2022). "Statistical inference with non-probability survey samples". *Survey Methodology,* v. 48, n. 2, pp. 283-311. *https://www150.statcan.gc.ca/n1/pub/12-001-x/2022002/article/00002-eng.htm*