CARMA 2024

*6th Int. Conf. on Advanced Research Methods and Analytics*
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17818

# A Comparative Analysis of Companies Missing from the SABI Database through BORME Gazette Web Scraping

**Xin-Hui Huang[1], Josep Domenech[2]**
[1]ETSINF, Universitat Politècnica de València, Spain, [2]DECS, Universitat Politècnica de València, Spain.

How to cite: Huang, X.; Domenech, J. 2024. A Comparative Analysis of Companies Missing from the SABI Database through BORME Gazette Web Scraping. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17818

***Abstract***

*This study aims to uncover parallels between the data issues observed in ORBIS and those in SABI, highlighting the need for cautious interpretation of both databases. By examining differences between entities present in SABI and those absent, insights into database representativeness are gained. Results indicate that SABI, like ORBIS, may not fully represent Spain's business population. Furthermore, analysis suggests that newer, smaller companies are less likely to appear in SABI, impacting data comprehensiveness. Extending this analysis, further variables will be explored to enhance understanding. This study underscores the importance of careful data scrutiny and the consideration of database limitations in research and decision-making processes.*

***Keywords:*** *SABI; BORME; ORBIS; bias; data; Python.*

## 1. Introduction

The Sistema de Análisis de Balances Ibéricos (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. Research using SABI illustrates its broad applicability across many different topics and studies (Martínez-Matute & Urtasun, 2022; Rizov et al. 2022; Sánchez-Infante, et al. 2020).

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. They present a comprehensive reach and frequently serve as a proxy for the total firm population in research (Opazo-Basáez et al., 2024; Garcés-Galdeano et al., 2024; Martinez-Sanchez and Lahoz-Leo, 2018). However, these databases do not exhaustively represent the corporate landscape, as they offer limited coverage, especially for small and micro firms (Bajgar et al. 2020, Almunia et al. 2018, Pinto Ribeiro et al. 2010). Therefore, the practice of considering companies listed there as the population may overlook the fact that they constitute a 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.

Simultaneously, the BORME[1] is the official gazette for business registrations and updates in Spain, providing a 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. Although the information that BORME contains for each firm is limited, the intersection of records between SABI and BORME reveals a unique opportunity to assess and enhance the completeness of business databases. The challenge lies in BORME's format—a text-based PDF 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 essential resources.

Our study aims to identify and describe companies that appear in the 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 (Liu 2020; Pinto Ribeiro et al. 2010; Bajgar et al. 2020; Culot et al. 2023), studies specifically focused on SABI are lacking. 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.

The remainder of the paper is structured as follows. Section 2 reviews some literature about the data quality issues in business databases. Section 3 explains the methods followed to obtain the data and their description. Section 4 the results obtained from the data with a brief explanation. Finally, Section 5 presents some concluding remarks.

## 2. Related work

The challenges related to data quality in business databases, particularly those like Orbis, are critical considerations for researchers relying on such sources for comprehensive firm-level

[1] Boletín Oficial del Registro Mercantil

information. This analysis delves into the prevalent issues, ranging from missing values to data errors and biases, affecting the reliability and representativeness of datasets like Orbis.

Biases in databases are widely discussed. Survivorship bias and selection bias were reported in Datastream and Orbis (Andrikopoulos et al., 2007; Ince & Porter, 2006; Kalemli-Ozcan et al., 2019). Biases can happen for various reasons. Overstatement by a statistical measure or index can result in an upward bias; when observations are excluded from the sample due to a selection rule other than random sampling, it can create selection bias and survivorship bias is an example of the selection bias driven by the disproportionate exclusion of stocks that were delisted over time.

Missing values are one of the most prevalent data quality problems. In Orbis, missing value can also occur due to the cap on the amount of data allowed to be downloaded (Kalemli-Ozcan et al., 2019) where the research emphasizes that “In spite of the extensive use of the Orbis database for research, firm-level data downloaded from this database are not nationally representative…” and finally they provide a guide for researchers on how to download and organize the data such that it ends up being nationally representative or comes close to being so. Also researchers sometimes take special procedures or filters to exclude missing values from the research sample. However, this practice may inevitably create omission bias or selection biases (Elton et al., 2001; Weiß & Muhlnickel, 2014; Liu, 2020). Dropping all observations that contain missing values is a naïve strategy and can have a marked effect on the statistical power of the tests (Hribar, 2016, p. 63) and excluding these missing values can create misleading results. This great number of missing values may make a database not usable for specific research (Francis et al., 2016; Lee, 2017).

There are also other problems like data errors (Monasterolo et al., 2017), inconsistencies (Kalemli-Ozcan et al., 2019), static header data issue or vintage issue (Kalemli-Ozcan et al., 2019) and reporting time issues (Kalemli-Ozcan et al., 2019). Also, it is worth noting that when making comparisons and using ORBIS data, caution is required, especially when dealing with different countries. This is because some variables or data may not refer exactly to the same thing, as mentioned on the ORBIS website: "Our reports are in standardized formats to accommodate regional variations in filing regulations and accountancy practices..." Although this is secondary to the issues with the data itself. Therefore, when conducting studies with these types of databases, careful attention must be paid to all the potential issues they may present. It is crucial to approach research with a clear understanding of the challenges inherent in these databases. This issue is not exclusive to specific databases; similar problems can also be encountered in SABI.

## 3. Data

### 3.1. Data sources

Data for companies established between 2010 and 2023 have been collected from two sources: SABI and BORME. SABI is a commercial database and provides data in a convenient tabular format. After selecting those established in Spain within the period under study, a list of 1,911,775 companies was retrieved.

However, BORME is a set of daily publications in PDF format, available on the official website. By means of web crawling and scraping techniques, those publications were downloaded and converted to text. The structure of each publication is a list of entries in the registry where each entry corresponds to a company. Each entry is associated with specific registry events, such as the establishment of the company, a capital increase or a declaration of bankruptcy.

To construct the dataset, around 100,000 publications were downloaded and 9,956,791 registry entries corresponding to 3,051,505 companies were processed. After filtering out companies not established in the selected period, 2,917,784 entries associated with 1,298,056 companies were kept. Each registry entry was transformed into tabular format and grouped by company. Data from SABI and BORME were finally merged to create the dataset used in the analysis.

### 3.2. Data description

The final dataset consisted of 1,298,056 companies. Table 1 describes the variables used in this paper, although some others representing various registry events were also collected and processed.

The downloaded database from SABI that is ultimately used consists of a total of 1,911,775 companies and the parameters/variables selected include the company name, the NIF code of the company, the BvD number that identifies the company in SABI, the province, and finally, the date of incorporation. After downloading and loading the database, certain process where applied to adjust to what is needed just like selecting only companies between the years 2010 and 2023, text elimination, date transformation, text transformations, etc.

As for the BORME database, it consists of much larger dataframes, just like mentioned before in Barcelona it has 1,5 million rows and a 33 columns. Each row in the dataframe corresponds to a record of an event that has occurred and was registered by the company in a certain province on a certain date. It's worth noting that a company can appear multiple times in the dataframe. And some columns are “nombre actual”, “constitución”, “fecha”, “provincia”, “extinción”, etc. Which some of them are dates data types or numbers but most of them are texts. This dataframe undergoes transformations, where, as mentioned earlier, the records are grouped by companies (using the current names of the companies). The variables have been transformed from texts to numerics and it been grouped by other parameters, such as the registration year or the year of incorporation.

After the merging of both datasets, we end up with a dataset where each row corresponds to a company, and each row contains different variables. For this time only“ extinción” and others variables, such as the year of incorporation for each company or whether that company is present or not in the SABI database are used.

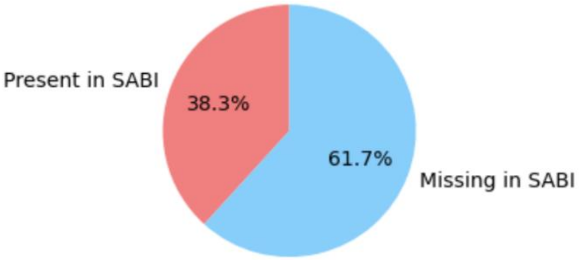
*Table 1. The variables used for the study explained and the type of the variable.*

| Variable | Definition | Units |
|---|---|---|
| Year | The year when the event was registered. | Date (Year) |
| Year of incorporation | The year when the company was created/ incorporated. | Date (Year) |
| Sabi | Whether the company is present in SABI or not. | Binary |
| Company | Name of the company | Text |
| Dissolution | If the company in that year has or not dissolved. | Binary |

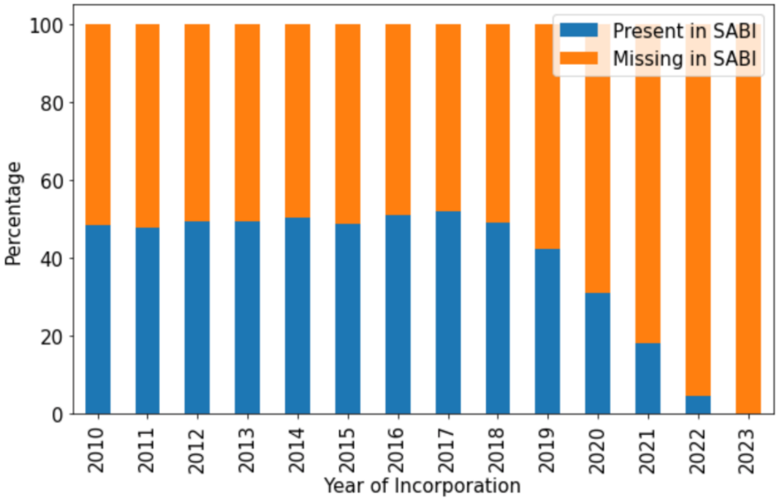
## 4. Results

### 4.1. First insight

In the initial study, a broad and generic analysis was conducted. To achieve this, a pie chart was created to observe the proportion of companies that are present in SABI and those that are not. Then, conducting a more in-depth study, a stacked bar chart where the X-axis reflects the year of incorporation of the companies, and the Y-axis represents the percentage. The 100% on the Y-axis corresponds to all the companies incorporated during that year. Each stacked bar is divided into two parts: the blue section denotes those present in SABI, and the orange section represents those absent in SABI.



*a)*



*b)*

*Figure 1. The proportion of companies present in SABI and missing in SABI (a) in general, and (b) by year of incorporation.*

As the graph depicts, in red, approximately 40% represents the percentage of companies registered in the SABI database, while in blue, around 60%, represents those not found in SABI. And observing the second graph, a distinction can be made between two periods. The first period spans from 2010 to 2018, where the proportion remains relatively constant, close to 50%. However, in the second period starting from 2018, there is a decreasing trend each year, with an average decrease of 7.2% per year, approaching the present.

## 4.2. Dissolution

Having explored the companies in a general context, we will delve deeper by studying more specific variables. In this instance, the focus is on analyzing the dissolution variable, aiming to identify potential patterns or differences between companies present and absent in SABI. To address this analysis, we have created two complementary graphs. The first is a temporal graph representing in blue the total dissolution of SABI-listed companies and in orange those not in SABI. The second graph is a stacked bar chart illustrating the proportion of both categories. It is essential to note that the years in these graphs do not represent the year of each company's establishment, but rather the year in which the dissolution event is recorded.

It can be observed from the first graph that, in both cases, the growth is positive. This is not surprising since, as mentioned earlier regarding the years, as the year increases, so does the number of companies, which can explain this growth. Although both show an increase, companies not present in SABI tend to dissolve more in the early years, excluding 2010 (a special case). Analyzing both graphs, and more prominently in the second one, it can be noted that as the year is more recent, SABI-listed companies also tend to dissolve more, approaching the percentage of those not in SABI. This phenomenon occurs with an average growth of around 1.41% per year. Although there is a balance in the most recent years, in subsequent years, companies outside SABI tend to dissolve more.

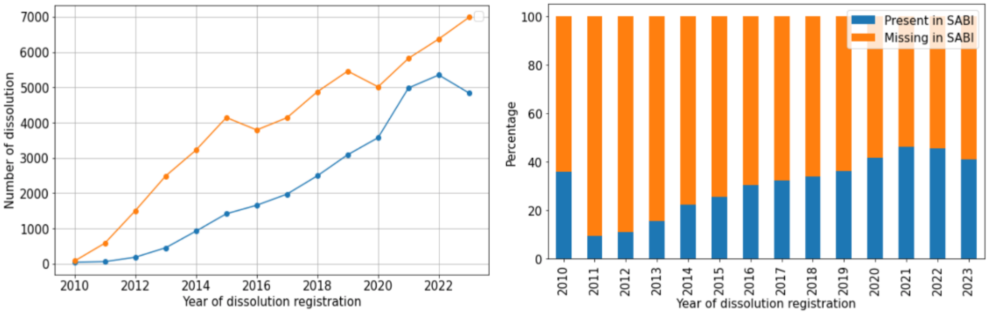


*Figure 2. Quantity and proportion of dissolutions per year for companies that are present in SABI and those that are missing from SABI.*

## 5. Conclusion

From the obtained results, it can be concluded that the issue present in ORBIS is also evident in SABI. The companies collected in SABI cannot be considered the complete population of Spain, emphasizing the need for caution when analyzing this data. Also observing the second graph from the figure 1, the decline may be attributed to the fact that, as the year of incorporation becomes more recent, companies are generally smaller and newer, making them less likely to be included in SABI. That can also be corroborated after studying the extinction where not in SABI companies tend to extinct since for smaller and newer companies is more challenging to sustain. Although we cannot conclusively state this until we have explored more variables, which will be done subsequently and not included here due to the limited number of pages that can be included.

## References

Andrikopoulos, P., Daynes, A., Pagas, P., & Latimer, D. (2007). UK market, financial databases and evidence of bias (Occasional Paper Series Paper No. 79). http://www.dmu.ac.uk/documents/business-and-law-documents/business/occasional-papers/paper79ukmarketfinancialdatabasesandrikopoulos.pdf

Annaert, J., Buelens, F., & Riva, A. (2016). Financial history databases: Old data, old issues, new insights? In D. Chambers & E. Dimson (Eds.). Financial market history (pp. 44–65). Charlottesville, VA: CFA Institute Research Foundation.

Arbelo, A., Arbelo-Pérez, M. & Pérez-Gómez, P. (2022). Are SMEs less efficient? A Bayesian approach to addressing heterogeneity across firms. Small Bus Econ 58, 1915–1929. https://doi.org/10.1007/s11187-021-00489-2

Bajgar, M., et al. (2020), Coverage and representativeness of Orbis data. OECD Science, Technology and Industry Working Papers, No. 2020/06. https://doi.org/10.1787/c7bdaa03-en

Blanco-Mazagatos, V., Romero-Merino, M.E. & Santamaría-Mariscal, M. et al. One more piece of the family firm debt puzzle: the influence of socioemotional wealth dimensions. Small Bus Econ (2024). https://doi.org/10.1007/s11187-024-00881-8

Bostwick, E. D., Lambert, S. L., & Donelan, J. G. (2016). A wrench in the COGS: An analysis of the differences between cost of goods sold as reported in Compustat and in the financial statements. Accounting Horizons, 30(2), 177–193. https://doi.org/10.2308/acch-51336

Chychyla, R., & Kogan, A. (2014). Does Compustat data standardization improve bankruptcy prediction models? Social Science Research Network. http://ssrn.com/abstract=2406136

Elton, E. J., Gruber, M. J., & Blake, C. 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, 56(6), 2415–2430. https://doi.org/10.1111/0022-1082.00410

Francis, R. N., Mubako, G., & Olsen, L. (2016). Archival research considerations for CRSP data. Social Science Research Network. https://ssrn.com/abstract=2608273

Hribar, P. (2016). Do Compustat financial statement data articulate? Journal of Financial Reporting, 1(1), 61–63. https://doi.org/10.2308/jfir-51329

Ince, O. S., & Porter, R. B. (2006). Individual equity return data from Thomson Datastream: Handle with care! Journal of Financial Research, 29(4), 463–479. https://doi.org/10.1111/j.1475-6803.2006.00189.x

Kalemli-Ozcan, S., Sorensen, B., Villegas-Sanchez, C., Volosovych, V., & Yesiltas, S. (2019). How to construct nationally representative firm level data from the Orbis Global Database: New facts and aggregate implications (No. w21558). National Bureau of Economic Research. https://www.nber.org/papers/w21558.pdf

Lee, J. (2017). How do firms choose their debt types? http://www.fmaconferences.org/Boston/P1_201608.pdf

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

Martín-Rojas, R., Garrido-Moreno, A. & García-Morales, V. J. (2020). Fostering Corporate Entrepreneurship with the use of social media tools. Journal of Business Research, 112, 396-412. https://doi.org/10.1016/j.jbusres.2019.11.072

Martínez-Matute, M., & Urtasun, A. (2022). Uncertainty and firms' labour decisions. Evidence from European countries. Applied Economics, 25(1), 220-241. https://doi.org/10.1080/15140326.2021.2007724

Monasterolo, I., Battiston, S., Janetos, A. C., & Zheng, Z. (2017). Vulnerable yet relevant: The two dimensions of climate-related financial disclosure. Climatic Change, 145(3–4), 495–507. https://doi.org/10.1007/s10584-017-2095-9

Opazo-Basáez, M., Monroy-Osorio, J. C. & Marić, J. (2024). Evaluating the effect of green technological innovations on organizational and environmental performance: A treble innovation approach. Technovation, 129, 102885. https://doi.org/10.1016/j.technovation.2023.102885

Pinto Ribeiro, S., Menghinello, S., & De Backer, K. (2010). The OECD ORBIS Database: Responding to the Need for Firm-Level Micro-Data in the OECD. OECD Statistics Working Papers, No. 2010/01. https://doi.org/10.1787/5kmhds8mzj8w-en

Rico, M., Pandit, N.R. & Puig, F. (2021). SME insolvency, bankruptcy, and survival: an examination of retrenchment strategies. Small Bus Econ 57, 111–126. https://doi.org/10.1007/s11187-019-00293-z

Rizov, M., Vecchi, M. & Domenech, J. (2022). Going online: Forecasting the impact of websites on productivity and market structure. Technological Forecasting and Social Change, 184, 121959. https://doi.org/10.1016/j.techfore.2022.121959

Sánchez-Infante Hernández, J. P., Yañez-Araque, B. & Moreno-García, J. (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, 151, 119774. https://doi.org/10.1016/j.techfore.2019.119774

Sánchez-Vidal, F. J., Hernández-Robles, M. & Mínguez-Vera, A. (2020): Financial conservatism fosters job creation during economic crises. Applied Economics 52:45, 4913-4926 https://doi.org/10.1080/00036846.2020.1751053

Segarra, A., Callejón, M. (2002). New Firms' Survival and Market Turbulence: New Evidence from Spain. Review of Industrial Organization, 20, 1–14. https://doi.org/10.1023/A:1013309928700

Weiß, G. N., & Mühlnickel, J. (2014). Why do some insurers become systemically relevant?. Journal of Financial Stability, 13, 95–117. https://doi.org/10.1016/j.jfs.2014.05.001

Yáñez-Araque, B., Sánchez-Infante Hernández, J. B., Gutiérrez-Broncano, S. & Jiménez-Estévez, P. (2021). Corporate social responsibility in micro-, small- and medium-sized enterprises: Multigroup analysis of family vs. nonfamily firms. Journal of Business Research, 124, 581-592. https://doi.org/10.1016/j.jbusres.2020.10.023