



WIN2025

Evaluating the completeness of business databases

Evaluating the completeness of business databases A comparison with official records using web scraping techniques

Xin-Hui Huang, **Josep Domènech**

Universitat Politècnica de València

Motivation


- SABI is widely used for business research
 - It is the local version of the ORBIS database
- BORME, the official business registry, provides a complementary source.
- **Aim:** Compare BORME and SABI to identify missing companies.

Research objectives

- Assess how many companies established 2010–2023 are missing in SABI
- Examine coverage by year of establishment, dissolution, province, sector, and legal form
- Uncover potential biases
 - (e.g., survivorship bias, size bias)

Data Sources

- SABI (Bureau van Dijk)
 - Commercial database with financial and registration data
 - Contains around 1.9 million Spanish companies (as of January 2024)



Las empresas con datos editados se presentan en azul [Modificar](#)

	Nombre	Fecha constitución	Número BvD
1.	REPSOL PETROLEO SA	27/06/1949	ESA28047223
2.	MERCADONA SA	18/01/1977	ESA46103834
3.	COMPAÑIA ESPAÑOLA DE PETROLEOS SA	26/09/1929	ESA28003119
4.	REPSOL COMERCIAL DE PRODUCTOS PETROLIFEROS SA	26/03/1992	ESA80298839
5.	ENDESA ENERGIA SAU	03/02/1998	ESA81948077
6.	NATURGY APROVISIONAMIENTOS SA.	30/12/1999	ESA82531062
7.	GAS NATURAL COMERCIALIZADORA SA	29/10/1998	ESA61797536
8.	ENGIE ESPAÑA SL.	29/11/1999	ESB82508441
9.	INDUSTRIA DE DISEÑO TEXTIL SA	12/06/1985	ESA15075062
10.	STELLANTIS ESPAÑA SOCIEDAD LIMITADA.	31/03/1995	ESB50629187
11.	IBERDROLA CLIENTES SOCIEDAD ANONIMA.	17/03/2014	ESA95758389
12.	CAIXABANK SA	12/12/1980	ESA08663619
13.	REPSOL TRADING SA.	29/10/1999	ESA82489451
14.	EL CORTE INGLES SA	02/01/1952	ESA28017895
15.	SEAT SAU	09/05/1950	ESA28049161
16.	PETROLEOS DEL NORTE SOCIEDAD ANONIMA	01/01/1968	ESA48053243
17.	BP ENERGIA ESPAÑA S.A.U.	04/12/1964	ESA28135846
18.	SOCIEDAD ESTATAL LOTERIAS Y APUESTAS DEL ESTADO SME SA	17/03/2011	ESA86171964
19.	CENTROS COMERCIALES CARREFOUR SA	01/01/1976	ESA28425270
20.	FORD ESPAÑA SL	26/09/1973	ESB46066361
21.	TELEFONICA DE ESPAÑA SAU	24/04/1998	ESA82018474
22.	PEUGEOT CITROEN AUTOMOVILES ESPAÑA, SA (EXTINGUIDA)	02/07/1957	ESA36602837
23.	ENDESA GENERACION SA	22/09/1999	ESA82434697

Data Sources

- BORME (Official Gazette in Spain)
 - Daily PDF publications by each registry
 - Essential legal information on all newly registered and dissolved companies

SECCIÓN PRIMERA

Empresarios

Actos inscritos

VALENCIA/VALÈNCIA

38512 - CITRISE SOCIEDAD LIMITADA.

Modificaciones estatutarias. Fecha de Cierre de Ejercicio: 31/12. Datos registrales. S 8 , H V 41910, I/A 12 (17.01.25).

38513 - YNNI FV2 SOLAR SL.

Ceses/Dimisiones. Adm. Unico: SELLES MARTINEZ MANUEL. LiqUnico: SELLES MARTINEZ MANUEL. Nombramientos. LiqUnico: SELLES MARTINEZ MANUEL. Otros conceptos: Cambio del Organó de Administración: Administrador único a Liquidador Unico. Disolución. Voluntaria. Extinción. Datos registrales. S 8 , H V 206280, I/A 3 (17.01.25).

38514 - AGRIPA EXPLOTACIONES AGRICOLAS SL.

Ceses/Dimisiones. Adm. Unico: ALMENAR BAVIERA TERESA. Nombramientos. Adm. Unico: AYARZA ALMENAR ANTONIO. Datos registrales. S 8 , H V 198068, I/A 5 (17.01.25).

38515 - JATIVA URBANA SOCIEDAD LIMITADA.

Ceses/Dimisiones. Adm. Unico: CARRASCOSA RICHART ERNESTO. LiqManc: CARRASCOSA MARTINEZ ERNESTO;CARRASCOSA MARTINEZ JAVIER ERNESTO. Nombramientos. LiqManc: CARRASCOSA MARTINEZ ERNESTO;CARRASCOSA MARTINEZ JAVIER ERNESTO. Otros conceptos: Cambio del Organó de Administración: Administrador único a Liquidadores Mancomunados. Disolución. Voluntaria. Extinción. Datos registrales. S 8 , H V 19489, I/A 13 (17.01.25).

38516 - BEAUTY ILDIVA SOCIEDAD LIMITADA.

Constitución. Comienzo de operaciones: 23.12.24. Objeto social: Actividad principal: Comercio al por menor por correspondencia o internet de productos diversos. Siendo el código de clasificación nacional de actividades económicas (C.N.A.E.) de la actividad principal el 47.91. - Otras actividades: Comercio al por mayor de productos de perfumería y droguería a ... Domicilio: C/ VALENCIA 9 (ALBALAT DELS SORELLS). Capital: 3.000,00 Euros. Nombramientos. Adm. Unico: SOVEN VIECO MIRIAM. Datos registrales. S 8 , H V 223050, I/A 1 (16.01.25).

38517 - AGENCIA NOVA CATARROJA SOCIEDAD LIMITADA.

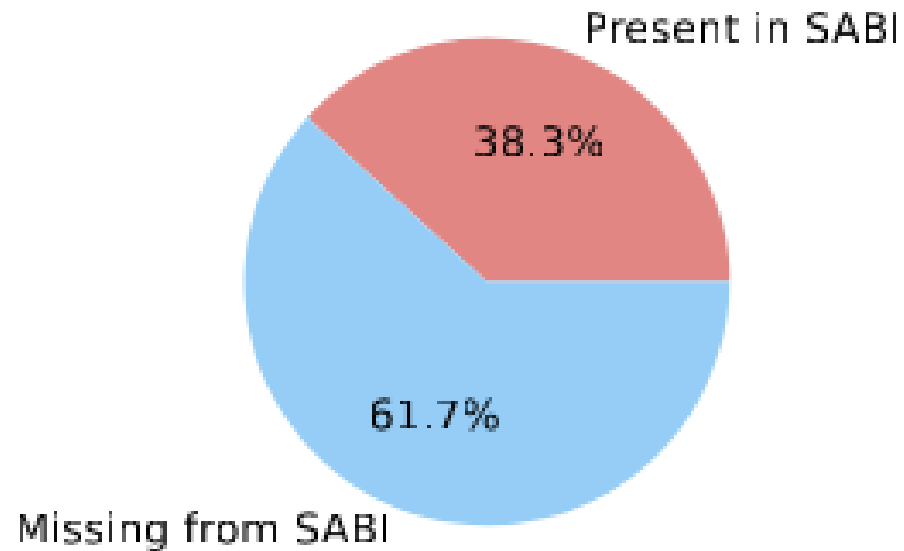
Ceses/Dimisiones. Adm. Unico: ESTEVE MARTINEZ PATRICIA. Nombramientos. Adm. Unico: ROMERO PENA BERNABE. Datos registrales. S 8 , H V 201283, I/A 6 (16.01.25).

Methodology

- Downloaded and parsed ~100k PDF publications from BORME
- Extracted ~9.96 million entries on 3.05 million firms
- Entries were divided in fields (establishment, dissolution, legal form...)
- Filtered for companies established 2010–2023
 - 1.30 million firms
- Matched with SABI records to identify missing entities

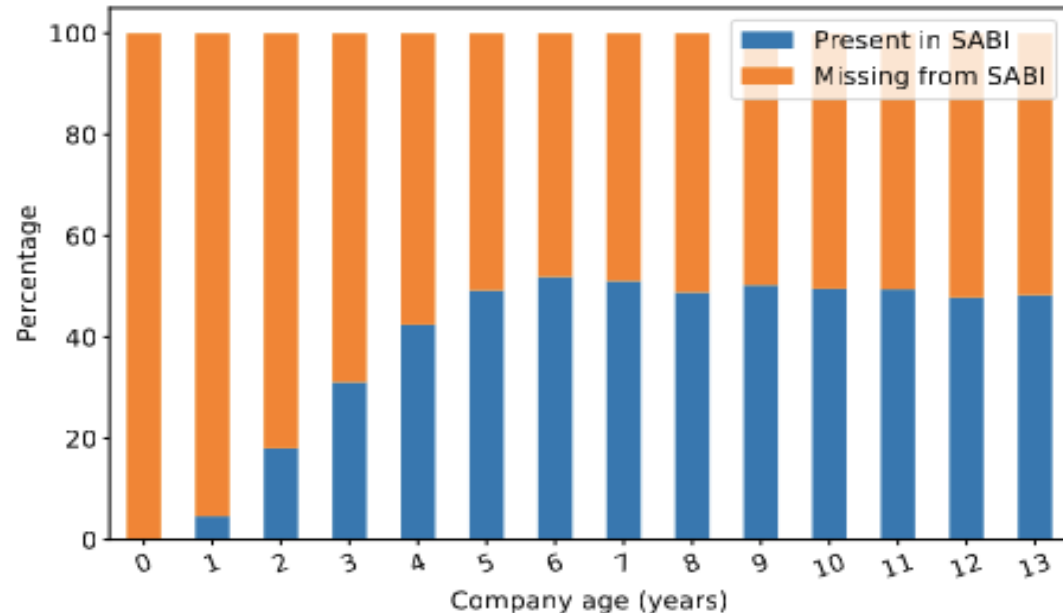
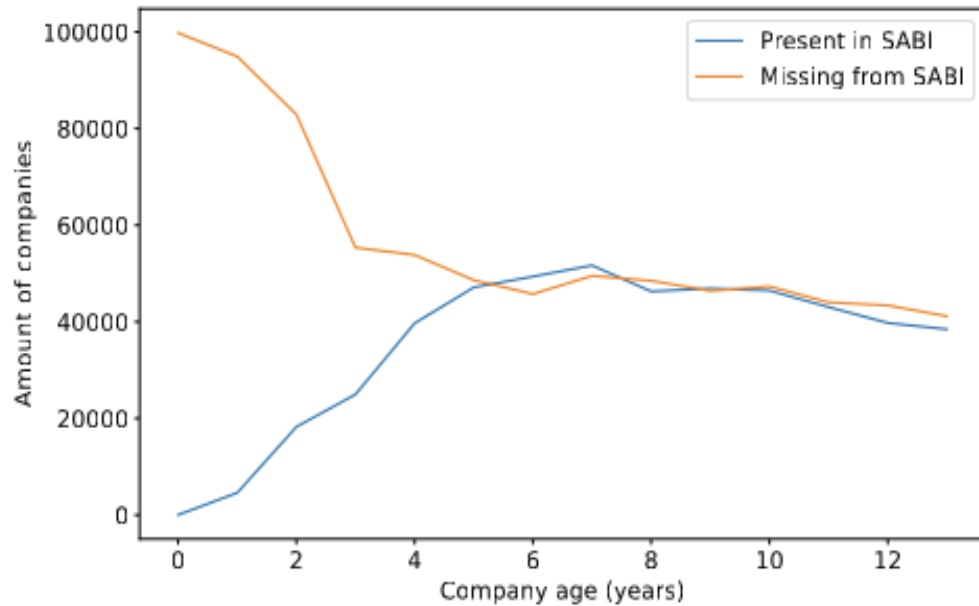
Coverage

- Only 38.3% of firms created 2010–2023 found in SABI



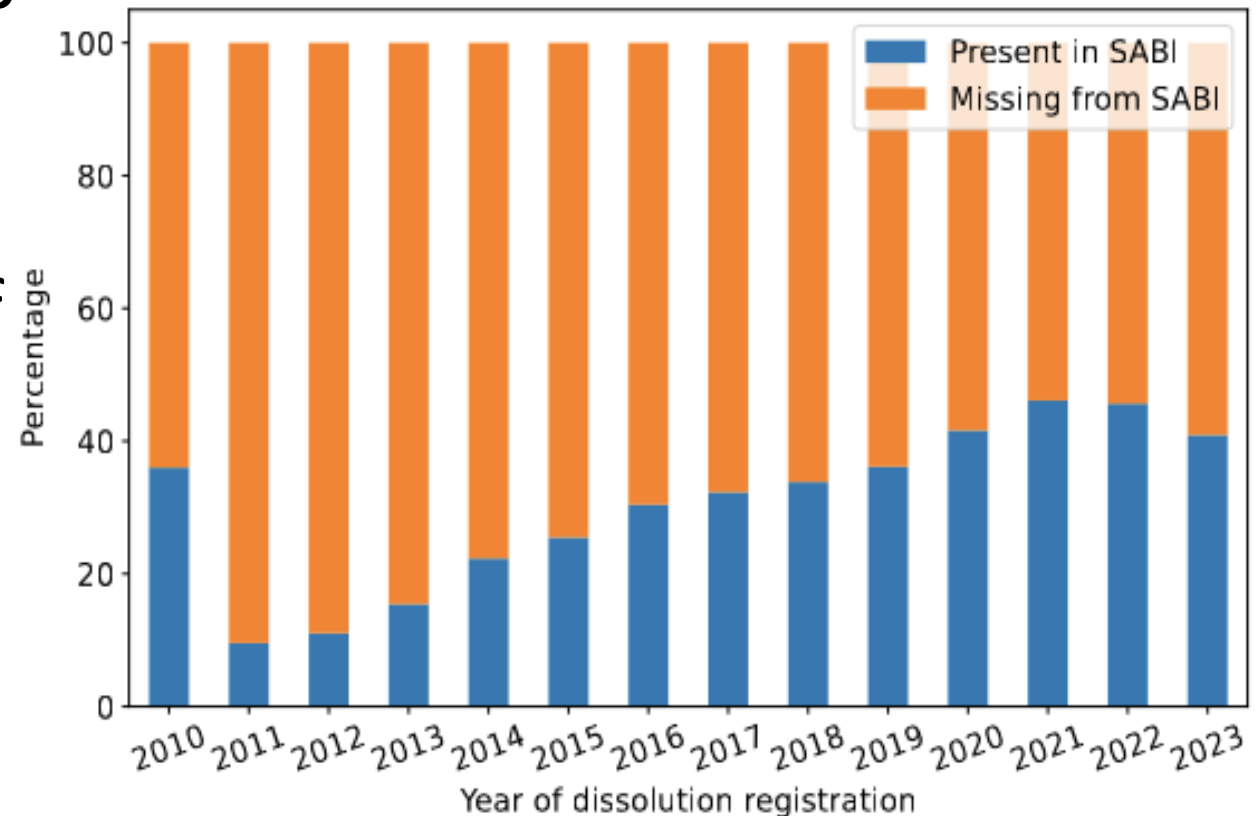
Coverage by Year of Establishment

- Marked underrepresentation of young firms
- ~50% coverage starts around age 5
- **Implication:** SABI underestimates recent entries



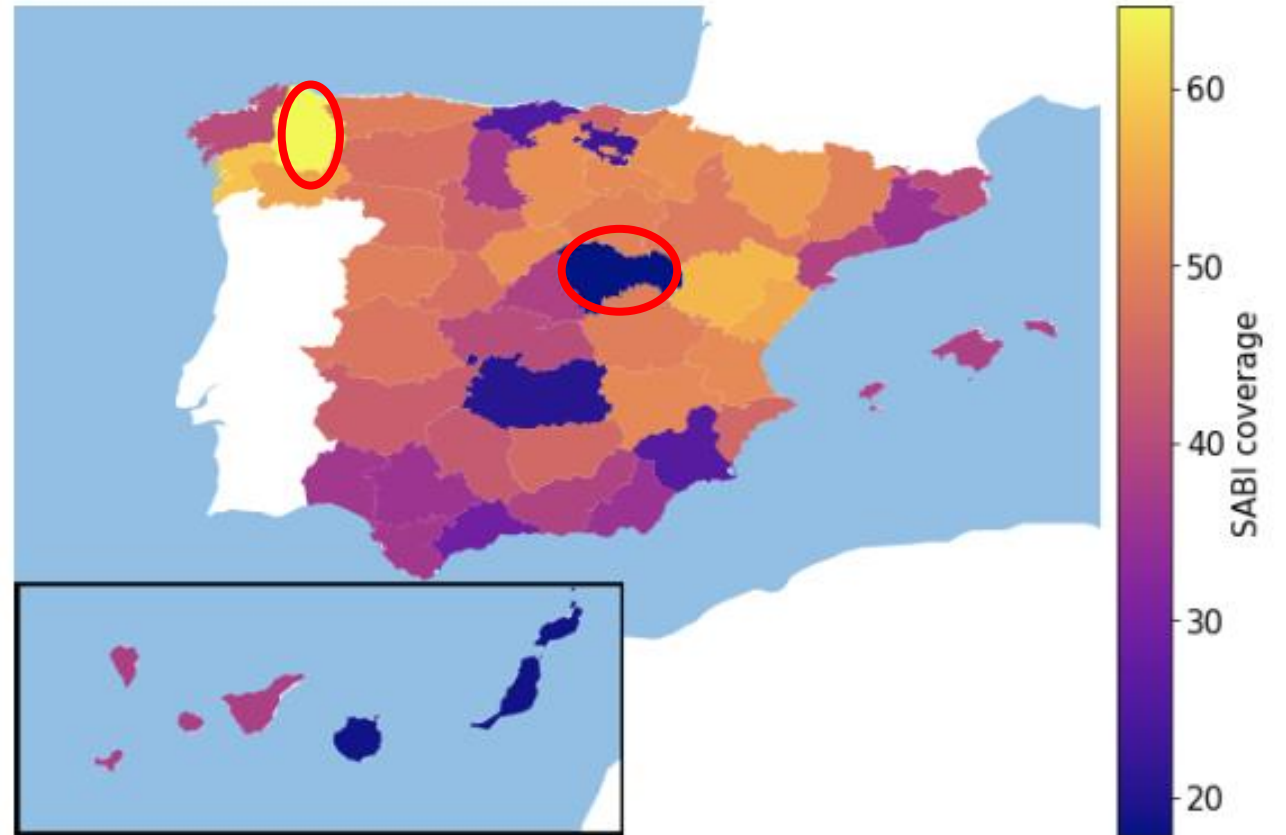
Survivorship Bias

- Dissolved firms gradually disappear from SABI
- Coverage drops around 3.7 pp per year after dissolution
- Results can overestimate the longevity and success rates of firms



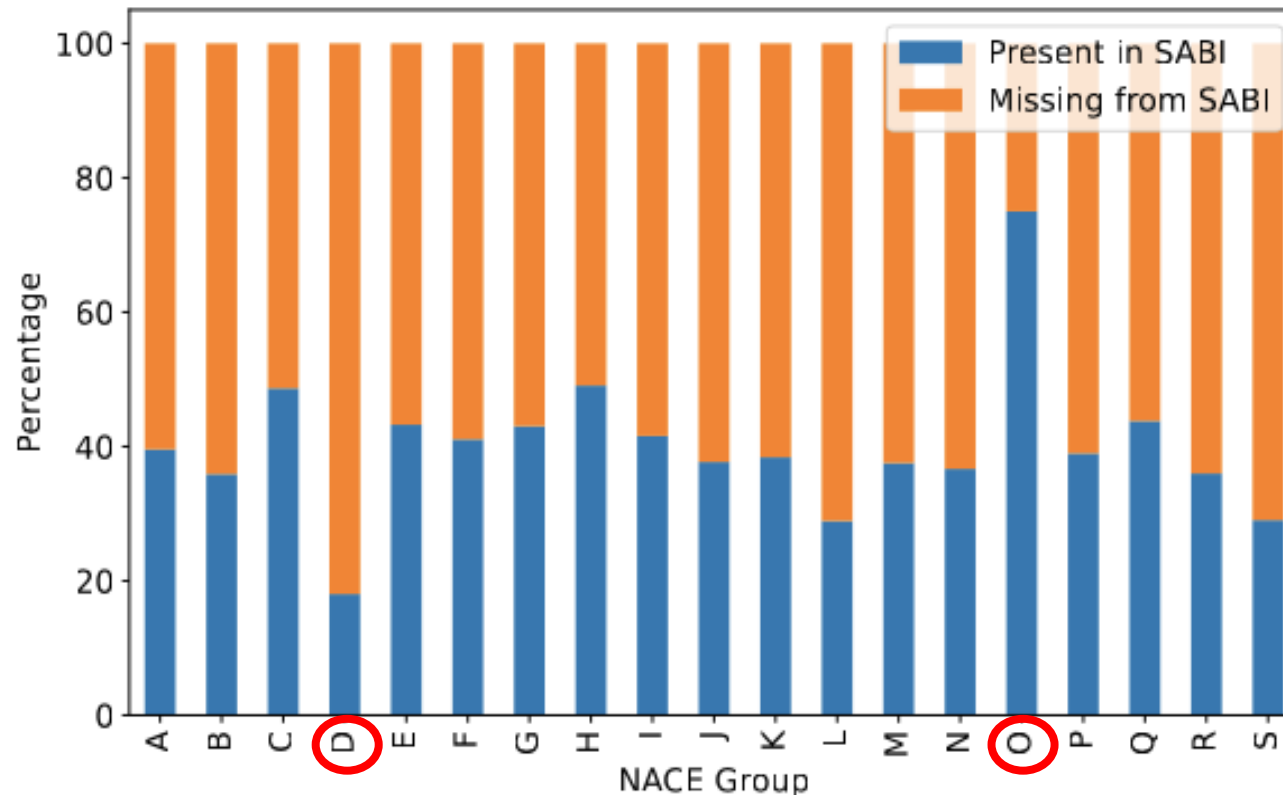
Geographic variation

- Coverage ranges from ~65% (Lugo) to ~18% (Guadalajara)
- No clear link between coverage and population or coastal proximity



Sectoral variation

- Some industries better covered (e.g., Manufacturing), others underrepresented (e.g., Electricity supply)



Legal form variation

- Legal forms with higher capitalization (SA similar to PLC) have better coverage
- SL forms (similar to LLC) are underrepresented

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%

Conclusion

- SABI, like ORBIS, also has coverage limitations.
- It includes well-known companies, but many newer and lesser-known companies are not registered here. Additionally, it may exclude companies dissolved long ago, thus presenting various biases.
- It is necessary to correct these biases, including geographic, sectoral, survival, or selection biases, when preparing a research sample.

Limitations

- Time focus: 2010–2023 (does not capture longer-term trends)
- Companies may fail to deposit statements in the registry
- SABI's own data collection and removal policies are not fully transparent

Conclusion

- SABI covers a limited share of the Spanish corporate landscape, especially recent or small firms
- Survivorship bias and other coverage gaps can affect research results
- Researchers should exercise caution and consider alternative data or methodological corrections



WIN2025

Evaluating the completeness of business databases

Evaluating the completeness of business databases A comparison with official records using web scraping techniques

Xin-Hui Huang, **Josep Domènech**

Universitat Politècnica de València

jdomenech@upvnet.upv.es