DEPARTMENT OF STATISTICS SINGAPORE
Empowering You with Trusted Data

Experimental Use of Machine Learning and New Data Sources in the Updating of the Statistical Business Register

Meeting of the Group of Experts on Business Registers Online
26-29 September 2022

Outline

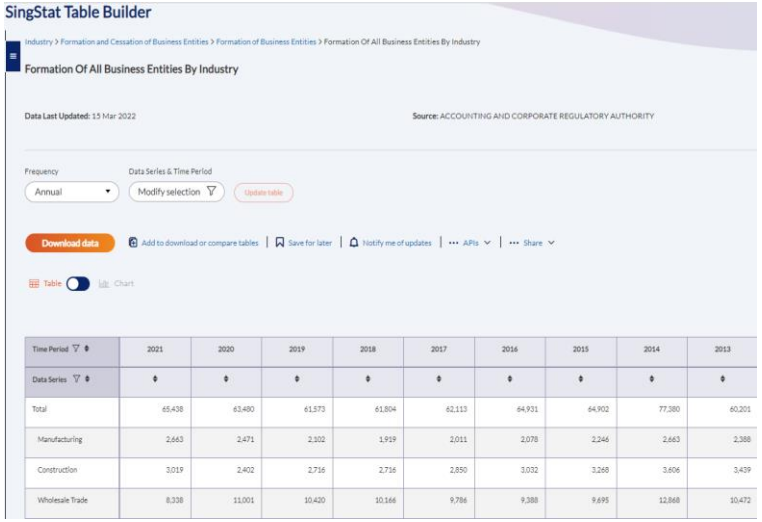
- Roles of Statistical Business Register (SBR)
- Data sources for updating of the SBR
- Experimental use of machine learning and new data sources

Roles of Statistical Business Register

The Statistical Business Register (SBR) serves as the foundational statistical infrastructure for the compilation of business and economic statistics, and contains key information such as enterprise name, Unique Entity Number (UEN), registration date and industrial classification (SSIC).

Provide comprehensive coverage for survey sampling frame and contact information for use in conducting of business surveys

Produce indicators on characteristics of firms (e.g. firm's formation and cessation by industry, number of startups) and support firm-level data integration with other data sources for in-depth analysis



Data Sources for Updating of the SBR

Multiple data sources are integrated in the update of the SBR:

- Primarily based on various administrative data, supplemented by statistical survey returns from DOS and Research & Statistics Units (RSUs)

Administrative Source	Administrative Data
• **Regulatory Authority of Business Registration & Financial Reporting**: Accounting and Corporate Regulatory Authority (ACRA) of Singapore	Identification and demographic information (e.g. UEN, business name, registration date, shareholder information, industrial classification)
• ACRA • **National Tax Authority**: Inland Revenue Authority of Singapore (IRAS)	Financial Information (e.g. Revenue, Profit)
• **Manpower Authority:** Ministry of Manpower (MOM), Central Provision Fund Board (CPF)	Employment and Wages
• **Authority for trade facilitation**: Singapore Customs / Enterprise Singapore	Merchandise Trade (i.e. imports, exports)

- Experimental use of AI/ML to text mine big data and unstructured data from admin sources

DEPARTMENT OF STATISTICS SINGAPORE
Empowering You with Trusted Data

Challenges and Solution

Despite the plethora of administrative data:

- Data may not be timely (e.g. corporate tax filings only available 1-2 years after firm's financial year ending) or not readily available (e.g. unstructured data)

- Data of increasing interest may not have been collected administratively

In response to these challenges and increasingly complex data demands, DOS has explored new data sources and developed innovative capabilities to supplement existing data sources:

- Using Machine Learning and Web-based Data to profile firms with internet presence

- Leveraging on Artificial Intelligence (AI) for data extraction and processing of unstructured data in financial accounts

Web-based data to profile firms with internet presence

Through **web-scraping** of the internet, **text-mining** firms' corporate website and the use of **machine learning** techniques, we were able to derive new firm characteristics on whether firms have website and their usage of their corporate websites.

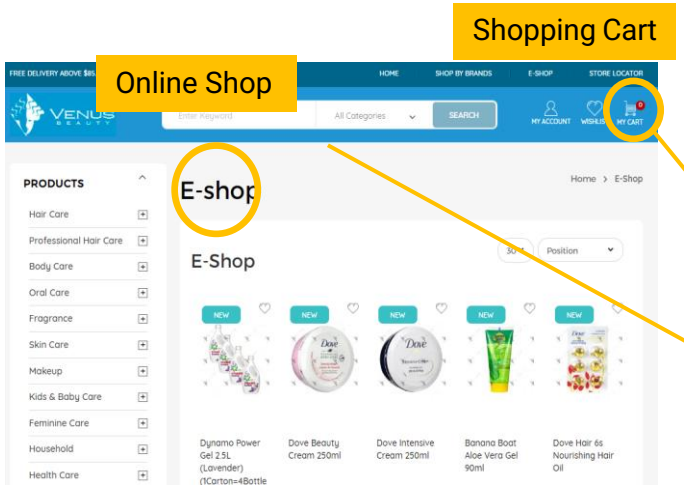
Sources

Business Surveys

Domain Registry

Step 1: Obtain corporate website address of enterprises

Extract Feature(s)



Step 2: Web scraping of corporate website and extracting features

Classify

Enterprises without websites (A)

Enterprises with websites (B)

Websites that generate revenue online (C)

Step 3: Machine Learning Classification

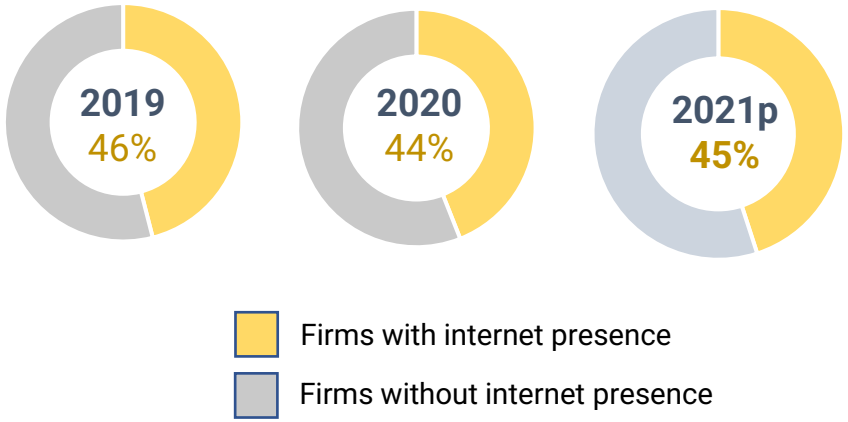
DEPARTMENT OF STATISTICS SINGAPORE
Empowering You with Trusted Data

Web-based data to profile firms with internet presence

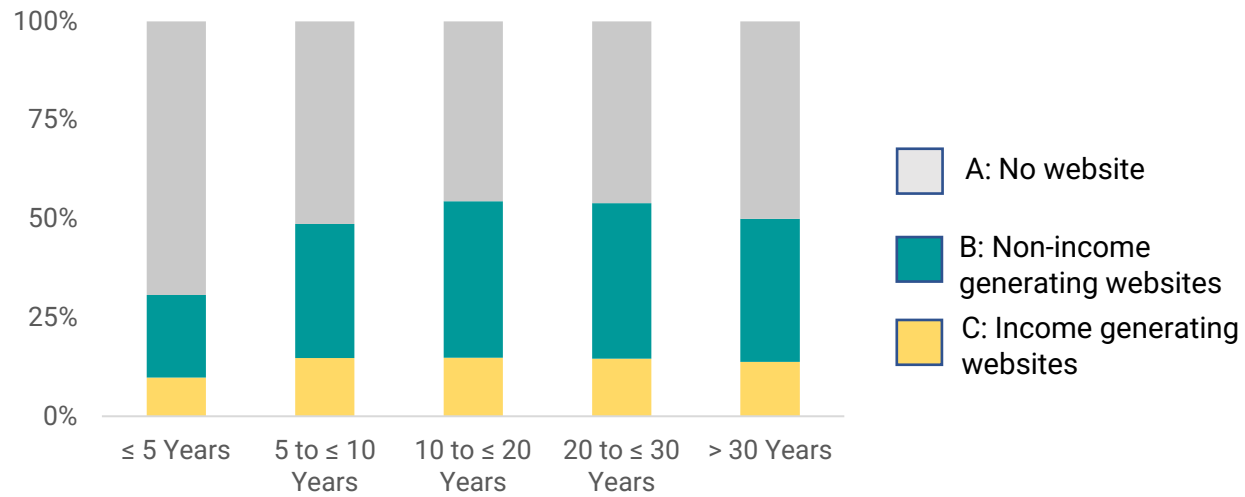
New insights can be generated by integrating information on the firms' internet presence with other firm characteristics data in the SBR:

- Almost half of all firms in Singapore had a website in 2021, with the proportion remaining relatively stable for the past three years.
- Only one in three firms aged 5 years or less had a corporate website in 2021.

Share of firms with internet presence, 2019 - 2021

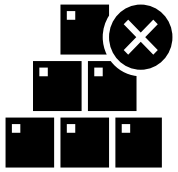


Share of firms with internet presence by age group, 2021



AI for data extraction of unstructured data

There is a **rich amount of financial information, but some of the data are in unstructured format in the financial statements** and **cannot be easily processed** by the system.



Significant manual effort required to **read, analyse and extract information from** the unstructured financial statements

Cannot **scale up** quickly to process a larger number of financial statements

4 PROPERTY, PLANT AND EQUIPMENT

	Freehold residential properties $'000	Freehold office properties $'000	Renovations $'000	Office equipment, furniture and fittings $'000	Computer equipment $'000	Motor vehicles $'000	Total $'000
Group							
Cost							
At 1 January 2020	183	3,331	432	1,467	1,221	616	7,250
Additions	–	–	61	5	60	55	181
Disposals	–	–	–	–	–	(55)	(55)
Write-offs	–	–	–	–	(13)	–	(13)
Effect of movements in exchange rates	–	(41)	(2)	(20)	(8)	(9)	(80)
At 31 December 2020	183	3,290	491	1,452	1,260	607	7,283
Accumulated depreciation and impairment							
At 1 January 2020	126	1,795	272	1,382	1,057	453	5,085
Depreciation for the year	3	184	48	34	54	67	390
Disposals	–	–	–	–	–	(55)	(55)
Write-offs	–	–	–	–	(13)	–	(13)
Effect of movements in exchange rates	–	(21)	(1)	(18)	(7)	(6)	(53)
At 31 December 2020	129	1,958	319	1,398	1,091	459	5,354
Carrying amounts							
At 1 January 2020	57	1,536	160	85	164	163	2,165
At 31 December 2020	54	1,332	172	54	169	148	1,929

31 FEE AND COMMISSION INCOME

	Group 2020 $'000	2019 $'000
Fee income	6,314	8,500
Underwriting commission income	183	132
	6,497	8,632

The fee income are service fees from provision of loans to the customers, received/receivable on the disbursement of the loans, subject to the loan agreements

32 NET INVESTMENT INCOMEE

	Group 2020 $'000	2019 $'000
Exchange gain/(loss), net	32	(160)
Dividend income	376	678
Loss on disposal of debt securities	(66)	(8)
Net change in fair value of financial assets through profit or loss	(528)	3,374
Interest income from bonds, fixed deposits and others	909	1,155
Amortisation of debt securities at amortised cost	(14)	(57)
	709	4,982

33 OTHER INCOME

	Group 2020 $'000	2019 $'000
Recoveries - loans, advances and receivables#	88	855
Gain on disposal of property, plant and equipment	15	–
Others	358	658
	461	1,513

Represents excess amount of loans, advances and receivables recovered.

Note: Information is sourced from the annual report made available on a firm's corporate website.

DEPARTMENT OF STATISTICS SINGAPORE
Empowering You with Trusted Data

AI for data extraction of unstructured data

- Artificial intelligence (AI) solution uses advanced semantic and reasoning algorithms to automatically **identify**, **extract**, **cleanse** and **validate** the required information from financial statements.

- The AI model is developed based on training datasets (i.e. a small set of financial statements) and deployed for data extraction from a large volume of financial statements.

AI for data extraction of unstructured data

- A **Proof-of-Concept (PoC)** was conducted to assess the ability and accuracy of the AI solution in analyzing and extracting the required information.

- DOS is currently working with the awarded vendor on the development of the system, and Phase 1 is expected to be rolled out by the end of the year.

- The new AI capability enables DOS to improve **operational processes** in data collection and processing and ensure that **more detailed data are available** for analysis.

AI for data extraction of unstructured data

Examples of AI-extracted data:

- Detailed assets information only available in unstructured format in the notes of the financial statements can be extracted to support more in-depth analysis on firms' asset structure and investment.

- More detailed shareholding information can be extracted from the financial statements to supplement existing machine-readable data for ownership analysis.

Example 1: Detailed Assets Information

	Note	Group 2020 $'000	Group 2019 $'000	Company 2020 $'000	Company 2019 $'000
Non-current assets					
Property, plant and equipment	4	1,929	2,165	247	213
Intangible assets	5	769	990	637	773
Investment properties	6	2,730	2,981	–	–
Subsidiaries	7	–	–	86,663	86,163
Other investments	9	18,819	25,096	54	14
Loans, advances, hire purchase and leasing receivables	10	82,332	83,092	75,837	69,368
Deferred tax assets	12	3,692	3,856	–	–
Right-of-use assets	38	2,525	2,839	1,834	2,020
		112,796	121,019	165,272	158,551

Extract value '247000' based on interpretation of column names (i.e. year 2020 and units ('000) and row name (i.e. Property, plant and equipment)

Example 2: Shareholding Information

As at 31 December 2019, the Company's immediate holding company is AB Limited, a company incorporated in the Republic of Singapore. The Company's intermediate holding company is ABO, a company incorporated in Denmark, and the ultimate holding company is AB Foundation, an enterprise foundation registered in Denmark.

Extract name and country of the immediate, intermediate and ultimate companies (highlighted)

Note: Information is sourced from the annual report made available on a firm's corporate website.

Updating the SBR

Machine Learning and Web-based Data to profile firms with internet presence

The information on firms' internet presence serve as a new indicator on firm's characteristics and can be merged with other firm-level data in the SBR to derive new insights or support in-depth studies.

AI for data extraction of unstructured data from financial statements

The unstructured data extracted by the AI can supplement the existing financial information residing in the SBR, for compilation of economic and business indicators.

Learning Points

Machine Learning and Web-based Data to profile firms with internet presence

Big data is increasingly an important data source in additional to traditional data. As data and technologies are evolving and their potential and limitations are researched, new data sources and methodologies would supplement conventional data collection and statistical methodology in the production of official statistics.

AI for data extraction of unstructured data from financial statements

Data exist in multiple forms, either structured or unstructured. With evolving technologies, it is possible to tap on the vast and potentially valuable resource of information and gain access to a much bigger pool of data to either derive new indicators or replace/supplement existing data collection/compilation.

DEPARTMENT OF STATISTICS SINGAPORE
Empowering You with Trusted Data

Thank You

Our Vision

National Statistical Service of Quality, Integrity and Expertise

Our Mission

We Deliver Insightful Statistics and Trusted Statistical Services that Empower Decision Making

www.singstat.gov.sg