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

by Chee Rong Can and Peh Li Lin
Business Statistics Division
Singapore Department of Statistics

Introduction

The Singapore Department of Statistics (DOS) manages the Statistical Business Register (SBR), which serves as the foundational statistical database for the compilation of business and economic statistics.

DOS's SBR provides a comprehensive coverage of the economic units for survey frame production used in sample selections for business surveys, facilitates the compilation of business indicators and supports in-depth analysis for policy insights.

In response to increasingly complex data demands, DOS has undertaken initiatives to transform and acquire innovative and new capabilities across the data value chain. This article shares DOS's experiences in leveraging machine learning (ML) and artificial intelligence (AI) techniques to enhance data availability and update the SBR.

Coverage of the SBR

The SBR covers all entities registered in Singapore. These comprise companies and businesses (including sole proprietorships and partnerships) registered with the Accounting and Corporate Regulatory Authority (ACRA) of Singapore, charities and societies registered with the Ministry of Culture, Community and Youth (MCCY) and the Registry of Societies (ROS) respectively, and other entities registered with their respective registration authorities[1].

Basic identification (e.g., Unique Entity Number (UEN), enterprise name) and enterprise characteristics (e.g., registration date, status) are readily available from the various administrative registration authorities.

Enterprises in the SBR are uniquely identified by their UENs, which are issued to them upon successful registration in Singapore. Enterprises use UENs in their interactions with the Government, such as the application of business licenses and permits, as well as filing of tax returns. It follows that the UEN enables DOS to process and integrate enterprise-level data from various sources efficiently and accurately.

Data Sources and Maintenance of the SBR

Administrative data are the primary sources of the SBR because of their comprehensive coverage of Singapore-registered enterprises. Table 1 presents examples of administrative data that are regularly used to update the SBR.

TABLE 1
EXAMPLES OF ADMINISTRATIVE DATA USED TO UPDATE THE SBR

Source Agency	Administrative Data
Accounting and Corporate Regulatory Authority (ACRA) of Singapore	▪ Identification and Basic Enterprise Information (e.g., UEN, Enterprise Name, Registration Date, Shareholder Information, Industrial Classification) ▪ Financial Information (e.g., Revenue, Profit)
Inland Revenue Authority of Singapore (IRAS)	Financial Information
Ministry of Manpower (MOM) Central Provident Fund Board (CPF)	Employment and Wages
Singapore Customs (CUSTOMS) Enterprise Singapore (ESG)	Merchandise Trade (i.e., Imports, Exports)

[1] More information on the UEN registration authorities can be found here .

Challenges with Using Administrative Data

DOS has been collaborating with data source agencies on the use of administrative data for statistical and analytical purposes as well as the coordination and streamlining of operational processes:

1. *Changes in administrative systems*: For example, agencies would notify DOS in advance of forthcoming data changes such as revisions in administrative filing requirements. This would allow DOS time to assess the potential impact of the data changes and implement necessary measures to minimise disruption to statistical production.

2. *Data consistency and quality*: DOS also provides regular feedback and suggestions on data quality improvement and assists data source agencies in data capability building via knowledge sharing. Through such communication and interactions, DOS and data partner agencies develop a shared understanding and mutual trust to optimise the use of administrative data for statistical and analytical purposes.

3. *Timeliness and coverage*: While administrative data are the primary data sources for the SBR, some may not be timely (e.g., corporate tax filings are only available 1-2 years after an enterprise's financial year ending), readily available in machine readable format or are unavailable. Therefore, data collected through surveys conducted by DOS or other government agencies are also used to supplement the updating of the SBR. Nonetheless, in view of respondent burden, it is not feasible to continually leverage surveys to collect data that are not available from administrative sources.

Experimental Uses of Machine Learning and New Data Sources

In response to these challenges and the need to meet increasingly complex data demands, DOS is building its capabilities to tap on new opportunities including big data and advances in technology such as AI and ML.

Big data offer potential advantages of higher data frequency, greater granularity as well as lower data collection cost. Big data also provide additional information not available in the existing administrative data and surveys.

Using AI and ML, DOS can extract and process such new data sources efficiently. The next section showcases two pilot projects that used AI and ML techniques to improve data availability in the SBR.

(I) Web-Based Data Sources to Profile Enterprises with Internet Presence

Over the last few decades, the internet provided growth opportunities for many enterprises. However, information on whether enterprises have an internet presence is not available from administrative sources.

Hence, DOS undertook a pilot project to text mine web-based data and use supervised ML to study how enterprises in Singapore utilise their corporate websites. In this project, enterprises are broadly classified into three major categories according to their internet presence and corresponding usage (Table 2).

Table 2
CATEGORISATION OF ENTERPRISES ACCORDING TO THEIR INTERNET PRESENCE

Internet Category	Definition
A	Enterprises without websites
B	Enterprises with websites/ online presence but do not generate income directly from their websites Example Websites with information on products/ services
C	Enterprises which generate income directly from their websites Example Online retail stores where customers can place orders directly

The target population was first identified through the enterprise information available in the SBR. The Uniform Resource Locators (URLs) or the website addresses of these enterprises were then gathered from various sources such as surveys, administrative data, online directory, and the Singapore Network Information Centre (SGNIC) which is the domain registry of website addresses ending with ".sg".

These information on enterprises' URLs were subsequently merged with the target population to generate a web crawling list. Web scraping technique was then applied to extract selected features from the website addresses on the crawling list.

To minimise burden on the websites, DOS only scrapped each domain once a year and included idle time when scrapping the pages within each domain.

Some examples of the extracted features include whether the website displays information on products and services or any online shopping facility. With the extracted features, a supervised ML classifier algorithm was applied to classify the enterprises into different categories of internet presence (Figure 1).

The indicator on enterprises' internet presence can be integrated with enterprise characteristics (e.g., economic activity, enterprise's age) available in the SBR to derive new insights for further analysis.

This pilot project demonstrated the feasibility of text mining web-based data and using ML to derive new indicators, which can be used to enhance the information in the SBR.

Compared to the traditional way of conducting a survey to obtain the data, this approach costs less and does not impose burden on survey respondents.

(2) Leveraging AI for Data Extraction of Unstructured Data from Financial Statements

While DOS has been relying on structured data to update the SBR and for statistical compilation, unstructured data from enterprises' financial statements are also a rich source of financial information and new insights.

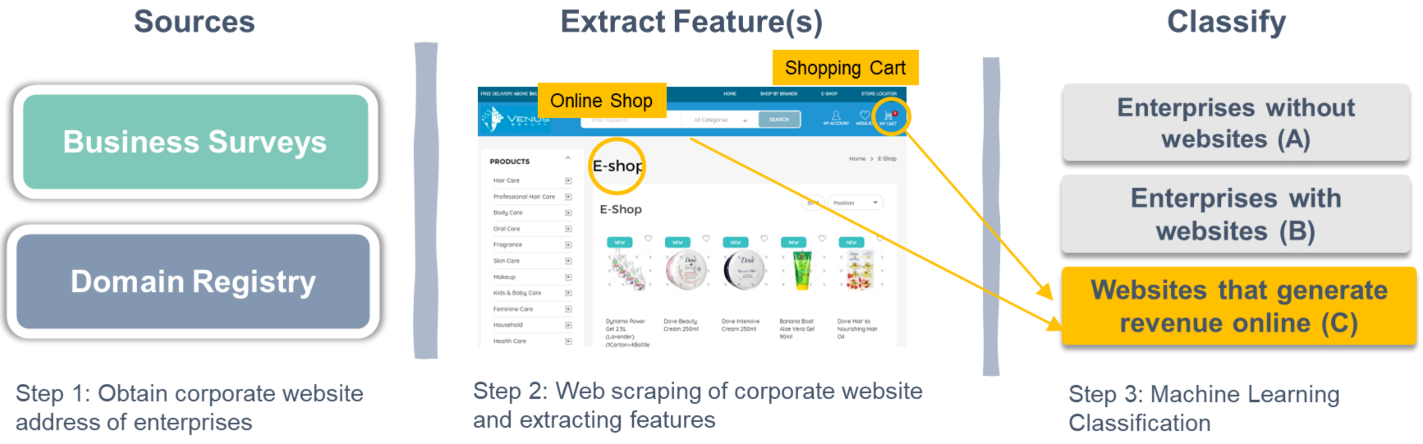
However, considerable manual effort is required to read, analyse and extract the relevant information. This curtails the number of financial statements and data points that can be captured and processed for statistical use.

To overcome this limitation, DOS is developing AI capabilities for data extraction and processing of unstructured data from enterprises' financial statements.

Advanced semantic and reasoning algorithms are used 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.

A Proof-of-Concept (PoC) of the AI solution, co-developed with a commercial AI solution provider,

Figure 1
WEB SCRAPING AND ML
FOR CLASSIFICATION OF ENTERPRISES WITH INTERNET PRESENCE


Sources
Business Surveys
Domain Registry
Step 1: Obtain corporate website address of enterprises
Extract Feature(s)
Shopping Cart
Online Shop
E-shop
E-Shop
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

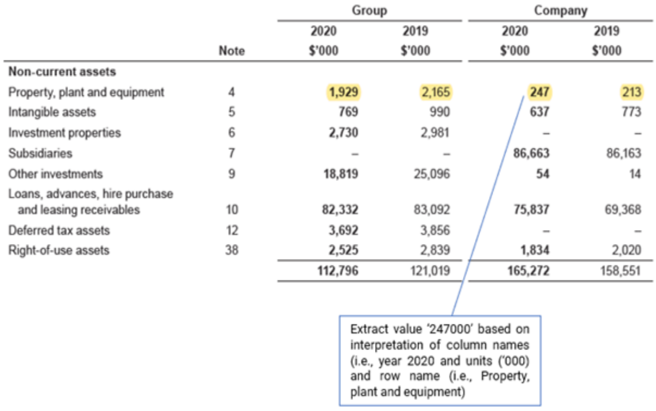
was conducted by DOS to extract data items such as the type of fixed assets, names of overseas subsidiaries and ultimate shareholders from financial statements. The AI solution successfully extracted the required data from the unstructured information in the financial statements with reasonable accuracy.

The information extracted by AI can be used to supplement existing financial information in the SBR for compilation of economic and business indicators. Two examples are presented below:

- Detailed assets information extracted from the notes of the financial statements can be used to support in-depth analysis on enterprises' asset structure and investment (Figure 2).

Figure 2

DETAILED ASSETS INFORMATION[2]



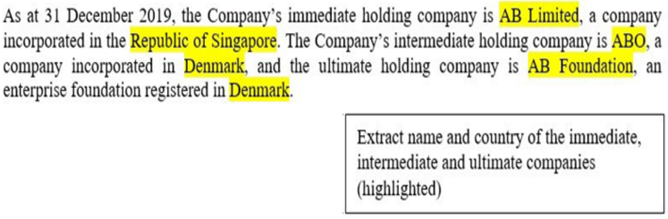
	Note	Group		Company	
		2020 $'000	2019 $'000	2020 $'000	2019 $'000
Non-current assets					
Property, plant and equipment	4	1,929	2,165	247	213
Intangible assets	5	769	990	637	773
Investment properties	6	2,730	2,981	–	–
Subsidiaries	7	–	–	86,663	86,163
Other investments	9	18,819	25,096	54	14
Loans, advances, 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)

- More detailed shareholding information supplement existing machine-readable data available in the SBR, for ownership analysis (Figure 3).

Figure 3

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)

The experience and knowledge gained in the PoC helped DOS plan and scale up the actual implementation of the AI solution that will be rolled out in production by end 2023. The new AI capability enables DOS to improve operational processes in data collection, statistical editing, imputing and estimation, and ensure that more detailed data are available for analyses.

Conclusion

The SBR plays an important role in the data collection, compilation and analysis of economic and business statistics.

While administrative data remain the primary sources for the SBR, new data sources such as big data and unstructured data are becoming increasingly important. Leveraging AI and ML opens up newer data sources for DOS to enhance the information in the SBR.

References

EuroStat (2021)
European Business Statistics Methodological Manual for Statistical Business Registers

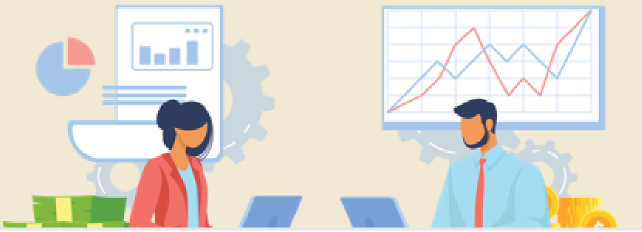
Singapore Department of Statistics (2022)
Using Big Data to Profile Singapore's Internet Economy

Statistics Netherlands (2020)
Measuring the Internet Economy with Big Data

United Nations (2020)
United Nations Guidelines on Statistical Business Registers

United Nations Economic Commission for Europe (2015)
Guidelines on Statistical Business Registers

United Nations Economic Commission for Europe (2021)
Machine Learning for Official Statistics



[2] Information is sourced from the annual report made available on an enterprise's corporate website.