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

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

Chee Rong Can, Peh Li Lin

Business Statistics Division, Singapore Department of Statistics

Abstract

Singapore Department of Statistics (DOS) manages the Statistical Business Register (SBR), which serves as the foundational statistical infrastructure for the compilation of business and economic statistics, by providing comprehensive coverage of the economic units for survey frame production and business demography data. Information on firm characteristics in the SBR is also integrated with other administrative and survey data to compile statistics and indicators on Singapore's enterprise landscape and to support in-depth analytical studies for policy insights.

In response to increasingly complex data demands, DOS has embarked on transformation efforts to acquire innovative and new capabilities across the data value chain. This paper will share DOS' experiences in leveraging machine learning (ML) and artificial intelligence (AI) techniques to enhance data availability in the SBR through two pilot projects. Firstly, DOS explored the use of web-based data sources and supervised machine learning to profile firms with internet presence. Firms with internet presence is used as a new indicator on firm characteristics and allow the profiling of enterprises by type of internet presence. Secondly, DOS is also developing AI capabilities for data extraction and processing of unstructured data from firms' financial statements. The AI-extracted data supplements existing data sources to provide relevant and updated information on firm characteristics in the SBR and to facilitate the development of more timely business indicators.

1 Introduction

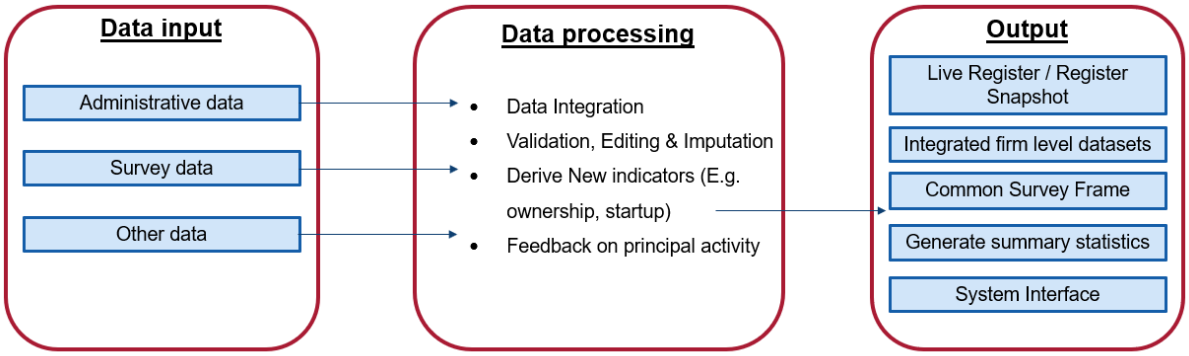
The SBR serves as the foundational statistical infrastructure for the compilation of business and economic statistics and contains key information such as enterprise name, Unique Entity Number (UEN), registration date and industrial classification (SSIC). Basic firm characteristics such as revenue and employee size are also available. The SBR serves as the survey sampling frame for conducting business surveys. The SBR is also used in producing business indicators (e.g. firm's formation and cessation by industry, number of startups) and supporting in-depth analysis of the enterprise landscape.

2 Data Sources for updating of the SBR

The SBR is updated primarily using administrative data, supplemented with statistical survey returns from DOS and Research & Statistics Units (RSUs) in government ministries and statutory boards. Recently, new data sources such as big data are also being explored as additional sources for use in the update of the SBR.

The diagram below depicts the process flow of the integration of various data sources for the maintenance and update of the SBR [Exhibit 1]. The SBR maintains a population of enterprises[1] and establishments[2] in Singapore. Input data are received from multiple data sources at different frequencies. These are processed and integrated in the SBR to produce outputs such as the 'Live Register' comprising up-to-date core information on firms, the 'Common Frame' for survey sampling as well as an integrated business dataset for firm-level data analysis.

Exhibit 1: Process flow of the integration of data sources in the SBR



Firms in the SBR are identified by the Unique Entity Number (UEN), a unique identification number issued upon firms' registration in Singapore. UENs are used by firms in their interactions with the Government, such as the application of business licenses and permits and filing of tax returns. The UEN is available in all administrative and survey data and enables DOS to process and integrate firm-level data efficiently and accurately.

Administrative data are the primary sources for the maintenance of the SBR because of its comprehensive coverage of Singapore-registered firms. Examples of administrative data include business data from the Accounting and Corporate Regulatory Authority (ACRA), tax

[1] Enterprise is defined as a registered business or organisation unit.
[2] Establishment is defined as a business or organisation unit engaged in a single activity and generally operating in a single location.

data from the Inland Revenue Authority of Singapore (IRAS), employment and wages data from the Ministry of Manpower (MOM) and Central Provident Fund Board (CPF) and merchandise trade data from Enterprise Singapore (ESG). These data are received regularly and used to update the various indicators of firm characteristics in the SBR.

DOS has been collaborating with data source agencies on the statistical uses of administrative data as well as the coordination and streamlining of operational processes. For example, the agencies would notify DOS of forthcoming data changes in advance such as changes in administrative filing requirements. This would allow DOS to assess the potential impact of the data changes and implement necessary measures to minimize disruption to statistical production. DOS also provides regular feedback and suggestions on data quality improvement and assists data source agencies in data capability building through knowledge sharing. Through such regular communications and interactions, DOS and data partner agencies have developed a shared understanding and mutual trust to optimise the use of administrative data for statistical and analytical purposes.

Despite the plethora of administrative data, some of these data may not be timely (e.g. corporate tax filings are only available 1-2 year after firm's financial year ending) or not readily available in machine readable format. In addition, data of increasing interest may not be available from administrative sources.

3 Experimental Use of Machine Learning and New Data Sources

In response to these challenges and increasingly complex data demands, DOS has embarked on transformation efforts to acquire innovative and new capabilities across the data value chain. Big data offers potential advantages of higher data frequency, greater granularity as well as lower data collection cost. Big data also provides additional information not available in the existing administrative data and surveys, enabling DOS to derive new indicators to illustrate emerging economic trends. It is also imperative for DOS to leverage on new technology such as machine learning (ML) and artificial intelligence (AI) to process such new data sources and extract useful information for statistical use. The next section will share two pilot projects that leverage machine learning (ML) and artificial intelligence (AI) techniques to enhance data availability in the SBR.

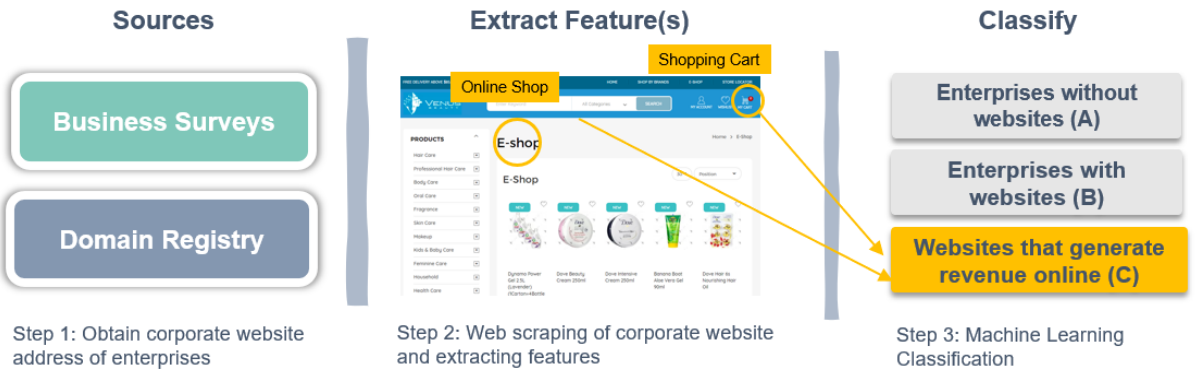
3.1 Web-based data sources to profile firms with internet presence

The internet permeates many aspects of our society and economy, from the way people interact to how companies and businesses operate. Over the last few decades, the internet provided growth and start-up opportunities for many companies. However, information on internet presence is not available from administrative sources. Hence, DOS undertook a pilot project to explore the use of web-based data sources and supervised machine learning to profile firms with internet presence, to better understand how enterprises make use of their corporate website.

The target population was identified through the enterprise information available in the SBR. The Uniform Resource Locators (URLs) or the website addresses of these enterprise were sourced from business surveys conducted by DOS or purchased from the private data provider, Singapore Network Information Centre (SGNIC) which is the domain registry of website addresses ending with ".sg".

These information on enterprises' URLs was subsequently merged back to the target population to generate a web crawling list. Web scraping technique was applied to extract selected features from the website addresses identified in the crawling list. Some examples of the extracted features include whether the website displays information on product and services or whether there is indication of online shopping. With the extracted features, a supervised machine learning classifier algorithm was applied to classify the enterprise into different categories of internet presence [Exhibit 2].

Exhibit 2: Web Scraping and Machine Learning for Classification of Enterprises with Internet Presence



Taking reference from a similar project undertaken by Statistics Netherlands[3], enterprises in this project are broadly classified into three major categories according to their internet presence and corresponding usage [Table 1].

Table 1: Categorisation of enterprises according to their internet presence

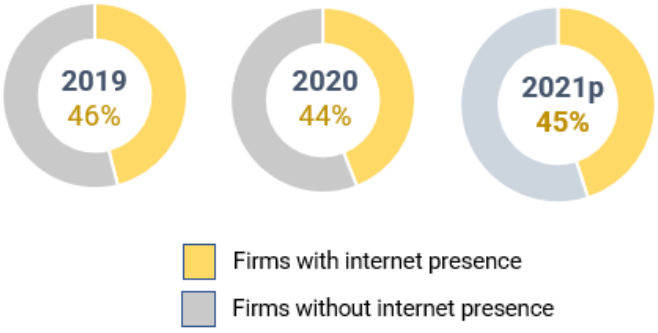
Internet Category	Definition	Examples
A	Enterprises without websites	-
B	Enterprises with websites / online presence but do not generate income	Websites with information on products/services
C	Enterprises which generate income directly from the internet	Online retail stores where customer can place orders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

The indicator on enterprises' internet presence can be integrated with firm characteristics (e.g. economic activity, firm's age) data available in the SBR, to derive new insights for further analysis. Some key findings are summarized as follows:

- 45% of enterprises in Singapore had internet presence in 2021, with the proportion remaining relatively stable for the past three years [Chart 1].

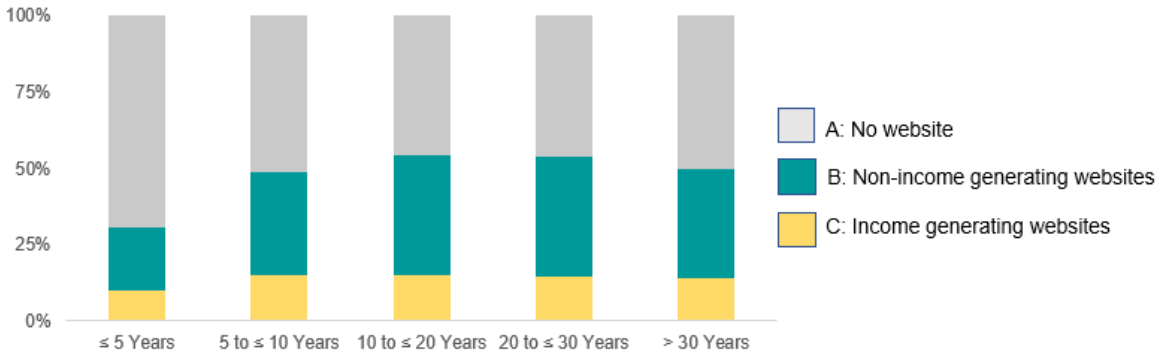
[3] Measuring the internet economy with big data. Netherlands: CBS; 2020.

Chart 1: Number of Enterprises with Internet Presence, 2019 – 2021



- In 2021, majority of enterprises above 10 years old had internet presence, while only one in three young firms (aged 5 years or less) had internet presence [Chart 2].

Chart 2: Share of Enterprises with Internet Presence by Age Group, 2021



This pilot project demonstrated the possibility of deriving new indicators using machine learning and new data sources (i.e. web-based sources) to enhance information in the SBR. These new indicators can also be merged with other firm-level data such as revenue and employment to derive other insights.

3.2 Leveraging on AI for data extraction of unstructured data from financial statements

While DOS has been relying on structured data for updating of the SBR and statistical compilation, there is a rich source of financial information and new insights that can be derived from data that are in unstructured format available in firms' financial statements. To use the unstructured data, considerable manual effort is required to read, analyse and extract the relevant information manually. This limits the number of financial statements and data points that can be captured and processed for statistical use.

DOS is developing AI capabilities for data extraction and processing of unstructured data from firms' financial statements, using 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.

A Proof-of-Concept (PoC) of the AI solution co-developed with a commercial AI solution provider was conducted by DOS to extract about 30 data items such as details on the type of fixed assets, names of overseas subsidiaries and ultimate shareholders from the financial statements, to assess the ability and accuracy of the AI solution in the analysis and extraction of the required information. The solution was assessed to be able to successfully extract 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 firms' asset structure and investment [Exhibit 3].

Exhibit 3: Details Assets Information[4]

		Group		Company	
	Note	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 [Exhibit 4].

Exhibit 4: 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 to plan and scale up the actual implementation of the AI solution that will be rolled out in production in 2023. It enables DOS to extract over 300 data points from over 50,000 financial statements annually. The new AI capability also facilitates improvements in DOS's operational processes, data quality and the compilation of business indicators.

[4] Information is sourced from the annual report made available on a firm's corporate website.

4 Conclusion

While administrative data remain to be the primary sources for maintenance of the SBR, new data sources such as big data and unstructured data are becoming increasingly important. 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. It is now also possible to tap on the vast and potentially valuable resource of information to derive new indicators to meet emerging data demands. DOS will continue our efforts in experimenting the use of advanced techniques and 'non-traditional' data in statistical production. This paper presented the pilot projects which demonstrate DOS's transformation efforts in acquiring new data capabilities to deliver insightful statistics and trusted statistical services that empower decision making.