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
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)

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

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

Extract Feature(s)
- Step 1: Obtain corporate website address of enterprises
- 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)

Online Shop
Shopping Cart

Step 3: Machine Learning Classification
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.
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

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

<table>
<thead>
<tr>
<th>Note</th>
<th>2020</th>
<th>2019</th>
<th>2020</th>
<th>2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-current assets</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Property, plant and equipment</td>
<td>4</td>
<td>1,929</td>
<td>2,365</td>
<td>247</td>
</tr>
<tr>
<td>Intangible assets</td>
<td>5</td>
<td>856</td>
<td>990</td>
<td>637</td>
</tr>
<tr>
<td>Investment properties</td>
<td>6</td>
<td>2,730</td>
<td>2,381</td>
<td>–</td>
</tr>
<tr>
<td>Subsidiaries</td>
<td>7</td>
<td>–</td>
<td>–</td>
<td>64,643</td>
</tr>
<tr>
<td>Other investments</td>
<td>9</td>
<td>18,819</td>
<td>25,096</td>
<td>54</td>
</tr>
<tr>
<td>Loans, advances, hire purchase and leasing receivables</td>
<td>10</td>
<td>82,332</td>
<td>63,092</td>
<td>75,837</td>
</tr>
<tr>
<td>Deferred tax assets</td>
<td>12</td>
<td>3,992</td>
<td>3,956</td>
<td>–</td>
</tr>
<tr>
<td>Right-of-use assets</td>
<td>38</td>
<td>2,525</td>
<td>2,839</td>
<td>1,834</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>112,796</strong></td>
<td><strong>121,019</strong></td>
<td><strong>105,272</strong></td>
<td><strong>158,551</strong></td>
</tr>
</tbody>
</table>

**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 addition 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.
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