Use of Administrative Data and New Data Sources in Industry Statistics

by Wong Yan Ru and Neo Soo Khee
Business Statistics Division
Singapore Department of Statistics

Introduction

The Singapore Department of Statistics (DOS) produces short-term and annual estimates of industry statistics for the services sector [1]. These statistics are used by policymakers, researchers, and the business community, to study the structure and performance of various activities in the services sector.

Surveys have traditionally been the primary method for collecting data to produce industry statistics. DOS is moving away from conventional business surveys by leveraging administrative and non-traditional data sources. Using artificial intelligence (AI) and machine learning (ML), DOS can extract and process new data sources more efficiently. This reduces respondent burden while enhancing the ability to collect and compile more comprehensive data on businesses.

This article shares DOS's efforts in using administrative and non-traditional data sources to streamline data collection and improve the data compilation process for both short-term and annual industry estimates. It also covers DOS's experience in producing new data products based on administrative data and new data sources, as well as the use of AI and ML to utilise new data sources.

Administrative Data

Administrative data are collected by organisations for administrative purposes (e.g., regulatory, accounting, or commercial). For the compilation of industry statistics, the administrative data relating to revenue, profit, and wages acquired from different government agencies in Singapore are listed below.

1 Regulatory Authority of Business Registration & Financial Reporting:
Accounting and Corporate Regulatory Authority (ACRA) of Singapore

Financial Information such as Revenue and Profit

Identification and demographic information of businesses such as:

- Unique Entity Number (UEN)
- Business name and address
- Registration date
- Shareholder information
- Industrial Classification

2 Manpower Authority:
Ministry of Manpower (MOM) & Central Provident Fund Board (CPF) [2]

Employment and Wages

3 National Tax Authority:
Inland Revenue Authority of Singapore (IRAS)

Financial Information such as Revenue and Profit from Goods & Services Tax (GST) [3] and Corporate Income Tax.

Stamp duty on tenancy agreements (e.g., UEN, property address, and tenancy start and end dates)

4 Housing Authority:
Housing and Development Board (HDB)

Rental of commercial properties under HDB management such as:

- UEN
- Property address
- Start and end dates of tenancy

[1] The Services Sector includes enterprises engaged in the Wholesale Trade, Retail Trade, Transportation & Storage, Accommodation & Food Services, Information & Communications, Real Estate, Professional Services, Administrative & Support Services and Community & Personal Services. Enterprises engaged in Finance & Insurance Services and Public Administration Activities are excluded. For more information on industry statistics of the services sector, please refer to the **SingStat Website**.

[2] The Central Provident Fund (CPF) is a mandatory social security savings scheme in Singapore, funded by contributions from employers and employees. Employers of local workers (Citizens and Permanent Residents of Singapore) are required to declare wage information and pay employees' CPF contributions monthly. Enterprises employing foreign workers must apply for relevant work passes from the MOM and inform MOM of changes such as revisions in employee's salary and updates to personal particulars.

[3] GST is a broad-based consumption tax levied on the purchases of goods and services in Singapore. Enterprises with annual taxable turnover exceeding S$1 million must register for GST, while those under S$1 million can voluntarily register for GST. GST data includes timely information on enterprises' revenue, as enterprises are required to file for GST on a quarterly basis, within one month after the end of the accounting period. GST is also known as the Value-Added Tax (VAT) in other countries. For more information, please refer to the **IRAS website**.

As administrative data may differ in coverage and/ or concepts from statistical concepts and definitions, additional work was undertaken to study the differences before they can be incorporated into the production of industry statistics for the services sector.

Short Term Indicators

DOS compiles the monthly Retail Sales Index (RSI), Food & Beverage Services Index (FSI), quarterly Wholesale Trade Index (WTI), and the Business Receipts Index (QBRI). These indices measure the short-term performance of services industries.

These indices are mainly compiled using revenue data collected via monthly or quarterly surveys. Administrative GST data and non-traditional data [4] such as Gross Merchandise Value (GMV) from third-party marketplaces are used to supplement the compilation.

For the quarterly compilation of WTI and QBRI, smaller enterprises are not surveyed as their GST data is sufficient to be used for estimation. Around 80% of the enterprises in the scope of WTI and QBRI are estimated using GST data currently. Larger enterprises will continue to be surveyed in the Quarterly Survey of Services due to differences in coverage and/ or definitions. For example, when reporting for GST, some enterprises may opt to report group level data instead of enterprise level data. Due to the differences in the definition of revenue, government subventions [5] for public and non-profit institutions required for the QBRI compilation are not included in GST revenue. Additional data items such as related party transactions and operating expenditure that are not in GST are collected via surveys.

The monthly compilation of the RSI and FSI includes data on the online proportions of retail and food & beverage (F&B) sales. Data from the monthly surveys are supplemented with the GMV data collected from third-party marketplaces. This ensures a more comprehensive coverage of smaller enterprises, which are more likely to sell their products through third-party marketplaces instead of setting up a website to do so.

Annual Industry Survey

The Annual Industry Survey (AIS) collects information to analyse the structure and performance of enterprises in the Services Sector. The data are used to compile the national accounts, input-output tables, and estimates on indicators such as annual Operating Revenue and Operating Expenditure. These data also support related studies by other economic agencies such as the Ministry of Trade and Industry and Enterprise Singapore.

Transitioning to Register-based Approach

Survey Approach

Previously, the AIS was entirely based on a survey approach. All large enterprises were sampled with certainty, while medium and smaller-sized enterprises were sampled using simple random sampling without replacement. Non-sampled enterprises were represented by sampled enterprises via sampling weights for the compilation of industry statistics.

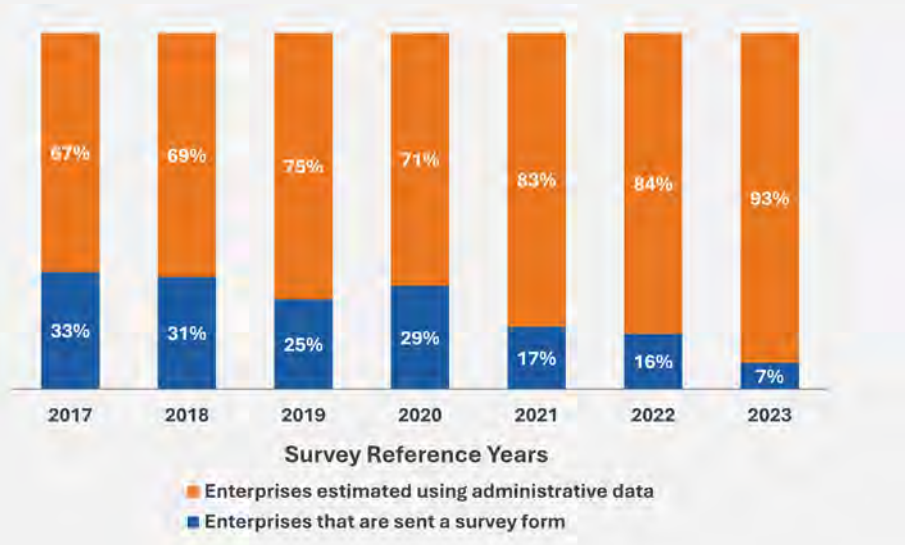
Register-based Approach

DOS is progressively adopting a register-based approach for the AIS where all enterprises are included and estimated using administrative data, compared to relying solely on survey returns. The transition begins with industries where administrative data are assessed to align with statistical concepts and has comprehensive coverage. Larger enterprises are still surveyed to collect detailed breakdowns of their revenue and expenditure items [6] that are not available in administrative data.

By adopting the register-based approach for selected industries, the overall sampling variability is reduced. This in turn improves the quality of the industry estimates developed from the information collected and facilitates data compilation with more granular breakdown. To date, DOS has implemented the register-based approach for around 50% of the services industries in the AIS.

As a result, while the overall number of enterprises covered in the AIS has increased in line with the population size of services enterprises in Singapore over the years, the proportion of enterprises surveyed decreased from 33% in 2017 to 7% in 2023 (Figure 1).

Figure 1: Breakdown of Enterprises by Data Collection Source for AIS, 2017-2023



[4] Non-traditional data refer to information that is not typically captured or analysed using conventional methods or sources.

[5] Government subvention is a revenue item for public and non-profit institutions. It is excluded from the calculation of taxable supplies.

[6] Examples of detailed revenue and expenditure items include freight charges, accounting, auditing and book-keeping fees, legal fees, consultancy fees, etc.

Use of XBRL Data and Machine Learning

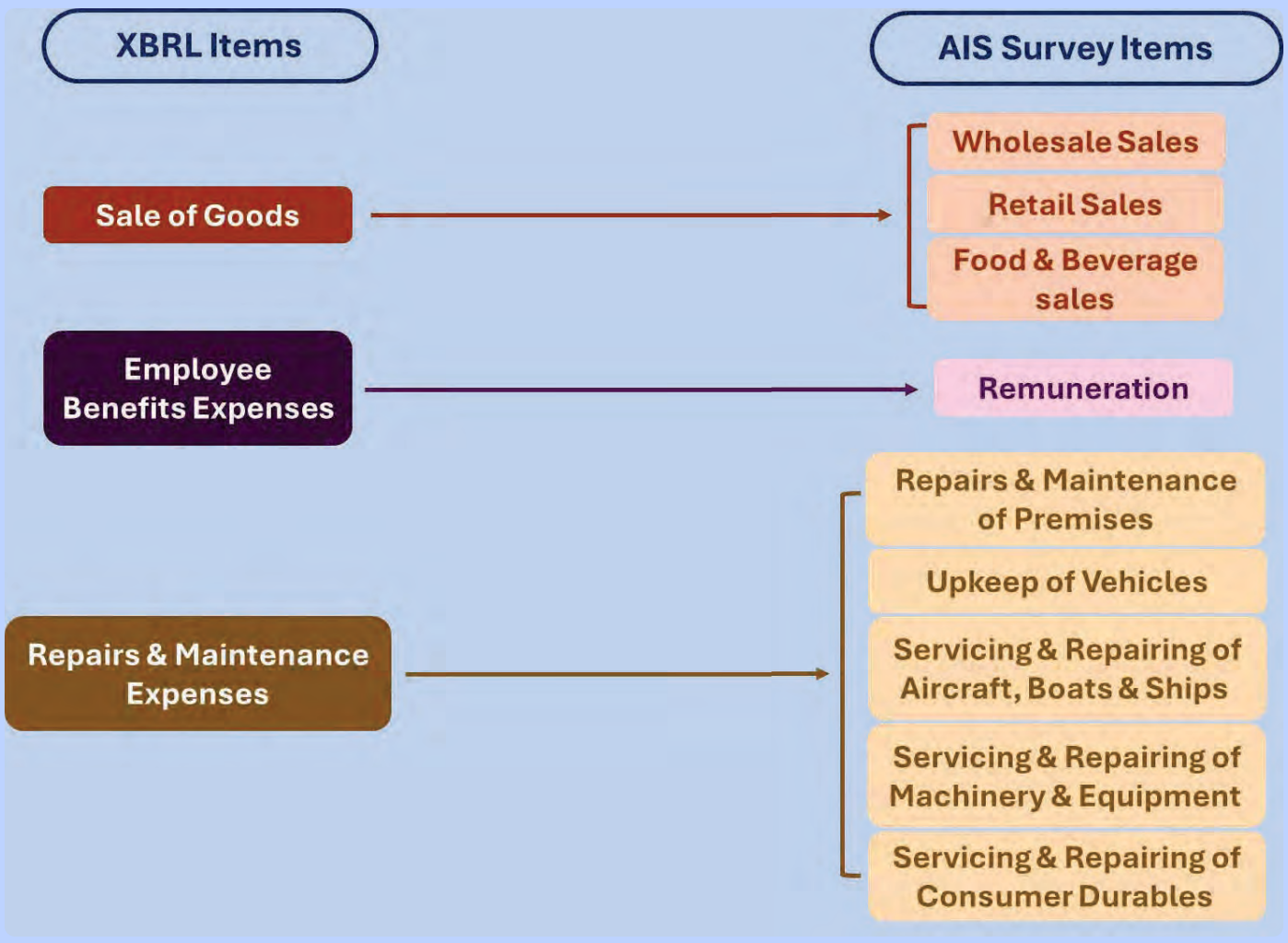
As most administrative data sources only include topline income or expense items at the enterprise level, additional data processing is required when administrative data is used to compile industry statistics. This involves splitting the enterprise-level administrative data to the establishment-level based on the industry proportion, mapping the administrative data items to the relevant income and expense items covered in AIS, and appropriately proportioning the less granular administrative data amongst these survey items.

An example of the data processing steps is illustrated using ACRA's eXtensible Business Reporting Language (XBRL) data [7]. The XBRL format provides a range of detailed income and expense items, such as revenue, profit, sale of goods, rendering of services, employee benefits expenses, repairs & maintenance expenses, and depreciation. Topline items, such as revenue and profits, are mandatory for all enterprises to report when filing financial statements in the XBRL format.

Some of the XBRL income and expense items are non-mandatory and enterprises may opt not to provide these data. In order to fill in the missing information, DOS implemented a random forest ML algorithm to predict their values. For each industry, a random forest algorithm is trained for each non-mandatory data item to predict a value based on the responses from other enterprises that have reported the data and its relationship with other data items.

After running the random forest model on the XBRL data items, these items are mapped as closely as possible to the AIS data items before using them to estimate the values (Figure 2). Some XBRL data items are mapped to more than one AIS item and require further proportioning based on the enterprises' industry classification and historical proportion. For instance, the XBRL income item '*sale of goods'* is mapped to three AIS items, i.e., wholesale sales, retail sales, and F&B sales. If the enterprise is a wholesaler, based on historical industry proportion, '*sale of goods'* could be proportioned into wholesale sales (99%) and retail sales (1%).

Figure 2: Example of Mapping XBRL items to AIS items



[7] Refers to data from financial statements filed in the XBRL format. XBRL is a language for electronic communication of business and financial data worldwide. For more information, please refer to the **ACRA website**.

New Data Products

DOS has produced new data products by leveraging administrative data and new data sources and is exploring the use of AI and ML algorithms to further utilise these sources.

Early Triggering with Employment and GST data

The Early Triggering project detects entries and exits of large enterprises or enterprises with significant changes in revenue and/ or employment in a timely manner. Higher-frequency administrative data on employment from MOM and CPFB, along with data on revenue from enterprises' GST filings with IRAS, are used to identify enterprises with changes in employment and/ or revenue exceeding sector-specific thresholds. Information of these identified enterprises is then shared with the relevant subject domain teams in DOS for incorporation into business surveys and the compilation of economic indicators.

Business Activity by Geographic Area

DOS has developed experimental estimates on retail and F&B activities by geographic area.

The estimates for retail and F&B activities are primarily based on enterprises' operating addresses filed with ACRA and the Singapore Food Agency respectively. The estimates for retail activity are further supplemented with data on rental of commercial properties from IRAS and HDB.

These data are integrated with the enterprises' characteristics (e.g., industry classification) before plotting onto a map within a dashboard [8], which provides insights on the distribution of retail and F&B activities across Singapore.

Weekly Economic Brief

DOS has access to a curation of business news from major media outlets based on keywords covering a range of topics such as business performance, mergers and acquisitions, expansion, and restructuring plans of enterprises. The selected news articles are categorised into pre-defined categories for easy reference.

Indicator on Enterprise's Internet Presence and Other Enterprise Characteristics

DOS text mines web-based data to study how enterprises in Singapore use their corporate websites. Keywords such as 'Shop' or 'Cart' are first extracted from these websites. Supervised ML is applied to categorise the websites into different categories (i.e., enterprises with websites that directly generate revenue, or those with websites but do not generate revenue from them) and derive an indicator on enterprises' internet presence. This indicator is then integrated with enterprises' characteristics from the Statistical Business Register to derive new insights for further analyses. For example, the identification of enterprises with corporate websites are early indications of business activities, and they will be included in the sampling frame for business surveys.

Moving forward, DOS is exploring the use of text mining and ML to obtain other enterprise characteristics like green enterprises, innovative enterprises.

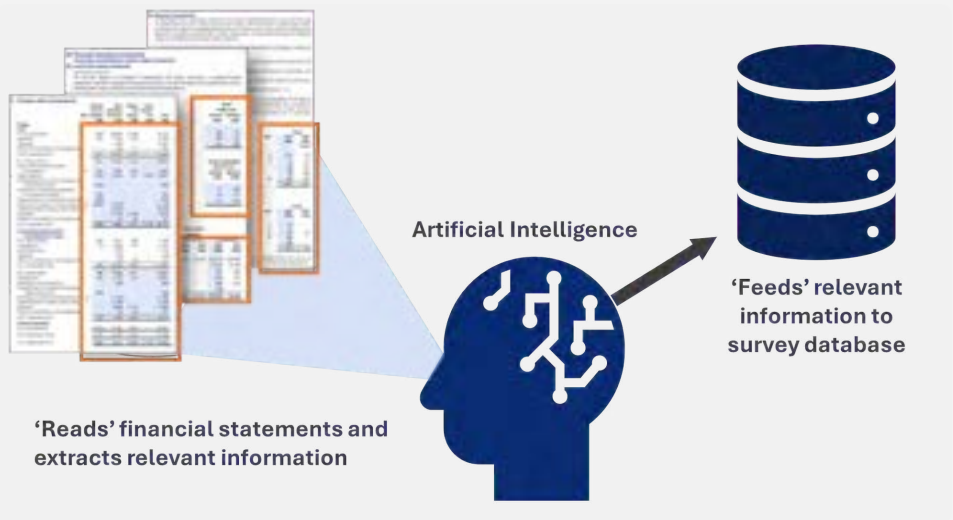
AI on Unstructured Financial Statements

Company financial statements submitted to ACRA or IRAS contain vast information but are presented in an unstructured format and varies significantly across companies. Hence, considerable manual effort is required to extract and interpret relevant business information. For example, one has to manually read, locate, then extract the required information from various sections of the financial statements.

To reduce manual effort, DOS launched a pilot project which uses AI to directly extract detailed information from unstructured financial statements to support the compilation of industry statistics for the services sector (Figure 3). This automates the extraction and interpretation of the unstructured information, which improves operational efficiency and supports the compilation of more comprehensive industry statistics.

DOS is still fine-tuning the algorithm for its eventual use in supporting data compilation. When fully implemented, this will further improve the quality and quantity of administrative data available for data compilation.

Figure 3: Illustration of Using Artificial Intelligence on Financial Statements



Conclusion

While surveys remain an important data source to obtain details not available from administrative sources, DOS has increasingly leveraged and optimised administrative data and non-traditional data sources to improve industry statistics and produce new data products. With the growing availability of administrative data coupled with advancements in data extraction and processing technologies, DOS will continue to explore the use of new data sources to produce relevant statistics for increasingly sophisticated users.

[8] To view the **dashboard** on Business Activity by Geographic Area, make the following selections:
a) For either the Retail or F&B industry, select 'Know My Industry'
b) Within the dashboard, select a Detailed Industry
c) Select 'How is my industry performing?' followed by 'How many firms are engaged in similar activities across Singapore?'

References

- Wallgren, A., & Wallgren, B. (2014). *Register-based Statistics: Statistical Methods for Administrative Data*. John Wiley & Sons.
- EuroStat (2021): **European Statistical System and handbook for quality and metadata reports**
- Singapore Department of Statistics (2022): **Using Big Data to Profile Singapore's Internet Economy**
- Singapore Department of Statistics (2023): **Experimental Use of New Data Sources for Prompt Identification of Changes in Firms' Status**