Use of AI-Enhanced OCR and Machine Learning for Data Processing in the Household Expenditure Survey 2023

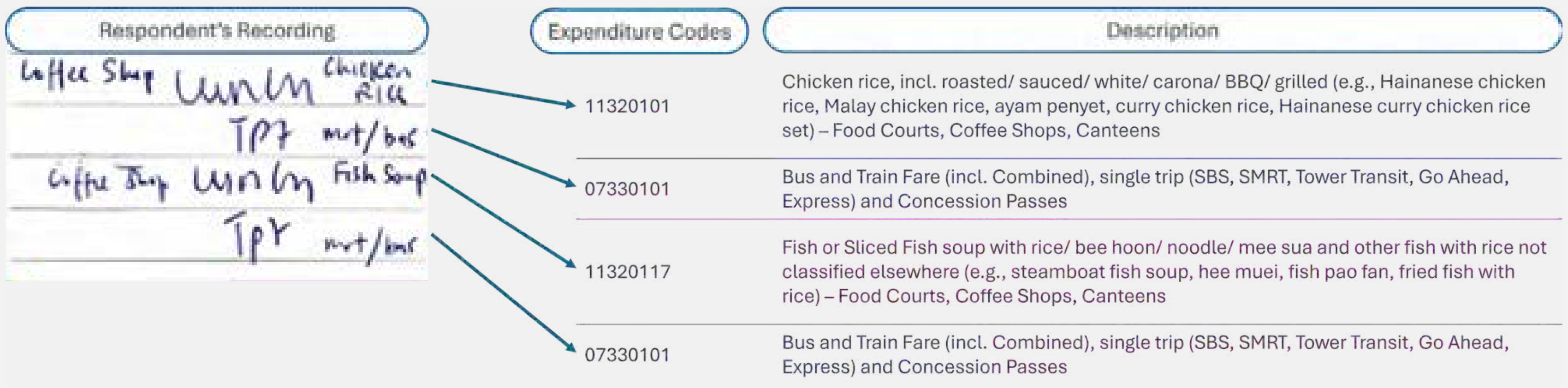
by Boon Kok Ann and Cheng Wan Hsien
Household Surveys and Expenditure Division
Singapore Department of Statistics

Introduction

The Singapore Department of Statistics (DOS) conducts the Household Expenditure Survey (HES) every five years, since 1972/ 73, to collect detailed information on households' expenditure, socio-economic characteristics and ownership of consumer durables. It is carried out over a one-year period to cover different festive and seasonal expenditure of households. Expenditure data collected include day-to-day expenses such as food, groceries, and transport; regular expenditure such as utilities and telecommunication subscription services; and ad-hoc big-ticket expenditure like the purchase of cars and household durables.

Data processing for the HES has traditionally been labour-intensive and time-consuming, requiring extensive manual checks, data entry, and coding. While respondents can submit their returns electronically since the 2017/ 18 survey, many prefer to submit handwritten returns in hardcopy booklets [1] or provide receipts of their regular and day-to-day expenses. In past surveys, data processing clerks manually entered the amount for each expenditure item into the system, then assigned an expenditure code [2] to each expenditure item (Figure 1).

Figure 1: Assigning Expenditure Codes to Expenditure Items



Expenditure Codes	Description
11320101	Chicken rice, incl. roasted/ sauced/ white/ carona/ BBQ/ grilled (e.g., Hainanese chicken rice, Malay chicken rice, ayam penyet, curry chicken rice, Hainanese curry chicken rice set) – Food Courts, Coffee Shops, Canteens
07330101	Bus and Train Fare (incl. Combined), single trip (SBS, SMRT, Tower Transit, Go Ahead, Express) and Concession Passes
11320117	Fish or Sliced Fish soup with rice/ bee hoon/ noodle/ mee sua and other fish with rice not classified elsewhere (e.g., steamboat fish soup, hee muei, fish pao fan, fried fish with rice) – Food Courts, Coffee Shops, Canteens
07330101	Bus and Train Fare (incl. Combined), single trip (SBS, SMRT, Tower Transit, Go Ahead, Express) and Concession Passes

System Redesign and Automation Initiatives

In the HES 2023, the data processing workflow was redesigned, by leveraging AI-enhanced optical character recognition (OCR) and ML modelling techniques, to reduce time spent on manual data entry and coding. For non-electronic returns, hardcopy booklets and receipts were scanned, and the images were processed by the OCR software to extract textual data. This information includes descriptions of expenditure items, dollar amount, payment indicators and date of recording. After verifying and amending any inaccuracies in the extracted information, the data was automatically sent to the data processing system in a machine-readable form.

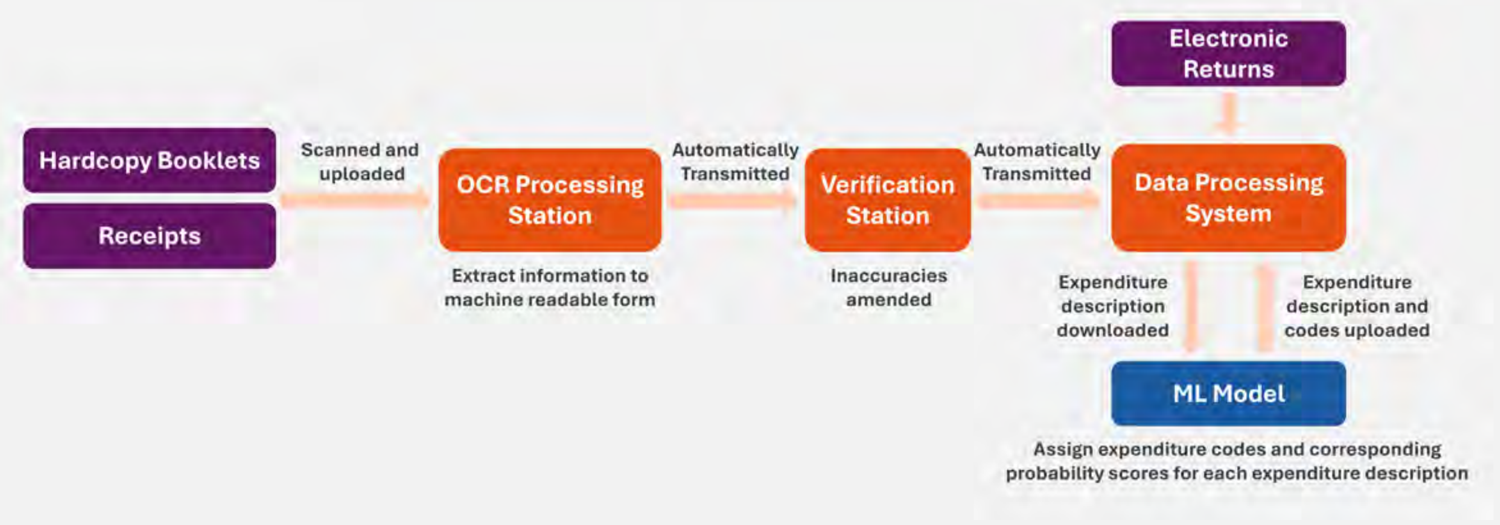
Together with electronic returns, expenditure descriptions were downloaded weekly from the data processing system and passed through an ML model. The model assigns expenditure codes and corresponding probability scores [3] to each expenditure description. If the probability score is above a predefined threshold, the expenditure description and code will be uploaded back into the data processing system (Figure 2). Those with a score below the threshold will not be automatically assigned a code and will require data processing clerks to manually input a code.

[1] In HES 2023, about 85% of respondents submitted some form of handwritten return.

[2] Expenditure codes are based on Singapore Standard Classification of Individual Consumption According to Purpose (S-COICOP).

[3] The probability score refers to the ML model's prediction of the likelihood that a particular expenditure code is the correct code for the description. For e.g., 'bus & train trip' could have a probability score of 90% to be coded as '07330101 - Bus and Train Fare (incl. Combined)' and a score of 8% to be coded '07320101- Bus/Coach fares'.

Figure 2: Redesigned Process for Assigning Expenditure Codes in the HES 2023


Hardcopy Booklets
Receipts
Scanned and uploaded
OCR Processing Station
Extract information to machine readable form
Automatically Transmitted
Verification Station
Inaccuracies amended
Automatically Transmitted
Electronic Returns
Data Processing System
Expenditure description downloaded
Expenditure description and codes uploaded
ML Model
Assign expenditure codes and corresponding probability scores for each expenditure description

Replacing Data Entry with AI-Enhanced OCR

90%

Improved OCR Accuracy for printed text

For faded receipts and poor handwriting, the accuracy rate was lower.

1 Recognising Handwritten Returns

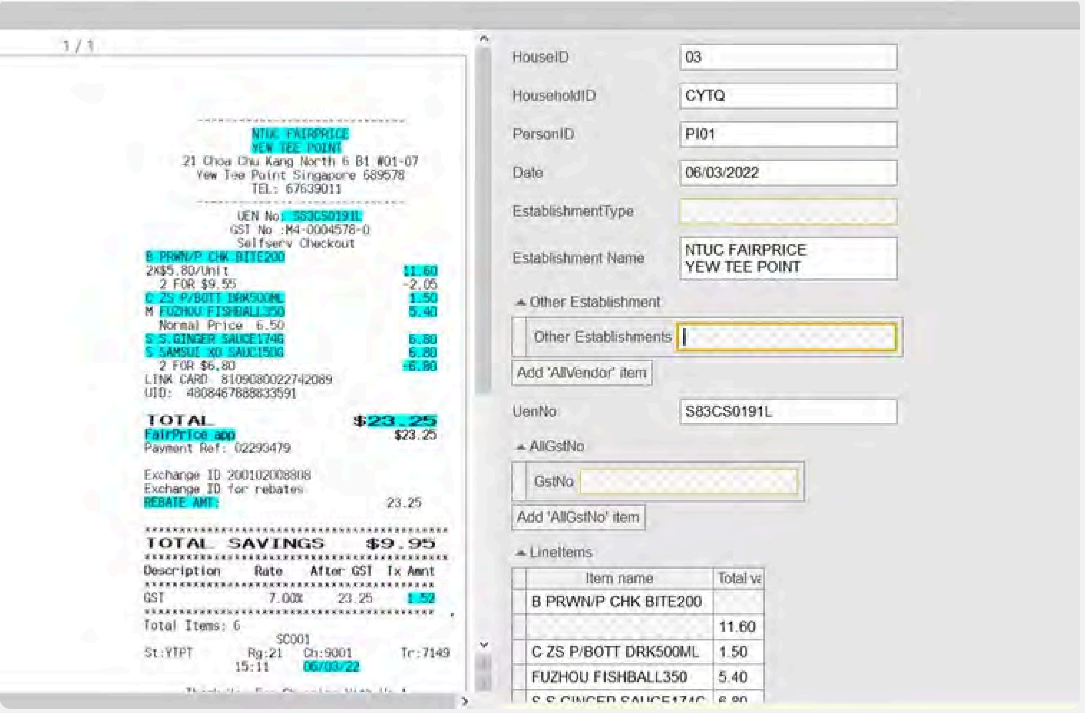
The AI-enhanced OCR software was pre-trained to identify on various handwriting styles, to better interpret respondents' handwritten returns.

2 Recognising Fields in Receipts

Given the different layouts of receipts from various establishments, AI was used to identify fields containing the amount, description, establishment name, payment mode, and other relevant information. The AI model was pretrained on sample receipts and keywords to improve the recognition prior to the data collection for the HES 2023. For example, words such as 'pte ltd' is associated with establishment names, while 'VISA' and 'MASTER' are associated with payment modes.

To reduce the risk of inclusion of such inaccurate inputs, data processing clerks used a verification software to compare the scanned image with the extracted information and correct any errors (Figure 3).

Figure 3: Screenshot of a Scanned Receipt in a Verification Software



1 / 1

```
              NTUC FAIRPRICE
              YEW TEE POINT
      21 Choa Chu Kang North 6 B1 #01-07
       Yew Tee Point Singapore 689578
             TEL: 67639011

            UEN No: S83CS0191L
           GST No :M4-0004578-0
            Selfserv Checkout
B PRWN/P CHK BITE200
2X$5.80/Unit                          11.60
  2 FOR $9.55                         -2.05
C ZS P/BOTT DRK500ML                   1.50
M FUZHOU FISHBALL350                   5.40
  Normal Price  6.50
S S.GINGER SAUCE174G                   6.80
S SAMSUI XO SAUC150G                   6.80
  2 FOR $6.80                         -6.80
LINK CARD  8109080022742089
UID:  4808467888833591

TOTAL                              $23.25
FairPrice app                      $23.25
Payment Ref: 02293479

Exchange ID 200102008808
Exchange ID for rebates
REBATE AMT:                         23.25

*****************************************
TOTAL  SAVINGS                    $9.95
*****************************************
Description     Rate    After GST  Tx Amnt
*****************************************
GST             7.00%    23.25      1.52
*****************************************
Total Items: 6
                 SC001
St:YTPT      Rg:21    Ch:9001       Tr:7149
            15:11    06/03/22
```

HouseID: 03

HouseholdID: CYTQ

PersonID: PI01

Date: 06/03/2022

EstablishmentType:

Establishment Name: NTUC FAIRPRICE
YEW TEE POINT

▲ Other Establishment

Other Establishments:

Add 'AllVendor' item

UenNo: S83CS0191L

▲ AllGstNo

GstNo:

Add 'AllGstNo' item

▲ LineItems

Item name	Total va
B PRWN/P CHK BITE200	
	11.60
C ZS P/BOTT DRK500ML	1.50
FUZHOU FISHBALL350	5.40
S S GINGER SAUCE174G	6.80

In the previous HES, the actual expenditure descriptions were not entered into the data processing system due to high resource demands of accurately capturing them. With the use of AI-enhanced OCR, unstructured textual data from booklets and receipts could be captured efficiently, allowing actual expenditure descriptions of expenditure items to be captured in the HES system for HES 2023.

The record of actual expenditure descriptions was very useful for data processing, as the expenditure items may need to be revisited when checking for consistency and accuracy of the data collected from respondents. Manhours were saved as the new process facilitated the review of summarised textual data which contained the expenditure descriptions and codes for the whole household. Whereas in the previous HES, checking involved navigating to stored images of booklets and receipts to view expenditure descriptions, which was time-consuming.

Automating Expenditure Coding with Machine Learning

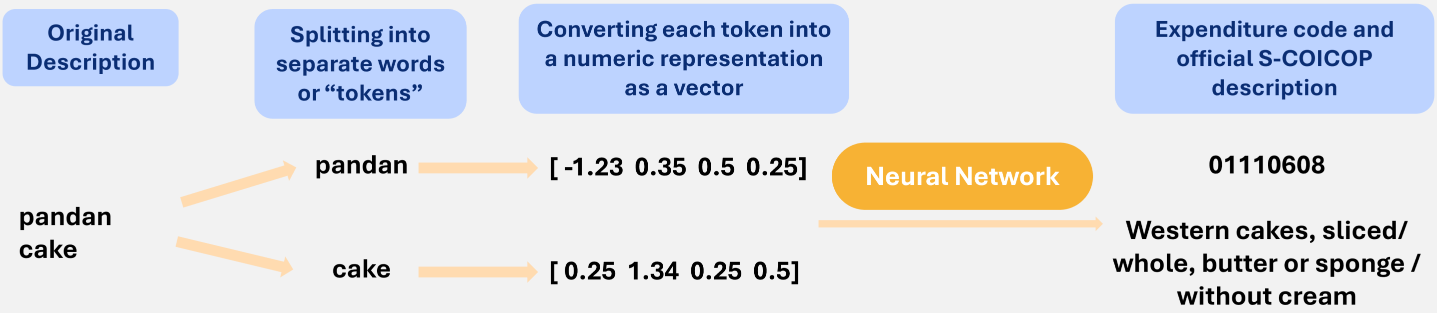
With the expenditure descriptions extracted via OCR, it became possible to apply ML models to automate the assignment of expenditure code to each expenditure item.

To improve the accuracy of the ML models, DOS developed in-house Python scripts to pre-process textual data into a standardised form that can be meaningfully tokenised. Pre-processing involves removing punctuation marks, special characters, unnecessary whitespaces, sizes and weights (e.g., Kg, XXL), and stop words; correcting typographical errors; converting acronyms (e.g., CS = coffee shop, FC = food court); and standardising all characters to lowercase.

After the textual data has been pre-processed, the Recurrent Neural Network (RNN) model is used to assign expenditure codes to each expenditure item. Different methodologies were explored, such as using natural language processing and cosine similarity to perform the coding, trained on data from past HES and the expenditure code dictionary. The RNN model was evaluated to be the best-performing model in terms of accuracy at the most detailed expenditure code level, and was designed for interpreting sequential or temporal information (e.g., text, time series, audio), compared to other neural network models such as Convolutional.

Figure 4 illustrates how the neural network takes in the vector representations of the words as inputs which effectively capture all the information in the sequence of words. The neural network then generates a probability score indicating the likelihood of the expenditure code being the correct code for the description.

Figure 4: Assignment of Expenditure Item to Expenditure Code with Neural Network


Original Description
Splitting into separate words or "tokens"
Converting each token into a numeric representation as a vector
Expenditure code and official S-COICOP description
pandan cake
pandan
cake
[-1.23 0.35 0.5 0.25]
[0.25 1.34 0.25 0.5]
Neural Network
01110608
Western cakes, sliced/ whole, butter or sponge / without cream

Expenditure Code Assignment Process Based on Probability Scores Generated

Automatic Code Assignment

When the probability score of the top-predicted code is very high, the entry will be automatically assigned an expenditure code. The threshold for the score was set to obtain an accuracy rate of 90% for the assigned code.

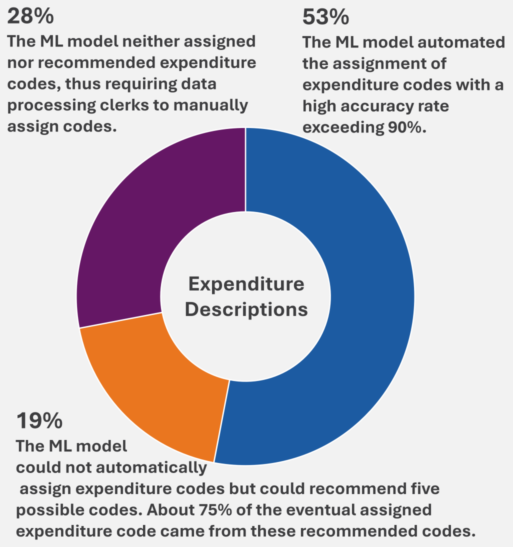
Code Recommendation

When the probability scores of the predicted codes is relatively high, the model will recommend five possible codes with the probability scores. Data processing clerks can either select one of the recommendations or choose a code outside the recommended. The threshold for the score was set such that for the majority of the records, the correct code is within the five recommended codes.

Manual Coding

When the score is low, data processing clerks will manually assign a code.

Effectiveness of ML Model and Conclusion



The RNN model was utilised from the start of the data processing operations and refined with subsequent batches of data. This refinement aimed to capture any nuances and characteristics unique to the HES 2023 data that were not present in the previous HES 2017/ 18. The ML model significantly reduced manual coding.

An additional benefit of for automated coding was the consistency of code assignment. In the previous HES, the accuracy of manual code assignment was largely dependent on the understanding and interpretation of the data processing clerks. In the HES 2023, the ML model ensured that all records with the same description were assigned the same code, making it easier to identify and rectify any incorrect codes.

The use of AI-enhanced OCR and ML techniques in the HES 2023 enabled DOS to automate processes that traditionally required considerable manual effort. As result, DOS achieved significant savings in time and resources. This accomplishment provides further impetus for DOS to explore the use of these tools in other projects.

Singapore's Trade in Services
by Industry:

Examining Shifts in Services Trade Patterns in Recent Years

Introduction

International trade in services play a significant role in Singapore's economy, amounting to a total value of about 127% of Singapore's Gross Domestic Product (GDP) at current prices in 2022, up from 84% in 2010. The expansion of trade in services over the years was spurred by globalisation and technological advances which facilitated easier access to services abroad.

The Singapore Department of Statistics (DOS) has been publishing Singapore's trade in services statistics with further breakdown by services categories, export markets, and import sources. In 2024, trade in services by industry breakdown was released, offering insights into the industries that contribute to services trade in Singapore and the type of services traded by the respective industries. This article highlights trends in Singapore's trade in services by examining Singapore's exports and imports of services by industry from 2017 to 2022.

Scope and Coverage

The International Trade in Services survey, conducted annually by DOS, is the main source of data for Singapore's **trade in services** statistics. The survey covers services transactions between firms domiciled in Singapore and overseas trading partners. However, data for certain services categories such as Travel services and Government Goods and Services are compiled via administrative sources instead. Since administrative data sources lacks information on industry breakdown, these services categories are excluded from the estimates on trade in services by industry.

Trade in services by industry statistics are presented based on eight main industries, with an 'Others' category encompassing the remaining industries. These main industries contribute to the majority of services trade in Singapore, whereas industries in the 'Others' category includes either small industries or those driven by domestic consumption.

Singapore Standard Industrial Classification (SSIC) [1] of Industries

Industry	SSIC 2020
Manufacturing	10 to 32
Construction	41 to 43
Wholesale Trade	46
Transport & Storage	49 to 53
Information & Communications	58 to 63
Financial & Insurance	64 to 66
Professional Services	69 to 75
Administrative & Support Services	77 to 82
Others	All remaining SSICs

1] The **SSIC** is the national standard for classifying economic activities undertaken by economic units.

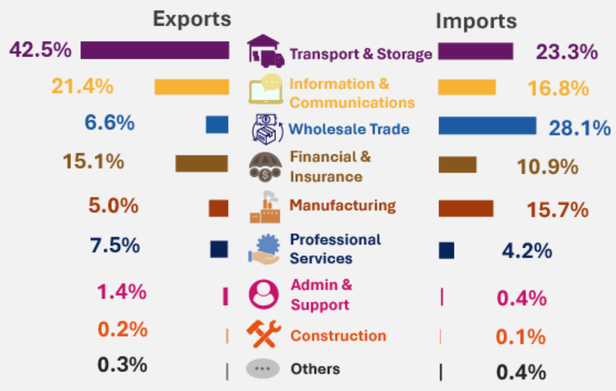
Findings

Services exports by the Transport & Storage Industry was the fastest growing industry in Singapore from 2017 to 2022, closely followed by the Information & Communications Industry.

From 2017 to 2022, Singapore's Transport & Storage industry was the top contributor to Singapore's services trade every year, recording a compound annual growth rate (CAGR) of 19.8%. This industry accounted for 42.5% ($184 billion) of services exports and 23.3% ($85.6 billion) of services imports in 2022 (Chart 1). Within the Transport & Storage industry, majority of its services exports was from the transport services [2], contributing 99% of the industry's total services exports in 2022.

Arising from the impact of the COVID-19 pandemic [3] and the accelerated pace of digitalisation in recent years, the Information & Communications industry had surpassed other industries to become the second largest contributor to Singapore's overall trade in services in 2022, recording a CAGR of 18.9% from 2017 to 2022.

Chart 1: Share to Total Services Exports/ Imports, 2022



With various measures such as remote working and social distancing introduced in response to the COVID-19 pandemic from 2020, the need for new digital tools to support these new arrangements rose sharply. Online conferencing for collaborative work, cloud storage, data security, and remote work solutions, together with digital entertainment such as streaming, proliferated. While these platforms and digitalisation trends were already present in the economy, the pandemic further compounded the demand for new digital products. This surge in demand led to a rapid growth in services exports by the Information & Communications industry from 2020 onwards (Chart 2A). Similarly, the Information & Communications industry experienced growth in services imports, overtaking the Manufacturing industry from 2021 as the third largest services importer (Chart 2B).

Chart 2A: Services Exports of Selected Industries, 2017-2022

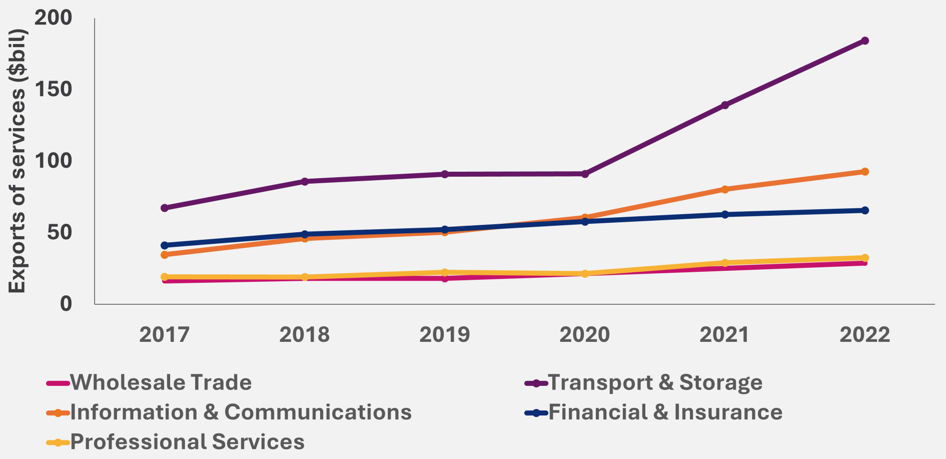
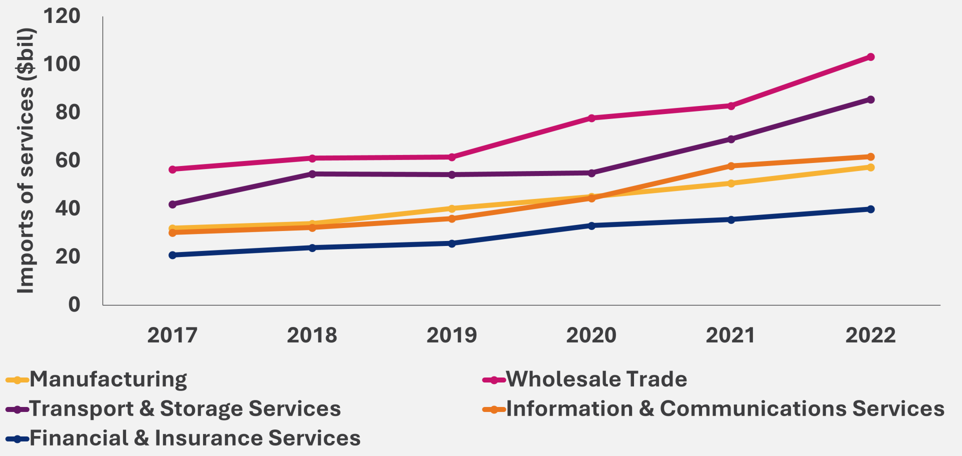


Chart 2B: Services Imports of Selected Industries, 2017-2022



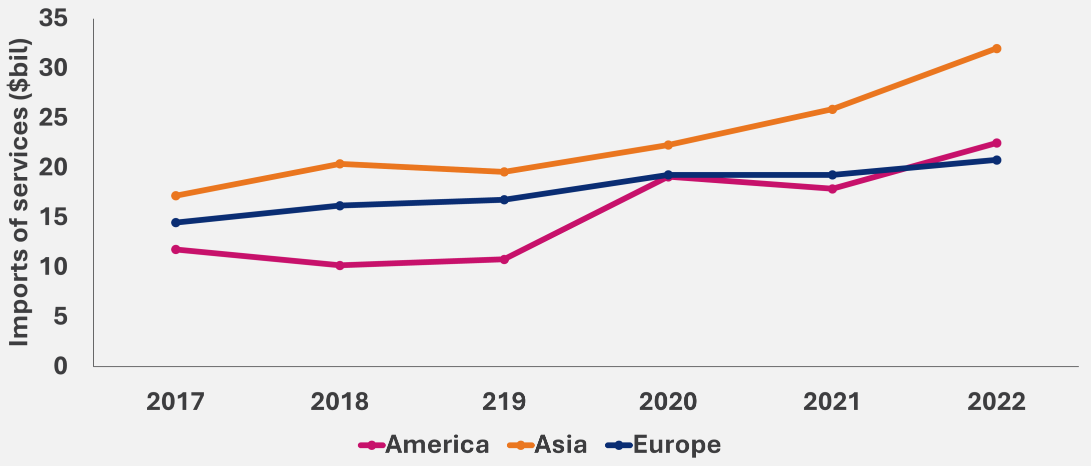
America Overtakes Europe as the Second Largest Source of Services Imported by the Wholesale Trade Industry

The Wholesale Trade industry was consistently the largest importer of services from 2017 to 2022. In 2022, it accounted for 28.1% of Singapore's services imports with transport services contributing 56.6% ($58.5 billion) to the industry's services imports. Trade-related services [4] was the second largest imported service at 12.1% ($12.5 billion).

The Asia region was the largest source of services imports for the Wholesale Trade industry from 2017 to 2022, due to its geographical proximity to Singapore (Chart 3).

Until 2021, Europe was the next largest source of services imports, closely followed by the America region. However in 2022, the Wholesale Trade industry imported more services from Asia and America which may be due to economic uncertainties in Europe stemming from the Russia-Ukraine war.

Chart 3: Services Imports of Wholesale Trade Industry from Selected Regions, 2017-2022



[2] Transport services include shipping of goods between countries and transport of passengers between countries.

[3] The World Health Organisation declared the COVID-19 outbreak a pandemic on 11 March 2020.

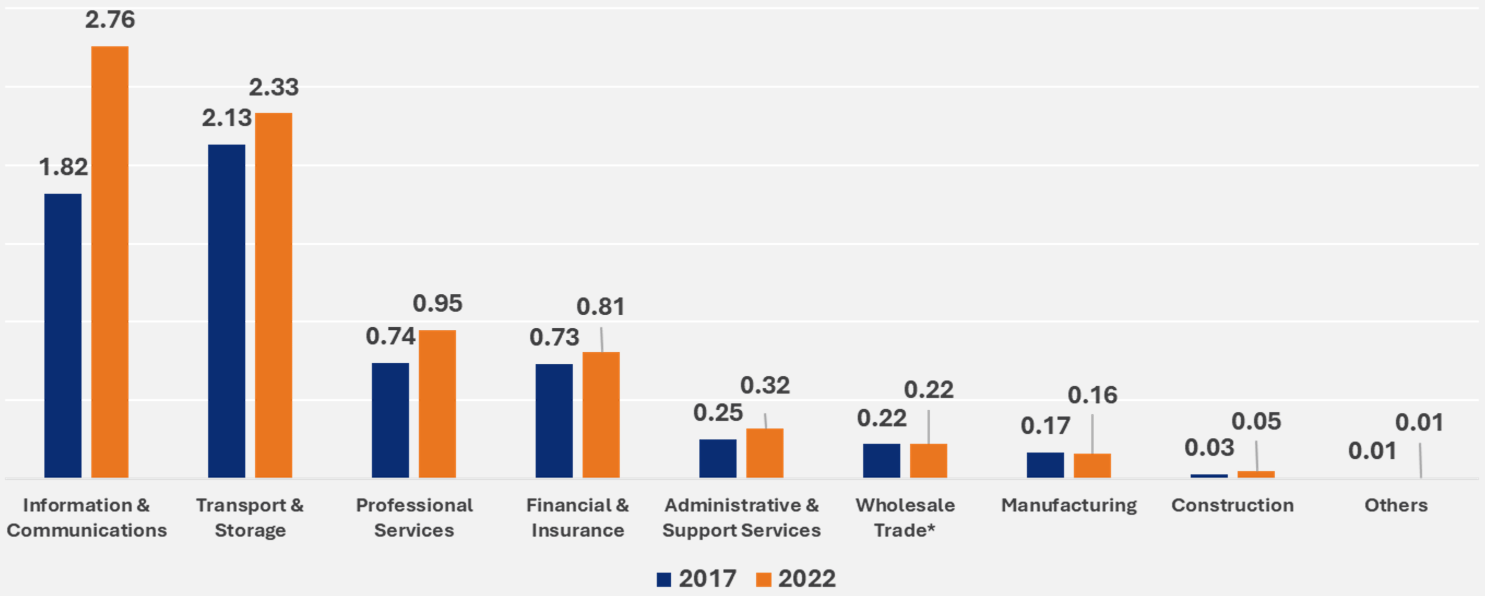
[4] Trade-related services consist of commissions, agency fees and distributor fees.

Export Propensity and Market Diversification: A Tale of Two Industries

By combining trade in services by industry statistics with other statistics such as value added by industry, additional insights can be derived. The services export propensity ratio (services exports by industry divided by **value added by industry**) measures the degree to which industries depend on external demand for services. The Information & Communications and Transport & Storage industries had export propensity ratios of 2.8 and 2.3 respectively in 2022, which suggested that these industries were more dependent on external demand. In contrast, other industries such as Administrative & Support Services had an export propensity ratio of 0.3 (Chart 4A).

A high export propensity ratio may indicate an industry's susceptibility to global economic trends and trade dynamics. For a more holistic picture, export propensity ratios should be analysed in conjunction with other indicators such as market concentration.

Chart 4A: Services Export Propensity by Industry, 2017 and 2022

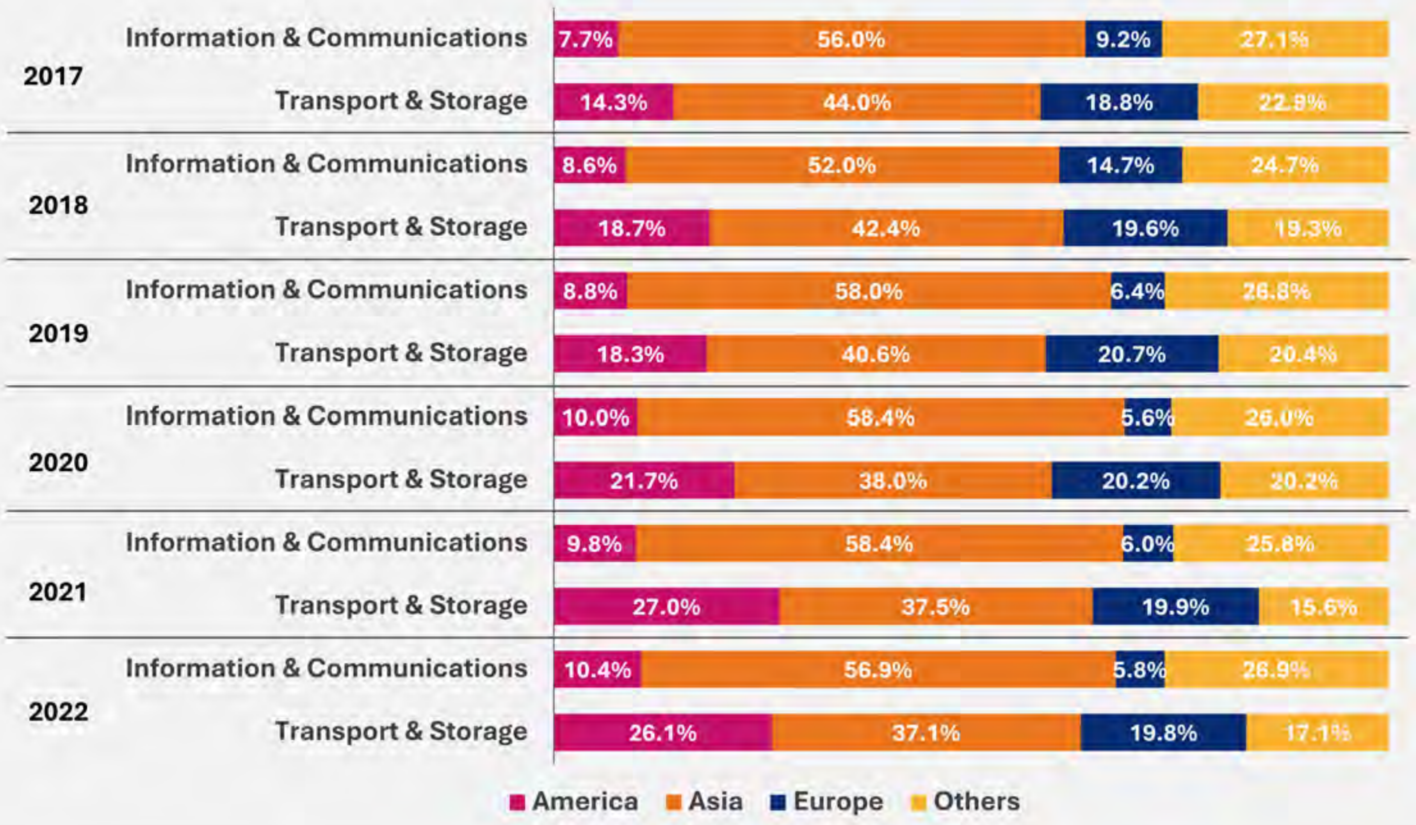


** The services exports propensity for the Wholesale Trade industry is low because the industry is primarily involved in the buying and selling of goods, and not so much in services. As such, the industry would have large exports of goods relative to its services exports.*

The Transport & Storage industry's services exports to the America region increased from 14.3% in 2017 to 26.1% in 2022 while exports to Asia steadily decreased from 44.0% in 2017 to 37.1% in 2022, showing diversification across the export markets (Chart 4B). The Information & Communications industry had high export propensity ratios from 2017 to 2022. On average, 56.6% of the Information & Communications industry's exports were concentrated in the Asia region during the period (Chart 4B).

Analysed together, this suggests that the Transport & Storage industry would likely be more resilient to adverse economic events compared to the Information & Communications industry, due to its more diversified exports.

Chart 4B: Concentration of Services Exports by Region for Selected Industries, 2017-2022



Conclusion

Singapore's trade in services has advanced substantially in recent years, driven by factors such as the services-oriented economy of Singapore, increasing digitalisation of services, and general post-pandemic recovery. Global events have caused disruption to the world economy and led to shifts in industry-specific trends. The newly published statistics on international trade in services by industry provide an additional lens to understand these changes in Singapore's services trade landscape.