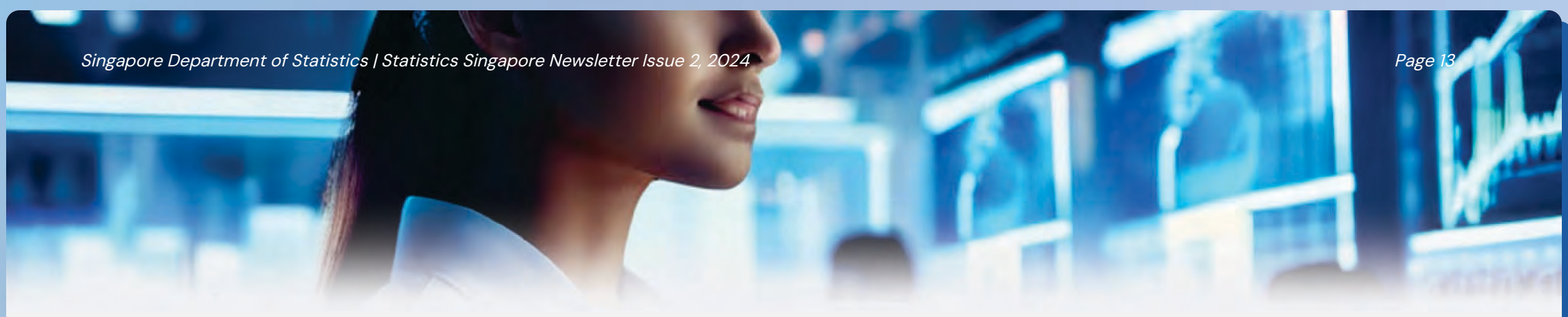


Use of AI-Enhanced OCR and Machine Learning for Data Processing in the Household Expenditure Survey 2023

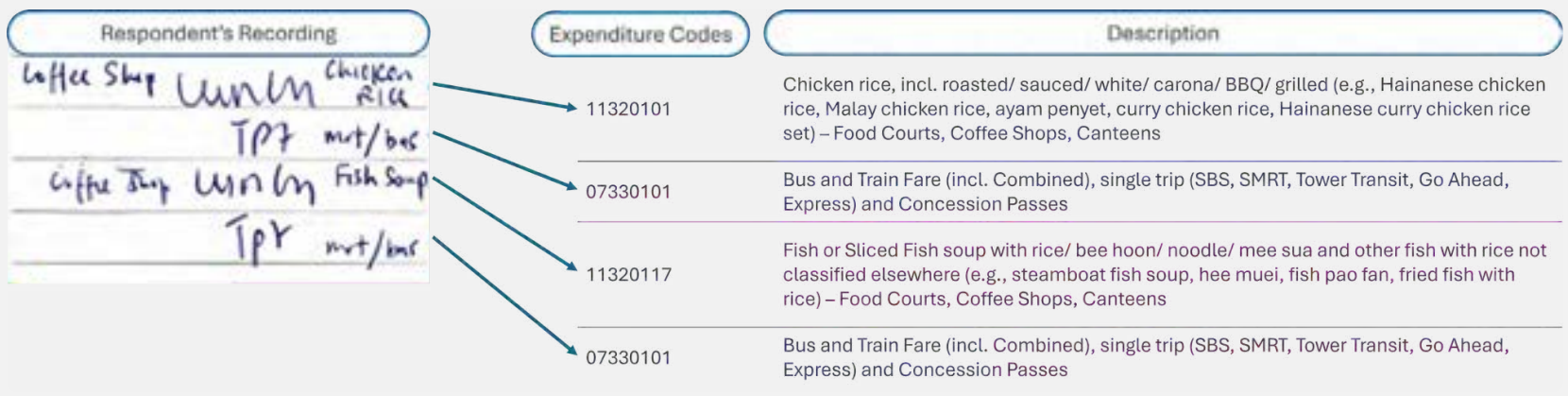
by Boon Kok Ann and Cheng Wan Hsien
Household Surveys and Expenditure Division
Singapore Department of Statistics

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



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.

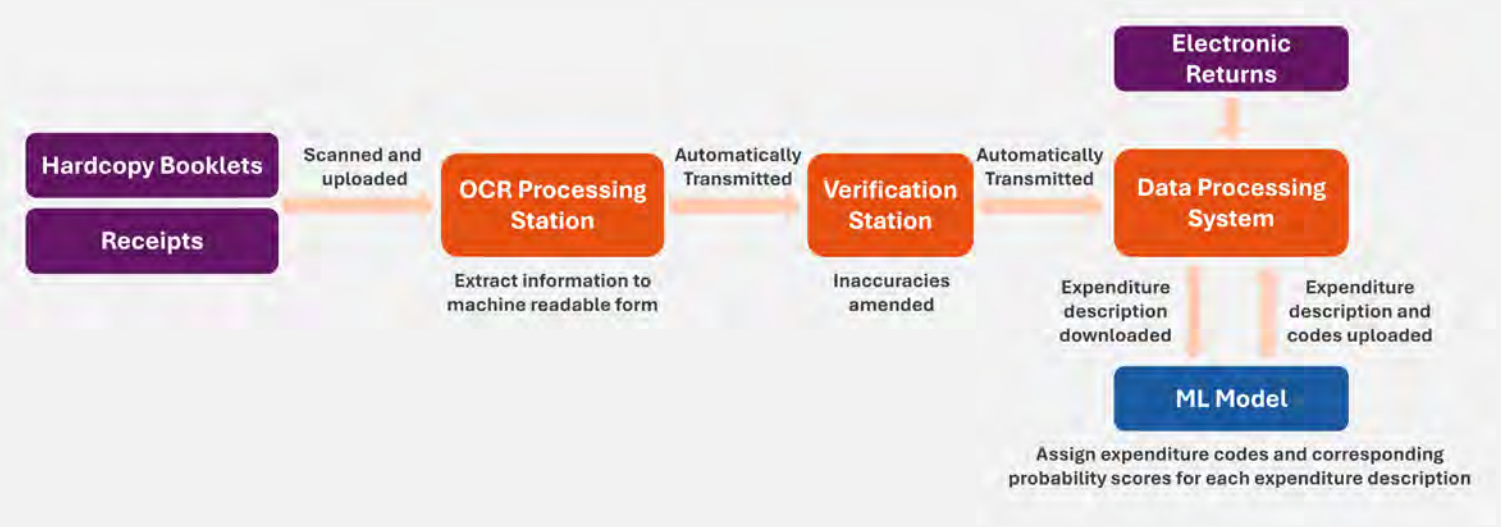
Together with electronic returns, expenditure descriptions were downloaded weekly from the data processing system and passed through an ML model. The model assigns expenditure codes and corresponding probability scores [3] to each expenditure description. If the probability score is above a predefined 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



Replacing Data Entry with AI-Enhanced OCR

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.

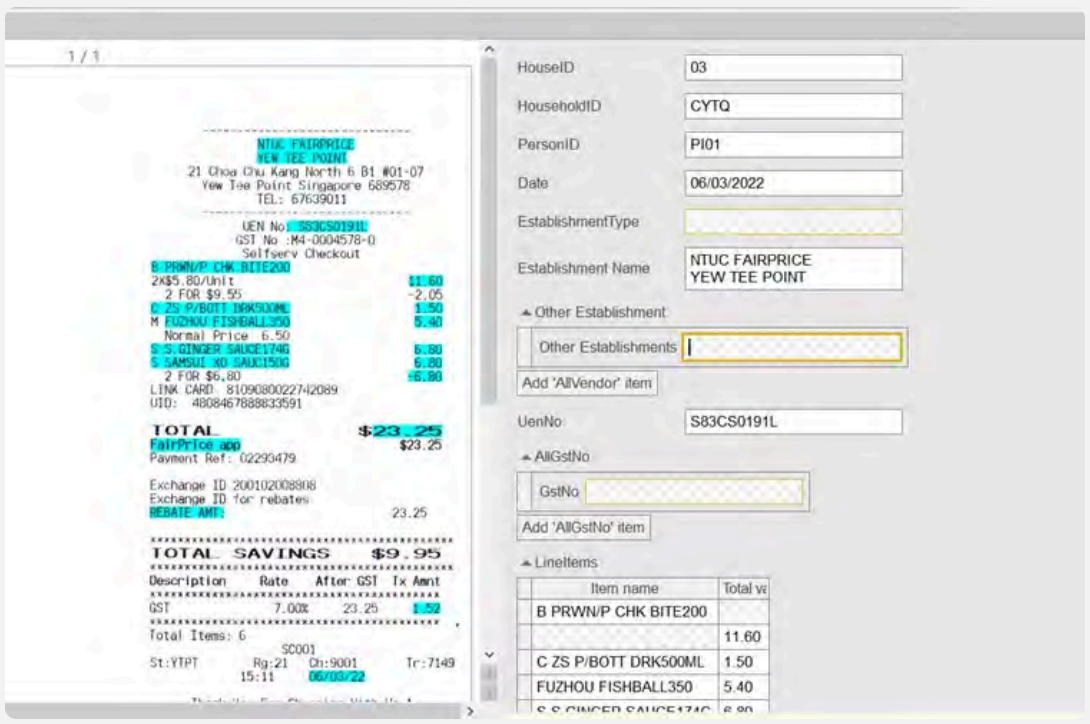
90%

Improved OCR Accuracy for printed text

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

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



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

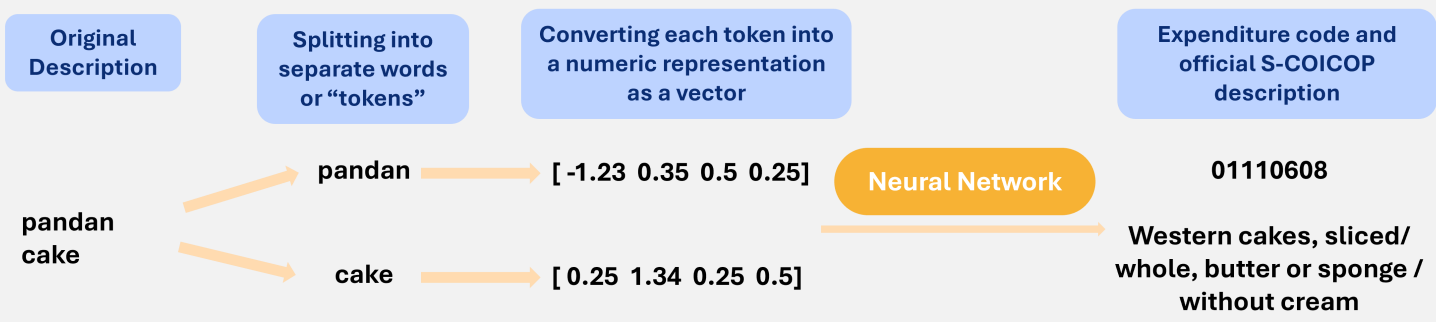
With the expenditure descriptions extracted via OCR, it became possible to apply ML models to automate the assignment of expenditure code to each expenditure item.

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



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.

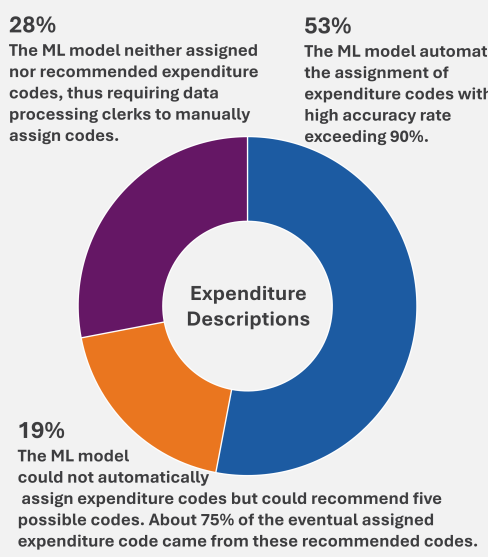
Code Recommendation

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.