

Coding of SSOC/SSIC in Census 2020 using Machine Learning

by Chen Tian Min, Teo Zhiwei, and Chan Wen Chang
Longitudinal Data Analytics Division
Singapore Department of Statistics

Introduction

The Singapore Census of Population 2020 (Census 2020) adopted a register-based approach, supplemented with a large-scale sample survey, to provide the most comprehensive source of information on the population and households of Singapore.

The processing of survey returns included the coding of free-text responses describing the occupations of respondents and the industries they work in, into Singapore Standard Occupational Classification¹ (SSOC) and Singapore Standard Industrial Classification² (SSIC) codes respectively.

The coding of occupations to SSOC codes was done using job titles and descriptions of main tasks and duties (Figure 1). The coding of industries to SSIC codes was performed using firm names and descriptions of principal economic activities (Figure 2).

Two methods of coding were used – batch coding and manual coding. All responses would first be processed using batch coding, which is an automated process that assigns appropriate codes using predetermined coding rules. Responses which could not be coded using the predefined rules were then manually coded. In general, batch coding was suitable for responses that were straightforward, while manual coding handled responses which needed human intervention.

This article presents the application of machine learning (ML) techniques in the processing of Census 2020 survey returns to improve the coding of free-text responses and reduce manual coding. The improvements were realised in two areas: 1) *automatic coding* of responses with strong predictions; and 2) *recommendation of codes* for responses with weaker predictions. The addition of a ML model into the coding process reduced the number of responses which were passed to the manual coders, and was estimated to have saved 5,600 man-hours.

FIGURE 1 EXAMPLE OF CODING OCCUPATIONS

Job Title	Main Tasks and Duties	→	SSOC Code
Head waiter	Manage waiters and ensure guest satisfaction	→	51311 – Captain waiter

FIGURE 2 EXAMPLE OF CODING INDUSTRIES

Firm Name	Principal Economic Activity	→	SSIC Code
Big Wheels	Manufacture non-motorized bicycles	→	30920 – Manufacture and assembly of bicycles, tricycles, trishaws and invalid carriages (including parts and accessories)

1 The SSOC is the national standard for classifying occupations. It consists of structured five-digit codes that classify occupations by their main tasks and duties. The SSOC publication is available on the SingStat Website at www.singstat.gov.sg/standards/standards-and-classifications/ssoc.

2 The SSIC is the national standard for classifying economic activities undertaken by economic units. It consists of structured five-digit codes that classify firms by their principal economic activities. The SSIC publication is available on the SingStat Website at www.singstat.gov.sg/standards/standards-and-classifications/ssic.

Using Machine Learning for Automatic Coding

ML models are capable of learning sophisticated rules for automatic coding. Hence, they can handle responses that batch coding cannot. If ML models cannot confidently code a response, they can provide suggested codes for the manual coders' reference, thereby expediting the coding process. These qualities make a ML model suitable for implementation as a step in between the batch coding and manual coding steps (Figure 3).

The ML coding step used in Census 2020 comprised four sub-steps (Figure 4):

- 1) *Data preparation*: Pre-processed training data into a standard format acceptable for model use. This sub-step was also applied to survey responses during model deployment.
- 2) *Model training and selection*: Decided on model specification for use in deployment.
- 3) *Coding of responses*: Generated predicted codes for the survey responses.
- 4) *Quality assessment*: Evaluated quality of predicted codes and identified issues to be addressed.

This was also the start of feedback loop to glean insights and incorporate feedback to improve the entire ML coding step.

Data Preparation

As the survey responses involved descriptive information in free-text format, it was necessary to pre-process them into a standard format so that a ML model could use them. This also helped to improve information quality.

Some of the text pre-processing techniques used in the data preparation sub-step included:

- i. Converted abbreviations to their full forms and standardised characters of the same field to the same letter case
- ii. Removed uninformative characters and words (e.g. 'is', 'the')
- iii. Corrected spelling, while keeping localised words (e.g. 'garang guni' was corrected to 'karung guni')

As an example, the sentence "I serve customrrs in a F&B business" was pre-processed into "serve customers food beverage business".

FIGURE 3 IMPLEMENTATION OF ML CODING IN THE CODING PROCESS

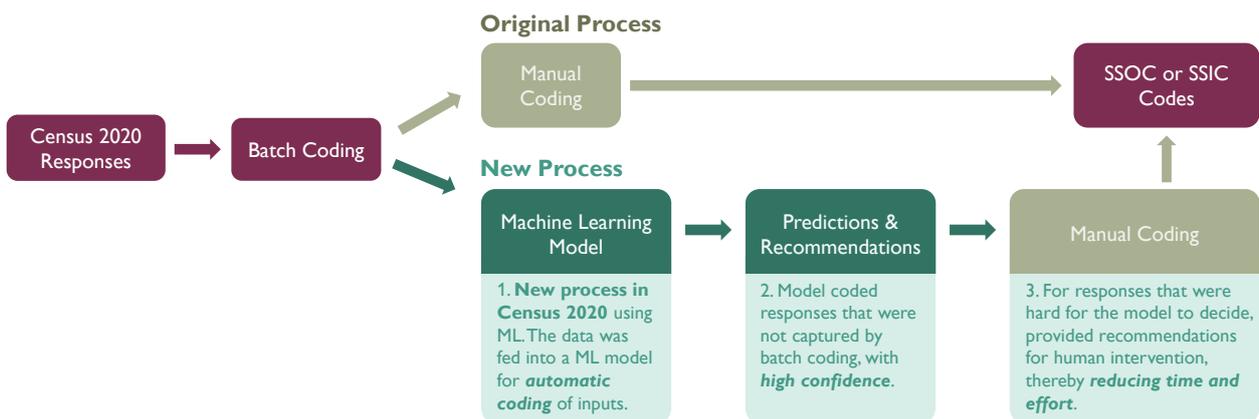
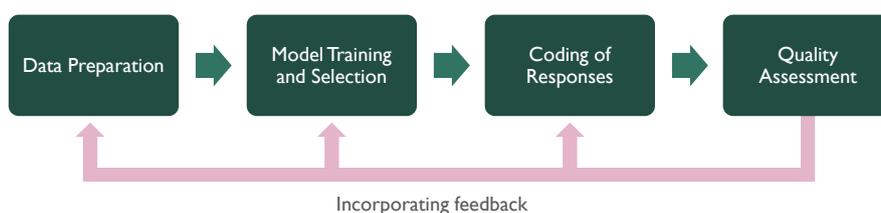


FIGURE 4 WORKFLOW FOR ML CODING STEP



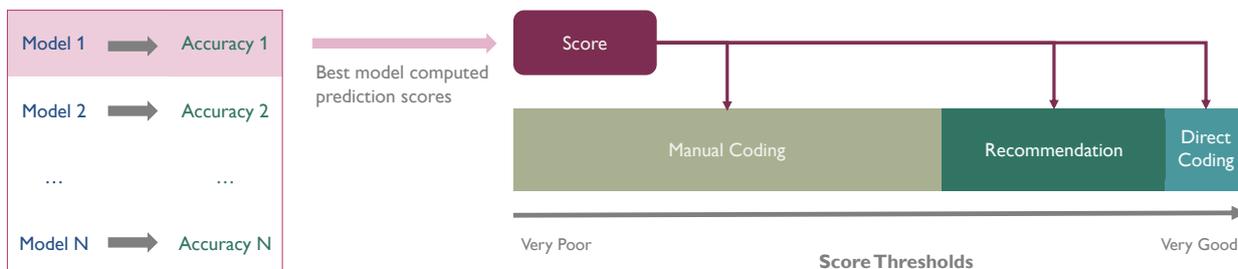
Model Training and Selection

In this sub-step, different supervised ML models were trained using data from past Comprehensive Labour Force Surveys conducted by the Ministry of Manpower. To ensure compatibility, the SSOC and SSIC codes in

the training data from past years were mapped to the latest codes, i.e. version 2020, before they were used.

After comparing the prediction accuracies³ of the various ML models (Figure 5), a neural network (NN) model was evaluated to be the best performing model.

FIGURE 5 ML MODEL TRAINING AND SELECTION, AND CODING OF RESPONSES



Basic Workings of Neural Network Models

NN models are a subset of ML models that mimic the way the human brain processes information. They typically consist of interconnected units, or “nodes” that resemble biological neurons, and connections between the units, or “weights” (Figure 5A). A NN model can be structured using three types of layers:

- **Input layer:** The first layer in which data are input into the network
- **Hidden layers:** Intermediate layers in which computations are performed on the data from the input layer, so that the data can be mapped into an output
- **Output layer:** The layer in which the output is generated

When an input, in numerical form, passes through a NN model, all node values from the previous layer are multiplied by the respective weights and added together (as reflected by the arrows in Figure 5A). The values are then transformed by a function at the respective nodes to produce the values for the next layer of nodes. This process is repeated until the output is generated. Figure 5B shows an example of how the node values pass from one layer to the next, in a simple feedforward NN with only one hidden layer. The bias terms are omitted for simplicity.

FIGURE 5A

EXAMPLE OF A NEURAL NETWORK SCHEMATIC

Nodes are represented by the circles and weights are represented by the arrows.

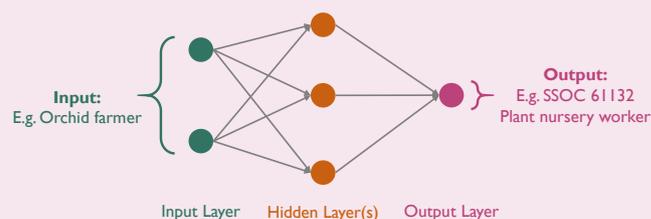
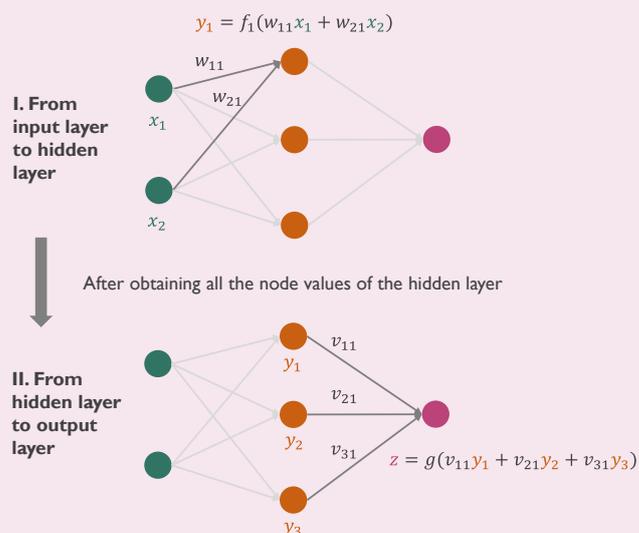


FIGURE 5B

EXAMPLE OF AN INPUT PASSING THROUGH A SIMPLE FEEDFORWARD NN WITH ONE HIDDEN LAYER



³ Prediction accuracy is a measure of how well a ML model is at providing the correct codes. It is obtained by using the model on responses that were not used for training and evaluating how many of those responses were coded correctly.

Coding of Responses

For each survey response, the NN model computed scores for all the possible SSOC/SSIC codes at the 5-digit level, which measured how likely the codes would be assigned to the response.

The code with the highest score was selected to be the predicted code. The score of the predicted code would then determine whether the response would undergo direct coding, recommendation or manual coding (Figure 5):

- *Direct coding:* When the score of the predicted code was very good, the response was coded with the predicted code.
- *Manual coding:* When the score of the predicted code was poor, the predicted code was not used. The response was passed on to manual coders to assign a code.
- *Recommendation:* When the score of the predicted code was neither poor nor good enough for direct coding, codes were recommended. Manual coders reviewed the recommended codes and assigned the most suitable code to the response. This process reduced the time and effort incurred as compared to manual coding directly.

Quality Assessment

The NN model’s ability at handling incoming survey responses was monitored via periodic checks and feedback. The insights gleaned from the checks and feedback were used to improve the coding performance of the model. The following lists two issues that were identified:

a. Abnormal and new data behaviour

Due to COVID-19, the responses received reflected new working circumstances. Numerous respondents indicated that they worked from home as the job description, while retaining their job titles.

There were also new occupations such as “safe distancing ambassador” and “swab test assistant” which did not exist previously. This shift in data behaviour affected the model’s performance, since it was trained based on pre-COVID-19 data.

b. Insufficient details in survey responses for certain occupations and industries

As with manual coding, if survey responses were not detailed enough, the NN model would not be able to provide accurate code predictions. Such cases usually required additional information from the respondents in order to be coded definitively.

As the example (Figure 6) illustrates, there could be many possible SSOC codes for a response on occupation that provided insufficient information to be allocated a code from numerous possible codes. More details on the driven vehicle would be useful in identifying the most suitable code.

To address the issues discovered, sub-steps 1, 2 and 3 were updated. Some of the updates included re-training the NN model with responses that were manually coded earlier, and reassigning predicted codes originally for direct coding to recommendations (for the affected groups of SSOC and SSIC codes).

FIGURE 6

EXAMPLE OF A RESPONSE THAT COULD NOT BE REASONABLY CODED FOR OCCUPATION

Job Title	Main Tasks and Duties	Possible SSOC Codes
Driver	Drive vehicle	83221 – Taxi driver 83222 – Chauffeur 83226 – Private-hire car driver

Effectiveness of Machine Learning in Census 2020

The ML coding step was estimated to have saved 5,600 man-hours in Census 2020. Most of the savings was for SSOC coding instead of SSIC coding, due to the following reasons.

Firstly, the task of a ML model predicting appropriate SSIC codes was harder than that for SSOC. The challenge in assigning SSIC codes stemmed from firm names not being analogous to job titles. For example, firm names might not be in proper English; even if they were, the words in the names did not necessarily carry their usual meanings.

Secondly, a large percentage of respondents were able to find and verify their firm names via a pre-defined list in the Census 2020 questionnaire; and this reduced the need for inputting free-text. As the list had a direct mapping to SSIC codes, responses that were selected from this list were batch coded without the need for the ML coding step. This efficiency of the pre-defined

list for SSIC resulted in less use for the ML coding step, and only the most challenging responses that eluded the dropdown list required ML and manual coding.

These two factors led to the decision of having the NN model to only provide recommendations for SSIC coding instead of performing direct coding, resulting in less contributions for the coding of industries.

Concluding Remarks

The use of ML in Census 2020 had demonstrated that standardised tasks, such as the coding of SSOC and SSIC codes, which traditionally required human effort, can be automated using appropriate ML techniques.

Given the rapid developments in ML, ML is expected to contribute significantly in future surveys and related data processing problems.

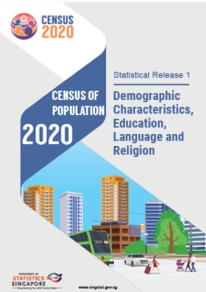
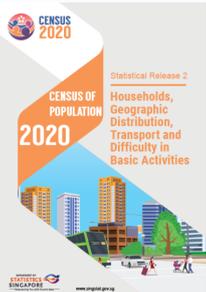
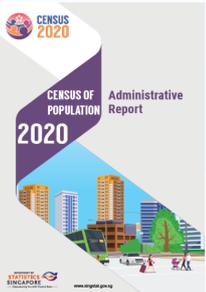
That said, manual monitoring and intervention are still integral to the deployment of a ML process, as they help the process adapt to real world issues, and ensure accuracy and usefulness of the outputs.

CENSUS OF POPULATION 2020

The Census 2020 provides comprehensive data on the characteristics of the resident population:

- Demographic Characteristics
- Education
- Language
- Religion
- Households
- Geographic Distribution
- Transport

Newly collected data on Place of Work and Difficulty Performing Basic Activities are also available.



DEPARTMENT OF
STATISTICS SINGAPORE
Empowering You with Trusted Data

Download

Download the [Census of Population 2020](#) publication for more information.