香港統計學會

Hong Kong Statistical Society

c/o Department of Statistics & Actuarial Science,
The University of Hong Kong, Pokfulam Road, Hong Kong
http://www.hkss.org.hk

Bulletin
Volume 44 No.1
April 2022

Editor's Foreword

Dear Members,

Welcome to the 2022 issue of the HKSS Bulletin.

It is a challenging year. The pandemic hits our city hard. This makes most of us work from home since the Lunar New Year. The outbreak has caused many events to postpone, reschedule or even cancel.

In this issue, Dr. Mehmood NAWAZ of The Chinese University of Hong Kong shares with us how deep learning can be used to detect objects from natural images with complicated backgrounds. Miss Natalie CHUNG and Mr Ian NG of the Census and Statistics Department introduce how they apply text analytics in the compilation of Hong Kong's external merchandise trade statistics. The Organising Committee shares with us some highlights of the 35th Statistical Project Competition for Secondary School Students.

Lastly, I would like to express my heartfelt thanks to all members of Editorial Board and contributors for their great contributions to this Bulletin.

Benson LAM

		Phone	Fax	Email
Editor	: Dr. LAM, Benson Shu-yan, HSUHK	3963 5450	-	bensonlam@hsu.edu.hk
Secretary	: Dr. LI, Billy Yeuk-goat, C&SD	2867 1068	2537 2575	-
Member	: Ms LAI, Carly Yuk-ling, C&SD	2582 4888	2802 1192	yllai@censtatd.gov.hk

CONTENTS

(Vol. 44/No.1, April 2022)

Page

President's Forum

Professor Alan WAN Tze-kin

Hong Kong has entered the fifth wave of the COVID-19 pandemic. Due to health restrictions, many of our society's events such as seminars and workshops have been postponed or cancelled. One such event was the workshop on data visualisation through PowerBI, a powerful interactive data visualisation software with user friendly interface for the creation of dashboards and reports. We had hoped to hold this workshop in 2021 as a sequel to a similar workshop held in 2020, where PowerBI for basic data analysis such as box and whisker plot, probability density plot and scatter plot were introduced to secondary school teachers. The workshop of 2021 is being postponed to 2022.

Similarly, the social distancing measures have necessitated the cancellation of the half-day eco-tour to Hoi Ha Wan Marine Park in Sai Kung, a joint event with the Association of Government Statisticians (AGS) originally planned for December 2021. The cancellation of this trip was a pity because Hoi Ha Wan is rich in marine biodiversity and is one of Hong Kong's six natural marine parks. The tour will be postponed to the second half of 2022.

Despite the difficulties caused by COVID-19, we did however manage to hold a statistical training course for occupational therapists of the Fu Hong Society on 17 December 2021. The course was delivered by Prof. May WONG, our Consultation Services Secretary .

The Bulletin contains two research articles. The first article by Dr. Mehmood NAWAZ discusses the challenges of detecting objects from natural images and introduces a hybrid method to obtain high-level features from images. The method entails developing a preliminary feature map in the first stage using a pre-trained convolutional neural network (VGG-16), which is a trained deep learning model from another database. The outcome is then refined by affinity background subtraction techniques in the second stage. The experimental results demonstrate that the method works well and is a promising tool for detecting objects from natural images. The second article by Miss Natalie CHUNG and Mr Ian NG discusses the application of text analytics models for compiling external merchandise trade statistics in Hong Kong. Examples are provided to illustrate the effectiveness of the technique and improvements in outcomes compared with other commonly used techniques.

Object Detection and Background Subtraction Application in Digital Images

Dr Mehmood NAWAS
The Chinese University of Hong Kong

Introduction:

In computer vision applications, object detection is becoming more prevalent. The primary purpose of object detection is to simulate human perception-based detection into relevant information for various applications. The human visual system acts as a filter, directing greater attention to the salient portions of an image, which are the most appealing and engaging parts of the image. It is a difficult task in computer vision to build a human-like perception-based object identification system in crowded images with noisy backgrounds. Recently, many object detection techniques have been described, using background information as border priors to find salient objects from images.

People are very attentive to critical information in digital images and videos. Computer vision specialists strive to recreate this human "technology" on computers to interpret and recognise image information. The object in an image that draws human attention is known as a salient object, and it occupies a certain spatial location. Identifying a prominent object in an image is a complex task. Because of the high resolution in the center of the retina, the human eye naturally directs its look to the center of the visual image. The contrast of an image inside the superpixel plane characterises the object that draws the greatest attention (i.e., intensity, color, or orientation, the most attractive factors for human vision).

The spatial information inside images affects the human visual attention process. The RGC (retinal ganglion cell) system is positioned on the retina's inner side. It collects spatial information from photoreceptors before transmitting it to the brain to be processed into natural images. The human visual system focuses on regions with a chromatic advantage and a spatial distribution that is highly compact. As a result, the primary goal of object detection approaches is to organise the image's most intriguing items in a fashion that mimics the human visual perception system. In [1], Nawaz et al. provide complete detail of several object detecting techniques.

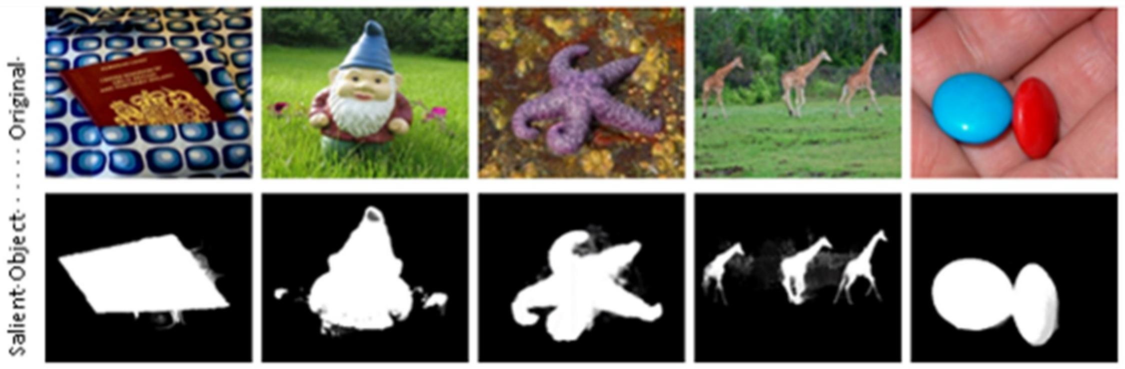


Figure 1: Examples of object segmentation

Challenges:

Although saliency detection has been extensively developed in recent years, there are still many ways to improve the application of saliency to more practical tasks. In terms of robustness and efficiency, this is more related to the limitations of existing methods. The challenges that need to be addressed are as follows:

(1) Image complexity is the most prevalent barrier to recognising prominent objects. It isn't easy to separate and characterise the foreground region from the background region (i.e., color, texture, contrast).

(2) When the background area is made up of many components, as seen in Fig. 1 (a). The basic object detection method cannot completely remove the backdrop at the object level.

(3) Detection algorithms face a difficult task when dealing with large amounts of data. Because saliency detection seeks to recognise a common prospect from different images, it may be used to find models from enormous amounts of data. The majority of joint saliency algorithms are developed for small and medium data sets (100-1000 images), but not for huge data sets. The dataset may have unsolved problems, including emission, noisy images, and category changes. The answer to these challenges with large-scale data is an area of research that should be pursued in the future.

(4) Another challenge is the detecting algorithm's efficiency and consistency. The majority of saliency detection approaches involve exploring and extracting contrast information from each image separately from the other images in the group.

(5) In visual scenes, illumination is an important part of perception. The interplay of light and shadow, and hence the placement of the light source, can have a significant impact on the salient object. Several light sources, such as a single light source highlighting any texture or function, can dilute these interactions in a real image.

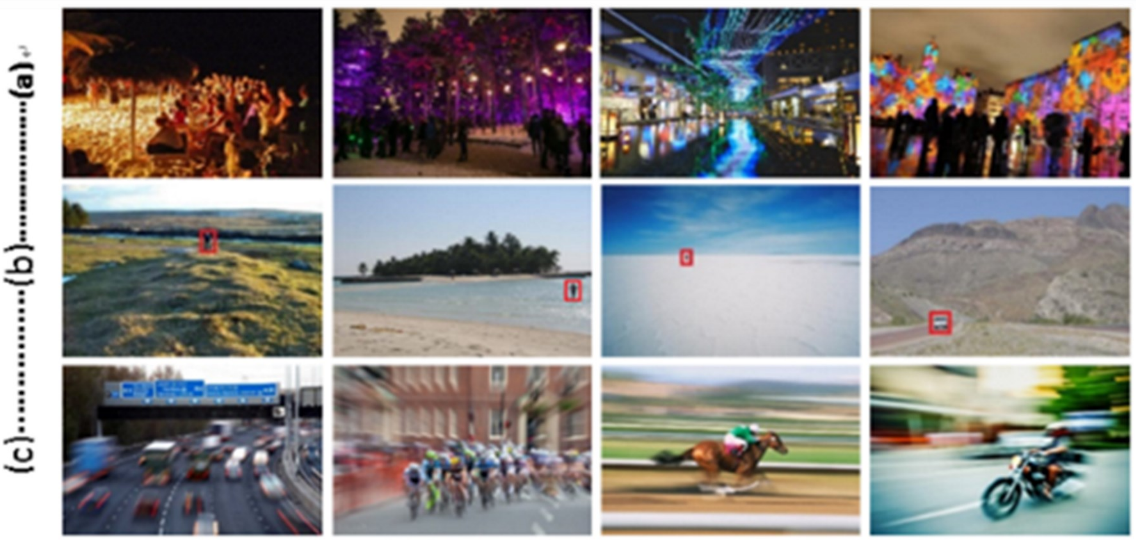


Figure 2: Challenges (a) shows the illumination in images (b) the detection of small size objects, and (c) images of fast-moving objects

(6) To extract the deep semantic properties of an image, existing object recognition algorithms based on highly condensed neural networks must execute multiple-level convolution and aggregation operations on the entire image. These models can produce better outcomes for multiple objects.

Methodology:

A hybrid feature-based framework is developed to detect the salient areas, which incorporates high-level features from both supervised and unsupervised approaches. A maximum attention map (MAM) based on characteristics of a pre-trained convolutional neural network (VGG -16) is developed, which is inspired by the attention map [3]. This MAM will create an expanded attention map using a multi-morphological reconstruction on different feature layers. The High-level feature maps are denoised using a reconstruction approach that removes the noisy and undesired pixels. An affinity-based background subtraction approach is used to remove foreground and background areas using color similarities and the color information flow between the foreground and background regions. This affinity-based background subtraction technique effectively deals with multiple and small sizes salient region structures relative to other background subtraction techniques, [9]. A consistency-based integration technique to extract dense and fully informative salient regions from different saliency maps is also developed.

An attention map is a system for assigning different weights to extract geographic region data and for creating a weighted heat map, which is helpful for object detection. Many attention maps have been included in object localisation and prediction into saliency detection algorithms [2], [4]. Khan et al. [3] suggested spatial neural attention for image segmentation by creating two attention maps. As illustrated in Fig. 3, both rows show the expansive attention map, which encompasses the whole salient object. Image captioning and question answering have both benefited from the attention map.

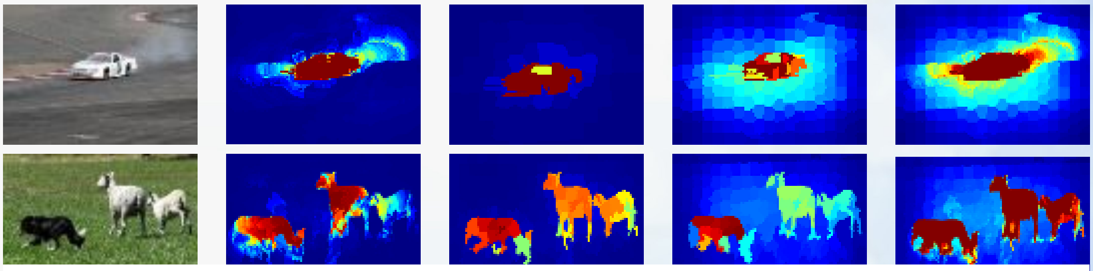
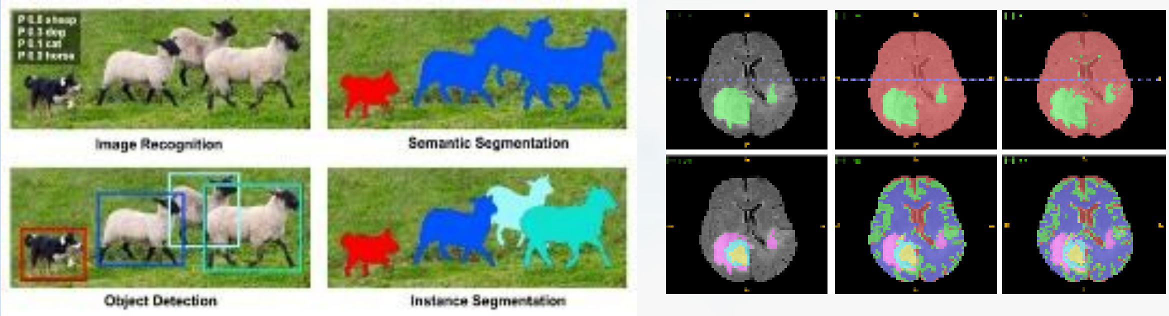


Figure 3: Examples of attention map

In object detection, accurate background removal is a difficult challenge. We noticed that the background superpixels had a distinct spatial arrangement contrast. The image border is related to the background area. Traditional background removal approaches such as threshold-based, probability-based, and linear combination-based techniques have all been employed. Weighted contrast background reduction approaches have attracted much attention [5], [6]. For example, Nawaz et al. [2] suggested a robust background approach based on border connectivity. It defines the spatial arrangement with the image boundaries and separates foreground and background pixels using a geometrical interpretation. Because of the probability-based threshold, this method has several disadvantages. The saliency accuracy decreases when the foreground cluster is smaller than the background cluster.

Applications:

Saliency detection distinguishes an object from its neighbors in the image. It has a wide range of applications such as object segmentation [1], image cropping [6], image retargeting [7,8,9], adaptive compression, image matching [4], and visual surveillance. Saliency detection is also used to predict eye fixation and detect small objects. It is hard to make a global framework to detect a salient object due to a textured background, non-uniform intensity, or low SNR (signal to noise ratio) in the image. It is widely used to detect a tumor in medical images, as shown below.



(b)

Figure 4: (a) shows the object recognition and detection in RGB images and (b) shows the tumor detection in brain images

To compare with the image segmentation results of the above methods, the results of detailed experiments are shown in Figs. 4&5. The six data sets ECSSD, MSRA10K, SED1, SED2, DUT-OMRON, and HKU-IS are used to indicate the segmentation results of the proposed and other methods. All saliency maps are converted into segmentation results by fixed thresholding with 0.5. Fig. 5 show the background subtraction comparison of the image segmentation results.

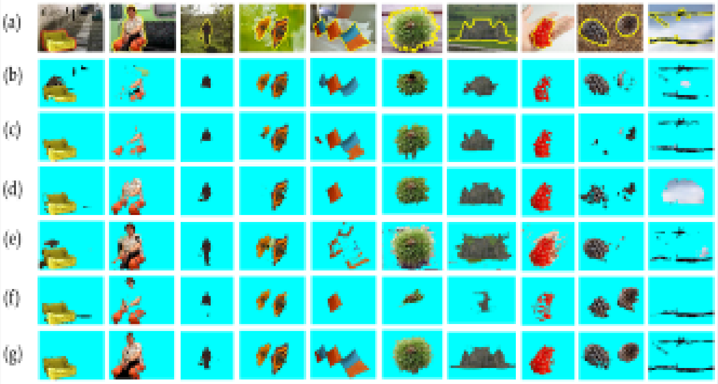


Figure 5: Results of image segmentation on different images. (a) original images with ground truth marked by yellow color counter. The results of different methods including (b) DSR, (c) RBD, (d) RRW, (e) MDF, (f) DISC and (g) Ours.

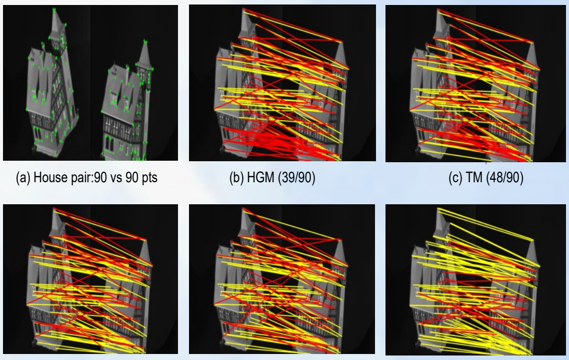


Figure 6: Matching output of CMU house dataset: Diagram (a) shows the 90 feature pts on image pairs. Diagram (b), (c), (d), (e) and (f) show the accuracy of HGM, TM, RRWHM, BCAGM and the proposed method, respectively. The yellow lines show the correct matching and red ones show incorrect matching. (Best viewed in color)

References:

1. Nawaz, Mehmood, and Hong Yan. "Saliency detection via multiple-morphological and superpixel based fast fuzzy C-mean clustering network." Expert Systems with Applications 161 (2020): 113654.

2. Nawaz, Mehmood, and Hong Yan. "Saliency Detection using Deep Features and Affinity-based Robust Background Subtraction." IEEE Transactions on Multimedia (2020).

3. Khan, Sheheryar, Mehmood Nawaz, Xu Guoxia, and Hong Yan. "Image correspondence with CUR decomposition-based graph completion and matching." IEEE Transactions on Circuits and Systems for Video Technology 30, no. 9 (2019): 3054-3067.

4. Nawaz, Mehmood, Sheheryar Khan, Rizwan Qureshi, and Hong Yan. "Clustering based one-to-one hypergraph matching with a large number of feature points." Signal Processing: Image Communication 74 (2019): 289-298.

5. Nawaz, Mehmood, Sheheryar Khan, Jianfeng Cao, Rizwan Qureshi, and Hong Yan. "Saliency detection by using blended membership maps of fast fuzzy-C-mean clustering." In Eleventh International Conference on Machine Vision (ICMV 2018), vol. 11041, p. 1104123. International Society for Optics and Photonics, 2019.

6. Nawaz, Mehmood, Rong Xie, Liang Zhang, Malik Asfandyar, and Muddsser Hussain. "Image super resolution by sparse linear regression and iterative back projection." In 2016 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1-6. IEEE, 2016.

7. Qureshi, Rizwan, Mehmood Nawaz, Avirup Ghosh, and Hong Yan. "Parametric models for understanding atomic trajectories in different domains of lung cancer causing protein." IEEE Access 7 (2019): 67551-67563.

8. Asfandyar, Malik, Mehmood Nawaz, Xie Rong, Liang Zhang, and Muddsser Hussain. "Block of Interest Based AVS to HEVC Transcoding with Resolution Conversion." In International Forum of Digital TV and Wireless Multimedia Communication, pp. 296-306. Springer, Singapore, 2016.

9. Qureshi, Rizwan, Mehmood Nawaz, Sheheryar Khan, Ali Raza Shahid, Avirup Singh, and Hong Yan. "Principal Component Analysis and Clustering to reveal the conformation dynamics of EGFR with L858R and T790M Mutation." In 6th International Conference on Bioinformatics Research andApplications (ICBRA 2019). 2019.

Application of Text Analytics in Compilation of External Merchandise Trade Statistics

Miss Natalie CHUNG and Mr Ian NG
Census and Statistics Department

1. Introduction

Hong Kong's external merchandise trade statistics are compiled based on administrative data given in trade declarations submitted by importers and exporters to the Government. With about 20 million electronic trade declarations received annually, the trade statistics system is among the largest statistical systems of the Census and Statistics Department (C&SD) in terms of data volume and velocity. The availability of a large corpus of textual commodity descriptions and a wide array of trade data items in electronic format provides opportunities for applying big data analytics in the quality checking of trade documents and thus improving the existing trade processing and statistical regime.

C&SD has initiated exploratory studies in this regard since 2018. The results are encouraging. The text analytics models developed have brought about impressive improvements in the quality checking mechanism of trade declarations in terms of efficiency and accuracy and have been incorporated into our regular trade declaration processing system as from early 2020. Enhancements to the models have been made on a continuous basis since then.

2. Trade declaration processing

Administrative data given in trade declarations include trade types (imports, domestic exports or re-exports), commodity codes classified according to the Hong Kong Harmonized System (or HKHS), commodity descriptions in free text format, trade values, trade quantities, countries/territories of origin, countries/territories of consignment, company names and transport modes.

To ensure the accuracy of data declared in trade declarations, in particular the HKHS commodity codes which are relatively error-prone because of the detailed classification involved, a rule-based Risk Management Model (RMM) is currently in place to identify trade records that are most susceptible to having reporting errors. While trade records with simple and straightforward errors are adequately identified and rectified under the RMM, those involving more complex issues, in particular those requiring reconciliation with the unstructured textual commodity descriptions provided, have to be handled through labour-intensive manual checking and rectification. In sum, the RMM has not fully utilised all unstructured data and the checking quality in some cases hinges much on the experience and product knowledge of individual checkers to discern reporting errors.

To address the limitations for RMM, C&SD has been exploring the application of text analytics using machine learning techniques to process trade declarations in parallel with the RMM, aiming to further automate the business workflow and increase the error-detection performance, as well as to save manpower resources.

Text classification has been used recently in spam checking [1], as well as sentiment analysis [2], where the prediction vector (the number of classification categories) is usually short. For example, in sentiment analysis, usually only several main categories of emotions are involved (i.e. positive, negative and neutral). In our study, however, the aim is to classify commodities into thousands of 8-digit HKHS codes, which is particularly challenging and complicated as the prediction vector is substantially longer.

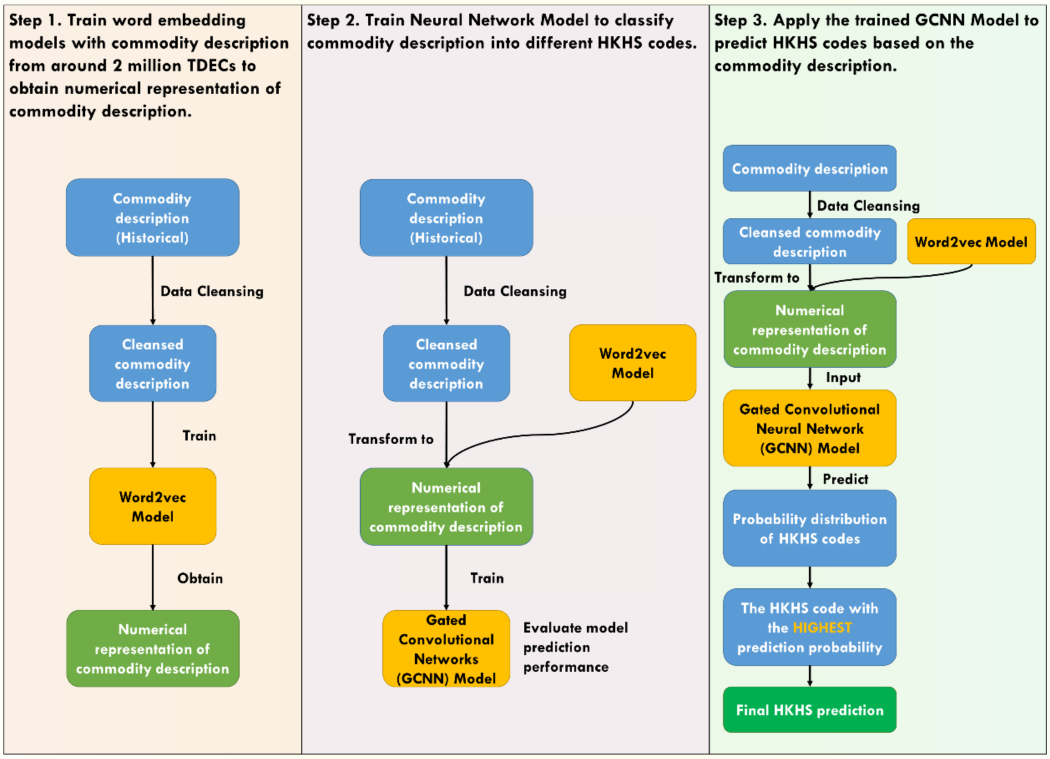
3. Model development

Stage 1: Text analytics on commodity description using Gated Convolutional Neural Network (GCNN)

A text analytics model using Word2Vec word embedding and Gated Convolutional Neural Network (GCNN) was first developed in 2018 on an exploratory basis to identify the most probable HKHS commodity code of a trade record based on the textual commodity description reported. This aimed to facilitate the verification of the reported HKHS commodity codes, which have all along been the main source of mis-reporting errors.

The flowchart for the whole procedure of training and applying text analytics is shown in Figure 3.1.

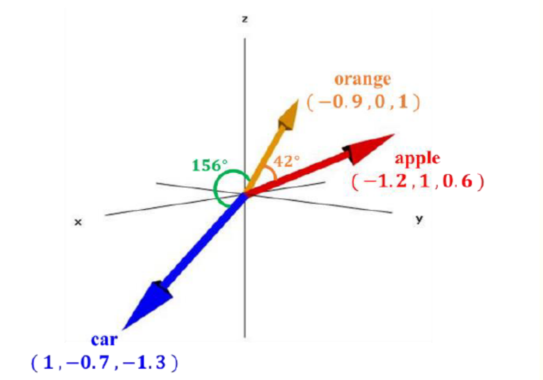
Figure 3.1: Flowchart


Step 1. Train word embedding models with commodity description from around 2 million TDECs to obtain numerical representation of commodity description.
Commodity description (Historical)
Data Cleansing
Cleansed commodity description
Train
Word2vec Model
Obtain
Numerical representation of commodity description
Step 2. Train Neural Network Model to classify commodity description into different HKHS codes.
Commodity description (Historical)
Data Cleansing
Cleansed commodity description
Word2vec Model
Transform to
Numerical representation of commodity description
Train
Gated Convolutional Networks (GCNN) Model
Evaluate model prediction performance
Step 3. Apply the trained GCNN Model to predict HKHS codes based on the commodity description.
Commodity description
Data Cleansing
Cleansed commodity description
Word2vec Model
Transform to
Numerical representation of commodity description
Input
Gated Convolutional Neural Network (GCNN) Model
Predict
Probability distribution of HKHS codes
The HKHS code with the HIGHEST prediction probability
Final HKHS prediction

3.1 Text Vectorisation

Word is the basic unit of text analytics. The purpose of text vectorisation is to transform words into numeric vectors, i.e. word vectors, as inputs for building Neural Network (NN) models. Word embedding is nowadays generally used, which transforms words into word vectors in continuous vector space. Instead of transformation into one-hot vector format, which would inevitably require an extremely long vector of dimension equal to the number of distinct words in the text corpus, word embedding process can generate vectors of much lower dimension, usually ranging from only 50 to 300. Another merit of word embedding vectors is their ability to encode semantic relationships amongst words. Figure 3.2 shows the word embedding vectors of the words “apple”, “orange” and “car” in a 3-dimensional vector space. The included angle between a pair of word embedding vectors reflects their semantic similarity; whereas one-hot vectors are by definition orthogonal to each other and thus all such word vectors are equidistant. Such nice property of word embedding vectors enables NN models to more easily differentiate words with different meanings.

Figure 3.2: Plot of word embedding Vectors



3.2 WordPiece Tokenisation

Word2vec [3], a word embedding model that computes the word vector for each of the *N* distinct vocabularies in a corpus, is used throughout this paper. In Word2vec, vectors are only trained for words with the number of appearances more than a minimum threshold. As such, misspelled or rare words are unlikely to have word embedding vectors trained because of their low appearances (such words are called Out-of-Vocabulary (OOV) words). A plausible choice is to perform word embedding at sub-word level since sub-words are more likely to be shared by a wider array of words. For example, the organic chemical compounds "butane" and "butanone" share the sub-word "butan", which suggests that they may be semantically similar.

WordPiece Tokenisation [4] is an algorithm to segment words into sub-word level. It can effectively handle misspelled or rare words in texts through breaking down a word into sub-words that are commonly shared by other words. The sub-words for segmentation are determined by using a greedy approximation to find a combination of sub-words that would maximise the likelihood of a language model.

Table 3.1: Examples which can be successfully classified using WordPiece Tokenisation

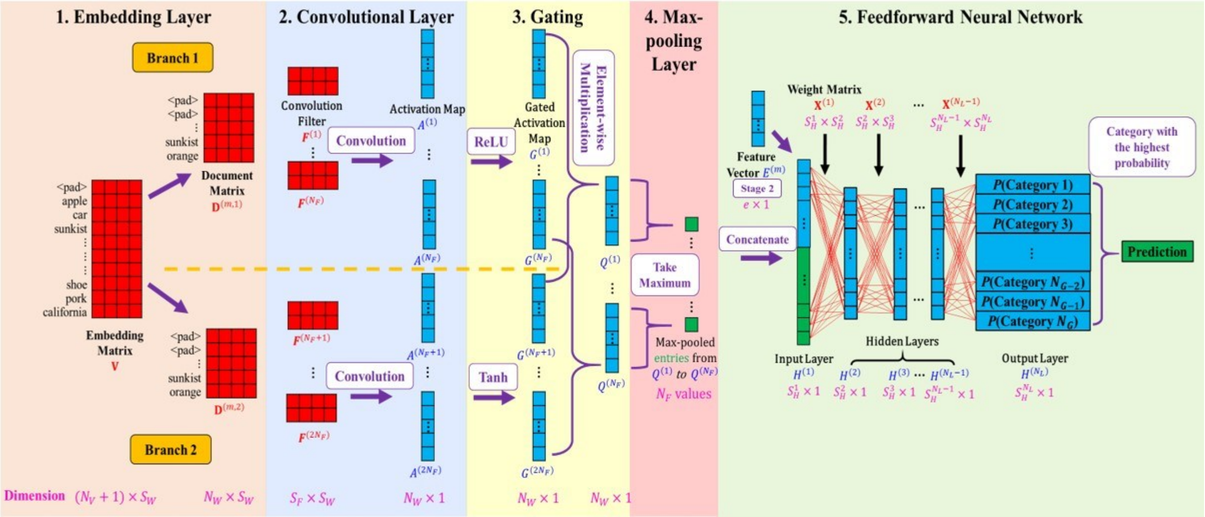
Commodity Description	Sub-words formed after WordPiece Tokenisation	HKHS Codes Classification (Prediction probability)
'Firee Extingguisher'	['fire', '##e', 'ex', '##ting', '##gui', '##sher']	8424 1000 - Fire extinguishers (71.3%)
'CAPACITOR 鋁介 '	['cap', '##ac', '##itor', ' 鋁 ', ' 介 ']	8532 2200 - Fixed capacitors nesoi, aluminum electrolytic (89.3%)

The first example in Table 3.1 is a misspelled text for "Fire Extinguisher". The misspelled words "Firee" and "Extingguisher" are OOVs by Space Tokenisation (using space character to delineate words) whilst they can be segmented into sub-words by WordPiece Tokenisation for correct classification. Our tokenisation can also handle multilingual texts, a feature especially required in Hong Kong since the commodity descriptions provided by traders can be in Chinese, English or a mix of the two. With our self-developed multilingual tokeniser, multilingual commodity descriptions can also be correctly classified. The concatenated multilingual text "CAPACITOR 鋁介" is separated into sub-words and results in correct classification.

3.3 Gated Convolutional Neural Network (GCNN)

GCNN is one of the several recently proposed machine learning models [5, 6]. We have deployed GCNN with an additional Gated Tanh ReLU Unit (GTRU) between the Convolutional Layer and the Max Pooling Layer in the Convolutional Neural Network (CNN). The outperformance of GCNN is attributed to the convolutional layer, which averages out the insignificant variations while keeping the crucial ones, thereby helping the latter features to stand out more clearly in the model. Also, with the use of GTRU, features that are crucial to the prediction are selected for model training. It also enables faster and more efficient training since the model can easily be parallelised in computation. Figure 3.3 shows the model architecture of GCNN we used.

Figure 3.3: Gated Convolutional Neural Network (GCNN)



3.4 Demonstration and Results

To build GCNN, one year (July 2019 to June 2020) trade declaration data, comprising 2.7 million records, were used and partitioned by the ratio of 8:2 for model training and validation respectively, while data in July 2020 were used as the testing dataset. The training data are used to fit the model parameters, which are the weights matrices in NN model, while the validation data are used for evaluating whether more complex model is required to improve the accuracy and for tuning the model's hyper-parameters to avoid over-fitting of the model. Classification rate was used to measure how well the model was in providing correct predictions. By definition,

$$Classification\ rate = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

Another angle to evaluate model performance is prediction probability, which is the level of confidence of a prediction. By studying the classification rate of predictions at different levels of prediction probability, we can understand the model performance in detail.

We use HS Chapters 71, 84 and 85 as a demonstration. There are 1 169 8-digit codes under these three Chapters.

Table 3.2: Results for GCNN

Performance statistics for the demonstration (At 8-digit HS code level)			
HS Chapter	Prediction probability	Proportion of records within the Chapter	Classification rate
71 (86 8-digit codes)	>90%	18.4% (1 728 records)	98.7% (1 705 records)
	>80%	49.4% (4 632 records)	98.2% (4 548 records)
	>70%	63.1% (5 925 records)	96.4% (5 710 records)
	Overall	**100%** (9 386 records)	**83.7%** (7 856 records)
84 (660 8-digit codes)	>90%	12.7% (4 475 records)	99.4% (4 449 records)
	>80%	33.7% (11 833 records)	98.1% (11 608 records)
	>70%	52.0% (18 268 records)	95.9% (17 523 records)
	Overall	**100%** (35 129 records)	**76.2%** (26 785 records)
85 (423 8-digit codes)	>90%	12.2% (18 094 records)	99.0% (17 908 records)
	>80%	36.4% (53 775 records)	98.1% (52 754 records)
	>70%	48.3% (71 385 records)	96.4% (68 851 records)
	Overall	**100%** (147 902 records)	**77.1%** (114 100 records)

From Table 3.2, GCNN performed satisfactorily with classification rates greater than 75% for all Chapters. Further analysing the predictions by prediction probability, the results were even more promising, with classification rates greater than 95% for records having high prediction probability (greater than 70%). This supports that comparing the HKHS codes reported by traders with the model prediction could help identify potential misreporting.

Stage 2 - Text analytics on more data fields using Multi-label input Neural Network (MNN)

During the development of GCNN, we discovered that the level of clarity of texts declared on trade declarations would significantly affect the model performance. Indeed, the commodity descriptions are free text data without any constraints or limitations in the choice of words or length of the entire phrase. Broad-brush descriptions in one or two words like “jewellery”, “polished diamonds” and “dresses” are not uncommon. Obviously, for these simple descriptions, more than one HKHS codes is possible. The vaguely or even incorrectly pre-labelled training data may mis-feed the model during the training process and affect the precision of the predictions.

Besides, while GCNN aims at highlighting key features in commodity descriptions, less prominent information may be masked in the process, leading to imprecise predictions in some cases. For example, in the text “Semi-precious Stone (Cut & Polished) – Opal”, the word “Semi-precious Stone” possibly masks the effect of “Opal”, while in the text “Stands for Shaver”, the word “Shaver” may mask the effect of “Stands”. Thus, GCNN in some cases cannot reach the correct HKHS codes due to loss of some information. This matches the observation that a large proportion of the discrepancies between reported and predicted codes at 8-digit level are in fact consistent at the higher 4-digit level.

3.5 Multi-label input Neural Network (MNN)

To mitigate the effect of less perfect commodity description data, a Multi-label input Neural Network (MNN) model was developed in 2021 to accept a host of data fields available in individual trade records as model inputs, including not only the textual commodity description, but also the country/territory of origin (CO) and the country/territory of consignment (CC). We developed the MNN model by concatenating Feature Vectors before the feedforward neural network of GCNN (see Figure 3.3). The Feature Vectors were constituted by fixed texts such as CO, CC and company name, which should provide additional information for the classification. In our initial trials, CO and CC were used as extra model inputs.

3.6 Comparison between two models

We compared the accuracy of GCNN and MNN to examine whether MNN has improved the overall performance.

Table 3.3: Results comparison of GCNN and MNN

HS Chapter	**Prediction Probability**	**Model**			
		GCNN		**MNN (with CO & CC)**	
		Classification rate	No. of Correct cases	Classification rate	No. of Correct cases
71 (9 386 cases)	>90%	98.7%	1 705 (21.7%)	99.5%	3 679 (45.8%)
	>80%	98.2%	4 548 (57.9%)	98.5%	5 610 (69.8%)
	>70%	96.4%	5 710 (72.7%)	97.1%	6 419 (79.9%)
	Overall	**83.7%**	**7 856 (100%)**	**85.6%**	**8 035 (100%)**
84 (35 129 cases)	>90%	99.4%	4 449 (16.6%)	99.4%	7 002 (24.3%)
	>80%	98.1%	11 608 (43.3%)	98.6%	13 382 (46.4%)
	>70%	95.9%	17 523 (65.4%)	97.0%	20 019 (69.5%)
	Overall	**76.2%**	**26 785 (100%)**	**82.0%**	**28 812 (100%)**
85 (147 902 cases)	>90%	99.0%	17 908 (15.7%)	99.4%	19 821 (16.9%)
	>80%	98.1%	52 754 (46.2%)	98.5%	53 458 (45.6%)
	>70%	96.5%	68 851 (60.3%)	96.8%	72 557 (61.9%)
	Overall	**77.1%**	**114 100 (100%)**	**79.2%**	**117 144 (100%)**

Table 3.3 shows that MNN outperformed GCNN in two aspects. First, the overall classification rates increased by 1.9% to 5.8% points for all the 3 HS Chapters. More importantly, MNN had a higher predictive power than GCNN, where the proportion of correctly classified cases having high prediction probability (>70%) achieved by MNN was higher than that by GCNN by 1.6% to 7.2% points.

Microscopically, MNN could also solve the problems associated with GCNN arising from ambiguous commodity descriptions. Table 3.4 lists some examples.

Table 3.4: Examples to illustrate the effectiveness of adding CO and CC

Commodity Description	**CO**	**CC**	**Predicted HKHS Code** **By GCNN** **(Prediction probability)**	**Predicted HKHS Code** **By MNN** **(Prediction probability)**
Semi-precious Stone (Cut & Polished) - Opal	AU	IN	7103 9990 - Other precious stones and semi-precious stones (73.7%)	7103 9920 - Opals, otherwise worked (98.2%)
Gold Jewellery (Bangles)	AE	CN	7113 1911 - Articles of jewellery and parts thereof, diamond mounted or set, of gold (54.2%)	7113 1919 - Articles of jewellery and parts thereof, diamond not mounted or set, of gold (78.4%)

After adding CO and CC, words with ambiguous meaning could be correctly identified. It is believed that the trading patterns of commodities from source to destination countries exist in past records have been successfully captured by MNN after incorporating CO and CC. Hence, a portion of simple or ambiguous texts can be dealt with after using MNN.

4. Applications

Amid the outbreak of COVID-19 in early 2020, the established checking process of trade declarations was seriously affected due to unexpected manpower constraint arising from the work-from-home arrangement of government staff as a measure to reduce the risk of spreading the virus. To maintain data quality with a reduced workforce, the text analytics model being developed at the time was applied as a contingency measure to supplement certain manual checking on trade declarations. Trade records under HKHS Chapters 84 and 85 (covering mainly machinery and electrical equipment, and accounted for around 70% of the total trade value and 40% of the number of trade records of Hong Kong) were selected for applying the model with a view to speeding up the checking process.

Subsequent evaluation indicated that, if the RMM was solely used (under the scenario of no reduction in workforce), around 12% of the total trade records would be selected for manual checking, of which 45% (or 5.4% in terms of all trade records) would be identified for amendment. On the other hand, when using the GCNN model, only around 7% of the total trade records were selected for manual checking, of which a high proportion of 90% (or 6.3% in terms of all trade records) required amendments. In other words, less manpower was required in the latter approach, which at the same time could achieve higher effectiveness and thus higher quality.

With the encouraging results achieved, in early 2021, the application of the GCNN model was extended to the checking of trade declarations under 27 HKHS Chapters, which accounted for 88% of the total trade value and 69% of the number of trade records of Hong Kong.

In mid-2021, C&SD started to apply the MNN model to replace the earlier GCNN model for checking the trade declarations under selected HKHS Chapters. Compared with the previous model, the overall classification rates of the MNN model increased by 1.9% to 5.8% points to a level of some 80%. More importantly, the MNN model achieved a higher predictive power. The proportion of cases with high confidence in the prediction result (prediction probability >70%), which defines the group of records for which the predicted commodity code can be meaningfully compared with the reported commodity code to identify potential erroneous cases, was higher than that of the GCNN model by 1.6% to 7.2% points. Besides, the new model was more effective in handling commodity descriptions with ambiguous meaning.

5. Further work

The application of text analytics models in processing trade declarations has successfully materialised the benefits of the new techniques. Plans are in hand to further expand the scope of application to trade records under more HKHS Chapters. The MNN model will also be further enhanced by including new data fields such as trade values and quantities with a view to improving its predictive power.

6. References

[1] R. Shams and R. E. Mercer, "Classifying Spam Emails Using Text and Readability Features", 2013 IEEE 13th International Conference on Data Mining (ICDM), p.657-666, 2013.

[2] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques”, In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), p.79–86, 2002.

[3] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). “Efficient Estimation of Word Representations in Vector Space”, International Conference on Learning Representations: Workshops Track, 2003.

[4] Mike Schuster and Kaisuke Nakajima, “Japanese and Korean voice search”, 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p.5149–5152, 2012.

[5] Dauphin, Y. N., Auli, M., & Grangier, D., “Language Modeling with Gated Convolutional Networks”, 2016 arXiv preprint arXiv: 1612.08083.

[6] Xue, W., & Li, T. (2018). “Aspect Based Sentiment Analysis with Gated Convolutional Networks”, In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), p.2514–2523, 2018.

2020/21 Statistical Project Competition for Secondary School Students

Organising Committee* of the 2020/21 Statistical Project Competition

The 2020/21 Statistical Project Competition (SPC) for Secondary School Students, the 35th round of the Competition since 1986/87, was successfully completed. SPC was jointly organised by the Hong Kong Statistical Society (HKSS) and the Education Bureau. The objective of SPC is to encourage secondary school students to understand the local community in a scientific and objective manner through the proper use of statistics, thereby promoting their social awareness and sense of civic responsibilities.

The SPC has two Sections for participants, namely Junior Section for Secondary 1 to 3 students and Senior Section for Secondary 4 to 6 students. Junior Section participants are required to submit their projects in the form of a poster on one of the following themes: population, district analysis with population statistics or transport and housing, while Senior Section participants in the form of a report with their own choices of themes. In addition to the First, Second, Third and Distinguished Prizes, each Section of the Competition also offers the Prize for the Best District Analysis using Population Census Statistics and Prize for the Best Graphical Presentation of Statistics.

To help interested participants prepare for the Competition, a Briefing Seminar for SPC for 2020/21 was held on 24 October 2020 at the Education Bureau Kowloon Tong Education Services Centre. To uphold the social distancing measure to reduce the risk of infecting the COVID-19, the Seminar was broadcast through the Internet for views by school teachers and students. An online exhibition of past winning projects was also carried out during 26 October 2020 to 8 November 2020.

Encouraging number of entries

Affected by COVID-19, the face-to-face lessons of secondary schools had been partially suspended in 2021, posting great difficulties for school teachers and participating students in discussing and preparing of their statistical projects. Despite the situation, 80 entries and 66 entries were submitted for the Junior Section and the Senior Section respectively from 50 secondary schools. The number of entries and secondary schools were substantially higher than the previous round. Demonstrating the diversity of topics, the entries covered various socio-economic issues of Hong Kong.

Adjudication panel led by Dr. CHEUNG Ka-chun

An adjudication panel, led by the Chief Adjudicator, Dr. CHEUNG Ka-chun of The University of Hong Kong, and comprised some 30 academics from local tertiary institutions and statisticians working in the Government, was set up for the Competition. Panel members scrutinised all the received projects stringently, shortlisted the more outstanding entries, and interviewed students of the shortlisted projects before determining the winning teams of the various awards. To reduce the risk of infection, this round of panel interviews was conducted through Zoom meeting.

Prize Presentation Ceremony

The Prize Presentation Ceremony for the SPC 2020/21 took place on 5 June 2021 at the Auditorium of the Hong Kong Federation of Youth Groups Building. Professor Alan WAN Tze-kin, President of HKSS, Ms Marion CHAN Shui-yu, Commissioner for Census and Statistics, and Mr Joe NG Ka-shing, Principal Assistant Secretary of Education Bureau were invited to the Ceremony to present prizes and trophies to the winning teams.

Regarding the results of the Competition, students of Stewards Pooi Kei College, who used official statistics to study population changes by district, won the First Prize of the Junior Section. Students of Diocesan Girls' School won the Second Prize, while students of Pooi To Middle School won the Third Prize. The Prize for the Best Graphical Presentation of Statistics and the Prize for the Best District Analysis using Population Census Statistics were won by the First and the Third teams respectively.

As for the Senior Section, the statistical report from students of Holy Family Canossian College was appraised as the best amongst all the projects. They applied official statistics to analyse the trend and influence of fertility rate in Hong Kong. Students of another team of Holy Family Canossian College won the Second Prize while students of Stewards Pooi Kei College won the Third Prize. Students of another team of Stewards Pooi Kei College won the Prize for the Best Graphical Presentation of Statistics and the Prize for the Best District Analysis using Population Census Statistics.

Gratitude

The Organising Committee would like to express sincere gratitude to the patrons of the Competition, Ms Marion CHAN Shui-yu, Commissioner for Census and Statistics, and Mrs HONG CHAN Tsui-wah, Deputy Secretary for Education, for their support to the event. Special thanks to helpers and the Adjudication Panel.

*Organising Committee for the 2020/21 SPC:

Mr Raymond TSE	Census and Statistics Department
Mr CHAN Sau-tang	Education Bureau
Mr Alex LI	Census and Statistics Department
Miss Carmen LO	Census and Statistics Department
Mr Michael CHU	Census and Statistics Department
Mr Oliver HO	Census and Statistics Department
Mr Proton NG	Census and Statistics Department
Mr Stanley TSANG	Census and Statistics Department

News Section

◆ Personnel Changes (New Appointments, Promotions and Retirements)

※ Professor Guodong LI of Department of Statistics and Actuarial Science of The University of Hong Kong (HKU) has promoted to Professor.

※ Dr. Eddy LAM of Department of Statistics and Actuarial Science of HKU has been appointed as Associate Dean of Faculty of Science (student affairs).

※ Professor Guosheng YIN of Department of Statistics and Actuarial Science of HKU was elected Fellow of Institute of Mathematical Statistics (IMS) (2021).

※ Dr. Lequan YU, Dr. Yuan CAO, and Dr. Kai HAN joined Department of Statistics and Actuarial Science of HKU as Assistant Professors.

※ Dr. Ben DAI joined the Department of Statistics of The Chinese University of Hong Kong as Assistant Professor.

※ Dr. LAM Shu-yan has promoted to Associate Professor, Department of Mathematics, Statistics and Insurance of The Hang Seng University of Hong Kong (HSUHK).

※ Dr. TSUI Wing-yan joined the Department of Mathematics, Statistics and Insurance of HSUHK as Lecturer.

◆ Master of Science in Artificial Intelligence launched in HKU

Master of Science in Artificial Intelligence (AI), jointly organised by Department of Mathematics (Host), Department of Statistics & Actuarial Science and Department of Computer Science, was launched in HKU. The programme provides students with foundational principles and knowledge in AI, and develops their practical skills and capabilities in applying AI to solve real world problems with ethical awareness. The programme is designed to prepare graduates for a wide range of career opportunities in AI-related fields. Details of the programme are available at https://hkumath.hku.hk/web/mscai/mindex.php.

◆ Professional Development Course at Fu Hong Society

Last December, Fu Hong Society (扶康會) invited HKSS to organise a professional development course on “Research Methodology and Data Analysis” to equip the occupational therapists working in their Society with adequate knowledge for evidence-based practice.

Professor May WONG, the Consultation Services Secretary of HKSS, delivered a full-day course on 17 December 2021. The course covered various topics: design of observational and clinical interventional studies, measurement validity and reliability, descriptive and inferential statistics, confidence interval and hypothesis testing, parametric and non-parametric tests, comparison of groups, correlation and regression. Research publications in occupational therapy were used as examples in the course.

In total, 14 occupational therapists attended the course, and favourable responses were received.

Research Methodology and Data
Analysis for Occupational Therapist
Data Analysis and Research Design
JOYFU

◆ CityU Day of Biostatistics

The HKSS co-hosted the “Biostatistics Day” with the (new) Department of Biostatistics of the City University of Hong Kong on 21 March 2022. The organising committee had hoped the event to be an in-person meeting but changed course as Omicron raged on. The (virtual) meeting was chaired by Professor Ian McKeague, Head of Biostatistics at City University, and Professor Alan WAN, President of HKSS.

The meeting began at 2:00pm with Professor Ian McKeague delivering a welcoming speech that was immediately followed by a short introduction of the HKSS by Professor Alan WAN. Then Professor Chun Sing LEE (Dean of Science at CityU) talked about of the key developments leading to the establishment of the Department of Biostatistics at CityU. This was followed by Professor Ngai Hang CHAN (Chinese University of Hong Kong) who gave an overview the role of Biostatistics in Health Science and Medicine.

The opening was followed by three scientific papers presented by Dr Can YANG (HKUST): *A fast and accurate method for genetic risk prediction by leveraging Blobank-scale data*, Professor Guosheng YIN (HKU): *Statistical Learning with Observational Studies*, and Dr. Tony SIT (CUHK) and Dr. George CHU (CityU): *Censored interquantile regressoion with time-dependent covariates*. These papers contain a mixture of methodological developments in Biostatistics and applied research to real life data. Originally the Committee had invited Mr Alan CHEUNG, Chief Statistician at the Hong Kong Hospital Authority (HKHA), to speak on the COVID-19 prognostication model used by the HKHA. Unfortunately, Mr CHEUNG could not attend due to work commitment.

In conclusion, the Biostatistics Day at CityU was an enjoyable event. The HKSS looks forward to co-hosting a similar academic event to be organised in 2023.

◆ The HKSS-John Aitchison Prize in Statistics

The HKSS is pleased to announce that a new prize in Statistics is being established from money donated by universities and other organisations in Hong Kong, in memory of Professor John Aitchison, one of 20th century's best-known statisticians, Chair Professor of Statistics at the University of Hong Kong between 1976 and 1989, and founder and first President of the HKSS.

The objective of the prize is to reward excellence in PhD research. The Prize is open to those who, on the closing date of applications, have within the past two years, full-time student status for a PhD degree in Statistics or a closely-related discipline at a Hong Kong university.

A working group comprising of Professor Ngai Hang CHAN, Professor Wai Keung LI and Professor Alan WAN has been formed to formulate the criteria for the award and related polices. Details of the award will be announced in due course. It is expected that the first award will be made in 2023.

◆ Tour to Hoi Ha Wan Marine Park

HKSS and the Association of Government Statisticians (AGS) originally planned to co-organise a half-day tour to Hoi Ha Wan Marine Park as social activity on 23 January 2022. However, the tour was cancelled due to the social distancing measures and the latest situation of COVID-19 epidemic.

◆ Workshop on Data Visualisation

The Workshop on Data Visualisation through PowerBI was originally planned to be conducted on 14 January 2022 in PolyU, with around 30 participants. The workshop was cancelled due to the latest situation of COVID-19 epidemic.