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 46 No.1
March 2024

Editor’s Foreword

Dear Members,

It is with deep gratitude that I acknowledge the significant contributions and support provided by Dr. Billy LI over the past eight years in the preparation of this bulletin. His dedication and service are truly appreciated. Furthermore, I would like to extend a warm welcome to Ms. Angela YEUNG who serves as the Secretary of the Editorial Board in 2023/24.

This issue of the Bulletin features three articles that promote discussions on further developments in statistics. The Accreditation Board briefs us the latest development of accreditation and mutual recognition of Graduate Statistician and Data Analyst professional qualifications between the Hong Kong Statistical Society (HKSS) and the Royal Statistical Society (RSS). The Organizing Committee of the 2022/23 Statistical Project Competition also takes this opportunity to report on the successful completion of the Competition.

I would like to express my deepest gratitude to all the contributors to this Bulletin and the esteemed members of the Editorial Board. Your expertise, dedication, and efforts have been instrumental in the creation of this publication.

Thank you for your continued support and engagement with the HKSS Bulletin. Your active participation and readership are invaluable in promoting the dissemination of statistical knowledge and fostering a vibrant statistical community.

Edmond CHAN

CONTENTS

(Vol. 46/No.1, March 2024)

President’s Forum

Professor Alan WAN Tze-kin

As we embark on the new year of 2024, I would like to provide you with an update on the recent activities of the Hong Kong Statistical Society (HKSS).

The Society plays a vital role in advancing the field of statistical sciences and their applications in Hong Kong and beyond. For over 45 years, we have brought together researchers, academics, and industry professionals to foster collaboration, stimulate research, and promote best practices. Through our work, we aim to ensure that statistics continue to provide valuable insights that benefit society at large.

In November 2023, we had the privilege of co-hosting a seminar with the Royal Statistical Society (RSS). Professor Rachel HILLIAM, Vice President of the RSS, delivered some thought-provoking perspectives on professionalising data science to an engaged audience at City University of Hong Kong. I would like to express my gratitude to our members who attended this event and extend our appreciation to the RSS for their collaborative efforts. The exchange of knowledge with other statistical bodies strengthens our profession and enhances our collective expertise.

The HKSS - John Aitchison Prize in Statistics 2024 received high quality submissions by the November 17, 2023 deadline. The panel, after careful deliberations, awarded the prize to Dr. Guohao SHEN (2022 PhD Graduate at the Chinese University of Hong Kong) for the co-authored paper titled “Deep Nonparametric Regression on Approximate Manifolds: Non-asymptotic Error Bounds with Polynomial Pre-factors”.

In the coming months, the Society will continue its efforts to support our members, foster innovation, and apply statistical methodologies to address various issues faced by Hong Kong. Please stay tuned for announcements regarding upcoming seminars, networking opportunities, and other initiatives. I would like to extend my sincere appreciation to all our members for your dedicated involvement and unwavering support in fulfilling the Society's important mission.

Integrated Time Series Analysis and Short-Term Forecasting On Carbon Credit Trading In China

Ricardo K WU, Department of Systems Engineering and Engineering Management, CUHK

Chun Man CHAN, Department of Statistics, CUHK

1. Background

In this article, we are playing the role of a Chinese traditional factory. It is predicted that the carbon quota for our factory will be run off within 3 months, so we are facing a dilemma at the moment: (1) purchase credit now, or (2) purchase credit within 3 months. A time series analysis and forecasting are needed for us to reduce our cost by referencing the past 7 years of data.

1.1 Situation in China

Starting in 2013, China started pilot testing of carbon credit trading in seven cities, such as Guangdong, Beijing, and Shanghai. After testing for a few years, a national carbon emissions trading market under the carbon trading scheme was finally initiated in 2021. This action also is an important step for realizing China government's goal, which is to become carbon neutral before 2060.

In February 2021, the city government of Shanghai implemented the scheme to 2,225 companies. As predicted by the Shanghai Environment and Energy Exchange, there would be 5 billion tons worth of carbon credits and around 8,000 to 10,000 companies would participate in the scheme.

In Guangzhou, until the beginning of December 2022, the national trading volume of the emissions exchange had exceeded 200 million tons.

1.2 New economic opportunities

Every country has its own quota for carbon emissions, which is set by the government in accordance with the emissions target committed under the international convention. Then the quota will be assigned to different companies. Each quota represents that a company could emit a ton of greenhouse gas. The company can reduce the emissions of carbon dioxide by introducing relevant technology, then they may sell the excessive quota to other companies through the carbon credit trading market. On the contrary, those firms which cannot reduce the emission of carbon dioxide could buy quotas from other enterprises through the carbon market. In other words, selling carbon permits is an opportunity for companies to increase their revenue. Tesla's carbon credit sales reached a new record of US$1.78 billion in 2022. When the firm gains more profit, the government may earn more profit tax to increase financial reserves, then the government has more funding to relieve livelihood issues.

2. Introduction

2.1 Methods

Most conventional statistical models analyze different types of data by linear and probabilistic statistical inference, with an aim of finding out the most likely parameters for the model. Though the diagnostics can be extensive, conventional statistical models have many limitations in analyzing other types of data and the process requires direct intuitions about sampling distributions and statistical assumptions, while machine learning provides a totally different way. The algorithmic and some probabilistic approach can extract patterns from messy data without much domain knowledge. It sometimes performs better than conventional methods, especially for non-linear and high dimensional or out-of-sample data. In this article, both traditional and innovative methods are adopted to make predictions.

2.2 Methodology

The data of average monthly unit price of carbon credits is first decomposed into traditional statistical parts consisting of trend, seasonality, and noises; the cyclic element is excluded due to the characteristic of the trend. Seasonal Extraction in ARIMA Time Series (SEATS), Seasonal and Trend decomposition using Loess (STL), and Multiple Seasonal-Trend decomposition using LOESS(MSTL) are adopted to help decompose the data and inspect the components of the data.

After decomposition, some naive forecasting methods like simple moving averages (SMA), centered moving averages (CMA), weighted moving averages (WMA), and exponential smoothing (EMA) are used for naive and simple forecasting, which are set to act as indicators for later modeling.

Besides, we use a trend projection model including linear regression and polynomial regression. We first find out the coefficient of the regression line and fit the data to do the forecasting. For polynomials, we first find out the best degree of the model and then do the forecasting. We also use Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) model. We first test the ARCH effects, then select the best model and do the forecasting.

2. Time series decomposition

As an emerging carbon trading market, we wonder about its performance.

3.1 Exploratory Time Series Analysis

There are 1,797 daily raw data defined as $Y = \{Y_1, Y_2, . . . , Y_{1797}\}$ collected from Guangzhou Emission Exchange in the format shown below.

Date	Type	Opening	Closing	Max	Min	%Change	Volume	Amount

As we aim to analyze the unit price per month and forecast the future trend due to missing data from the low monthly trading frequency, we do the operation to get the unit price by summing up Volume$_{day}$ and Amount$_{day}$ of the month to get the average price in RMB per volume defined as $\frac{Volume_{day}/Amount_{day}}{n_{month}}$. After the operation, the raw data are transformed into 96 monthly data defined as $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_{96}}\}$, the time series $\mathbf{X}$ is visualized in Figure 1.

After analysis, multiplicative model without seasonality is used to undergo decomposition.

$$\ln(X_t) = \ln(T_t) + \ln(N_t) \qquad N_t \sim N(0, \sigma^2) \qquad (1)$$

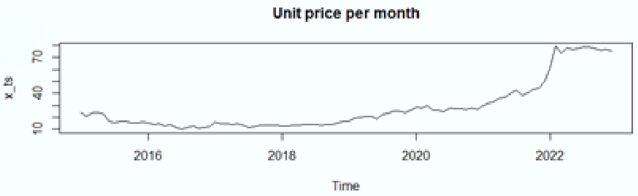
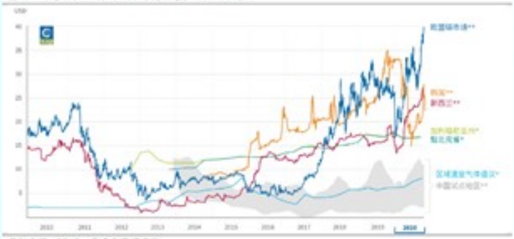


Figure 1: monthly unit price from 2018 to 2022

3.2 Decomposition

图 8: 近十年一级和二级市场碳价走势 (USD/tCO$_2$)

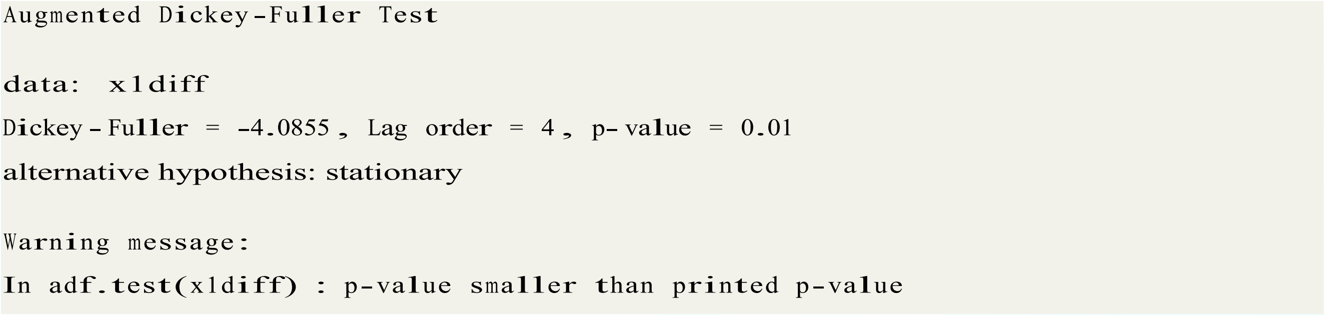


数据来源: ICAP 华泰期货研究院

Figure 2: Carbon credit price

By checking the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, the time series is not stationary. A second-degree differencing is suggested by ***ndiffs()***.

Then the ACF of X$_t$(1 − B) is checked, and the Augmented Dickey-Fuller test is used to check whether the differenced data is stationary or not.



```
Augmented Dickey-Fuller Test

data:  x1diff
Dickey-Fuller = -4.0855, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(x1diff) : p-value smaller than printed p-value
```

As the output shows that the p-value is smaller than α = 0.05, we cannot reject the null hypothesis. Thus, $X_t(1 - \mathbf{B})$ is stationary.

The trend is then got by calculating X_t - $X_t(1 - \mathbf{B})$, the smoothed trend is taken by applying filter [1/24, 1/12, 1/12, 1/12, 1/12, 1/12, 1/12, 1/12, 1/12, 1/12, 1/12, 1/12, 1/24]

3.2.1 Preparation for forecasting - HMA

Hull Moving Average (HMA) is used to see whether the combination with the models would give better performance. The HMA, developed by Alan Hull, is a fast and smooth-moving average. The HMA almost eliminates lag altogether and manages to improve smoothing at the same time. It is commonly adopted in stock analysis as an advanced industry smoothing method.

Considering our situation that we need to long carbon credits within three months with a large volume, while the unit price seems to become flat and may decrease, the smoother should be able to catch the inflexion signal to reduce the cost. Thus, a HMA with a short period is carefully chosen.

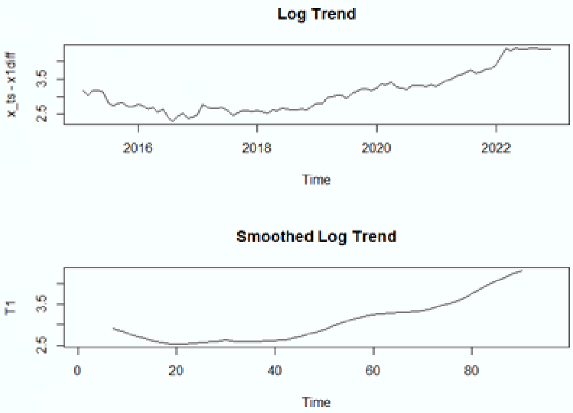
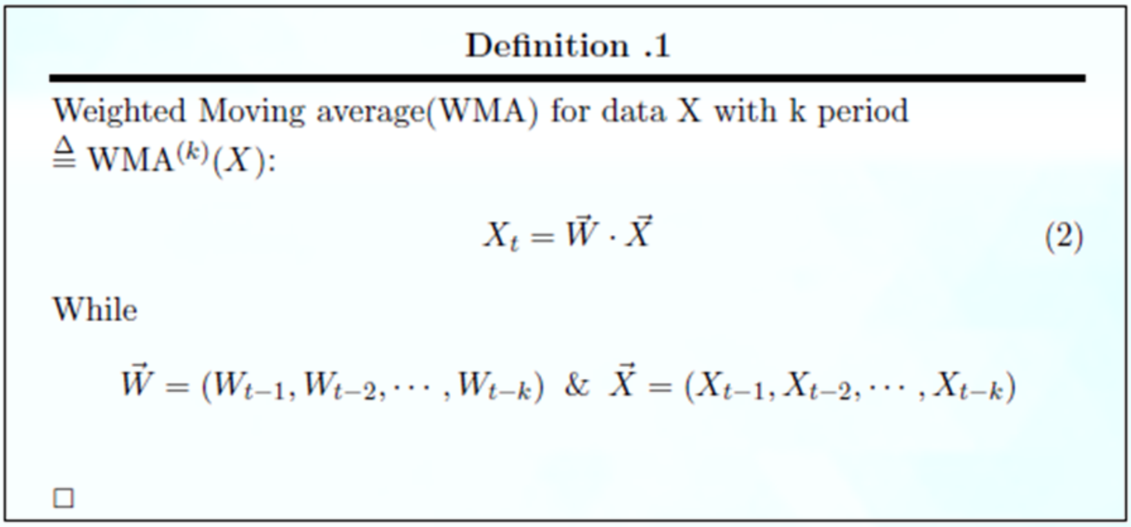


Figure 3: Log trend and smoothed log trend



Definition .1

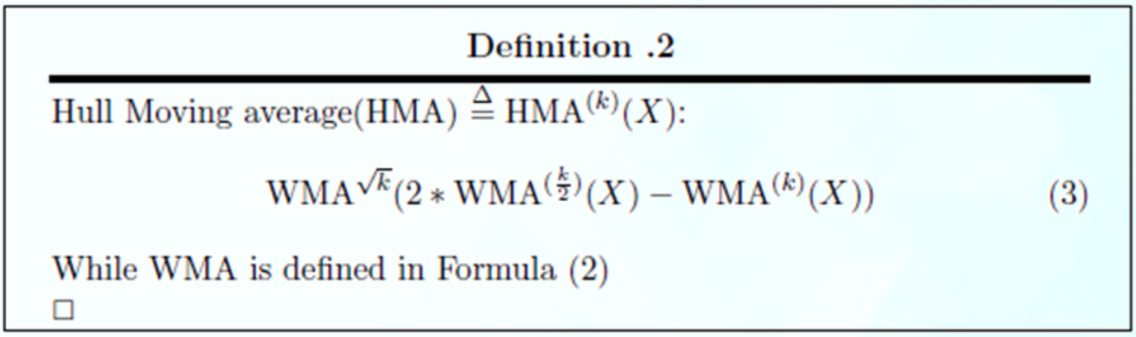
Weighted Moving average(WMA) for data X with k period $\triangleq \mathrm{WMA}^{(k)}(X)$:

$$X_t = \vec{W} \cdot \vec{X} \tag{2}$$

While

$$\vec{W} = (W_{t-1}, W_{t-2}, \cdots, W_{t-k}) \;\; \& \;\; \vec{X} = (X_{t-1}, X_{t-2}, \cdots, X_{t-k})$$

□



Definition .2

Hull Moving average(HMA) $\triangleq \mathrm{HMA}^{(k)}(X)$:

$$\mathrm{WMA}^{\sqrt{k}}(2 * \mathrm{WMA}^{(\frac{k}{2})}(X) - \mathrm{WMA}^{(k)}(X)) \tag{3}$$

While WMA is defined in Formula (2)
□

4. Data centric models

This section is for machine learning models and other modern models in complement to the conventional statistical models.

4.1 Naive machine learning

4.1.1 Cross validation

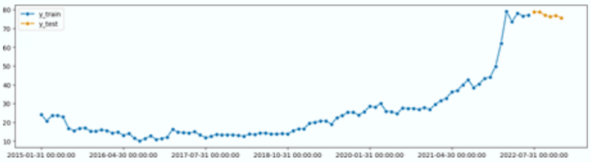


Figure 4: Temporal train test split (6 splits)

In this part, the data is split into 3 parts: training set, cross-validation set, and test set. Short data splitting is chosen to forecast a short period of data. For convenience, a Python library called Sktime is introduced. It is an open-source Python toolbox for machine learning with time series funded by the UK Economic and Social Research Council, the Consumer Data Research Centre, and The Alan Turing Institute. It extends the sci-kit-learn API to time series tasks. The function temporal train test split with 6 splits is used to split the data which means the last 6 data are used for validation. The split data is shown in Figure 4. ExpandingWindowSplitter (initial window=22, step length=13,fh=6) is used as a cross-validation splitter.

Then, Naive forecaster (sp=5) is used as the base model, the plot with 90% confidence in Figure 5 shows the in-sample prediction for cross-validation.

After fitting, AutoETS is the best basic model in Sktime with the performance shown in Figure 6, the plot is bounded by 90% confidence interval.

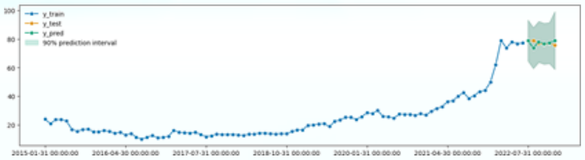


Figure 5: Naive forecaster (sp=5,CI=0.9) for cv

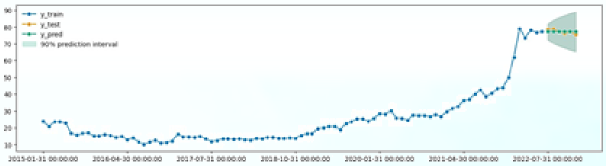


Figure 6: AutoETS (CI=0.9) for cv

Table 1: MSE of base models

Metric	Naive	AutoETS
MSE	6.587	1.456

AutoETS outperforms Naive forecaster for the selected dataset in cross-validation.

4.1.2 Prediction

After cross-validation and hyperparameters tuning, the prediction for the out-of-sample 6 data by naive forecaster and AutoETS is shown in Figure 7 and Figure 8.

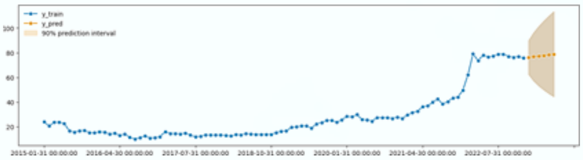


Figure 7: Naive forecaster(sp=5,CI=0.9) for cv

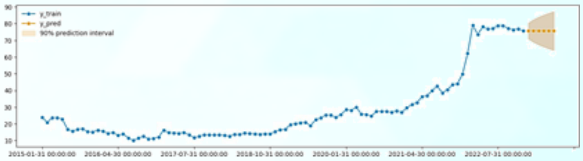


Figure 8: AutoETS (CI=0.9) for test

Apparently, AutoETS performs better with short prediction boundary and small error. The MSE for predicted data is shown below. However, AutoETS shows a flat trend with the same output for both January and February data.

Table 2: MSE of base models

Model	MAE	MSE
Naive	3.85	8.16
AutoETS	0.479	0.12

4.2 Neural network

For this part, long short-term memory (LSTM) and gated recurrent unit (GRU) are chosen to check their performance on the selected dataset. Due to the complexity and poor interpretability, only the training performance is shown here.

We use the min-max scalar to do feature scaling so that the algorithm can converge faster with less loss. The parameters for LSTM is: input dim = 1, hidden dim = 32, num layers = 4, output dim = 1, num epochs = 100, loss = MSE, learning rate = 0.005. The error scores are Train Score: 5.11 MSE, CV Score: 4.58 MSE. The performance graph is shown in Figure 9 while the in-sample cv dataset prediction is shown in Figure 10. In Figure 9, the prediction shows a flat trend. By checking the values, there are minor differences between them.

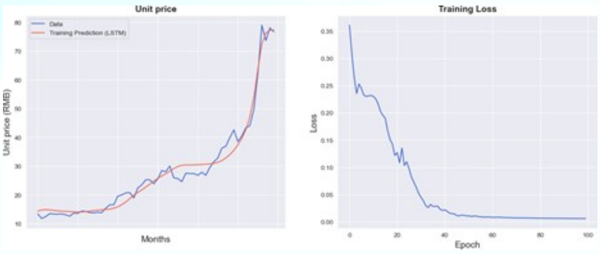


Figure 9: LSTM training performance

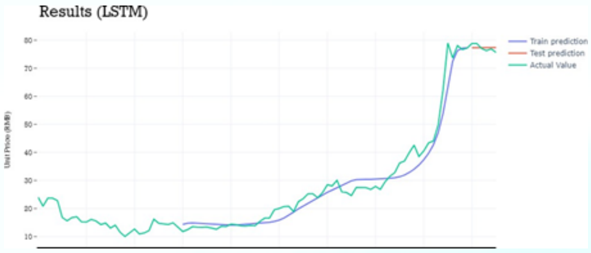


Figure 10: LSTM performance

Then we use the same hyperparameters to check the GRU model, which is also a neural network model for sequential data with a simpler architecture than LSTM. The performance is shown in Figure 11 while the in-sample cv dataset prediction is shown in Figure 12.

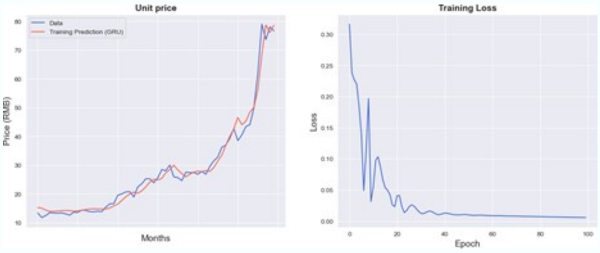


Figure 11: GRU training performance

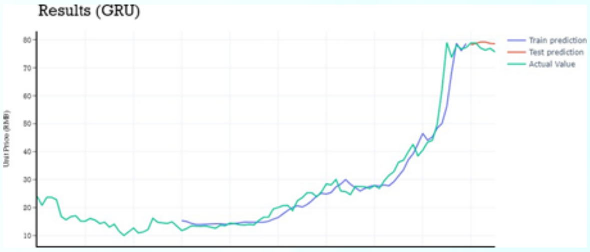
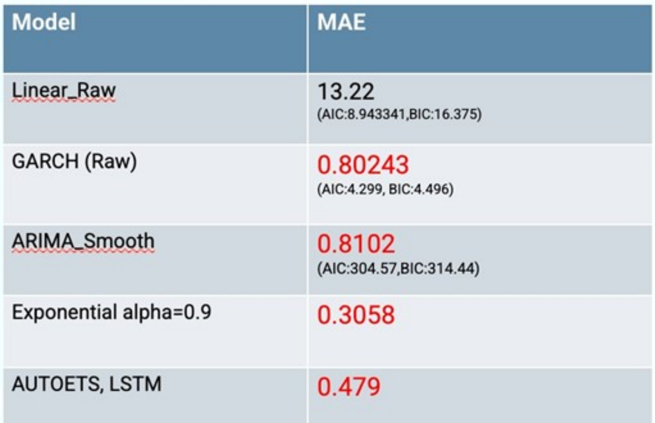


Figure 12: GRU performance

5. Conclusion

After comparison among the models, the models with outstanding performance are listed in Figure 13.



Model	MAE
Linear_Raw	13.22 (AIC:8.943341,BIC:16.375)
GARCH (Raw)	0.80243 (AIC:4.299, BIC:4.496)
ARIMA_Smooth	0.8102 (AIC:304.57,BIC:314.44)
Exponential alpha=0.9	0.3058
AUTOETS, LSTM	0.479

Figure 13: Final comparison

Though models' performance differs in smooth and raw data, it is believed the fluctuation is mainly due to the smooth original trend of the raw data. Unlike other countries, the selected data cannot show the seasonal fluctuation and rapid trend exhibited by other markets. However, the models still can give a satisfying result in this analysis.

References

[1] China in carbon trading experiment. (2013b, June 18). BBC.

About us - cnemission. (n.d.). https://www.cnemission.com/article/gywm/201907/20190700001675.shtml COP26climate

[2] Change and carbon trade. (2021, July 17). BBC.

The first carbon exchange has been launched... (n.d.-b). "https://asia.nikkei.com/Spotlight/Environment/Climate-Change/China-s-national-carbon-Trading-market-eyes-June-debut-in-Shanghai

[3] China's national carbon trading market eyes june debut in shanghai. (2021, March 28). Nikkei Asia.

[4] Zach. (2022b, April 19). Polynomial regression in R (step-by-step). Statology. https://www.statology.org/polynomial-regression-r

[5] Bevans, R. (2023, June 22). Linear regression in R: A step-by-step guide examples. Scribbr. https://www.scribbr.com/statistics/linear- regression-in-r

[6] 4.3 - residuals vs. predictor plot: Stat 501. PennState: Statistics Online Courses. (n.d.). https://online.stat.psu.edu/stat501/lesson/4/4.3

[7] F-tests and nested models. (n.d.-b). https://www.rose hulman.edu/class/ma/inlow/Math485/ftests.pdf

[8] Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2018). Short-Term Forecasting: When Do Statistical Models Outperform Judg- ment? International Journal of Forecasting, 34(1), 3-16. doi: 10.1016/j.ijforecast.2017.06.004.

[9] L, J. (2023, January 27). Tesla Carbon Credit Sales Reach record $1.78 billion in 2022. Carbon Credits. https://carboncredits.com/tesla- carbon-credit-sales-reach-record-1-78-billion-in-2022/

[10] Tan, S. T. Applied Mathematics for the Managerial, Life, and Social Sciences, latest edition, Brooks / Cole, Cengage Leaming.

[11] Anderson, D.R., Sweeney, D. J., Williams, T. A. Quantitative Methods for Business, edition, Thomson South-Western.

[12] Garch. RPubs. (n.d.). https://rpubs.com/Sharique16/garch

[13] Zhang, A., Lipton, Z., Li, M., Smola, A. J. (2024). Dive into deep learning. Cambridge University Press.

An analysis of Hong Kong's monthly sewage flow by a time series approach

Ching Hin YEUNG and Chun Man CHAN
Department of Statistics, CUHK

1. Background and introduction

Sewage monitoring has been considered an efficient tool for epidemiological surveillance since the outbreak of COVID-19 (Bar-Or et al., 2022) to the extent of revealing hidden prevalence rates under asymptomatic infection as well as the spatial distribution estimation, etc. (Ng et al., 2023). As proved efficient by Hendriksen et al. (2019) and Bar-Or et al. (2022), mass-scale surveillance of sewage has inevitably become a common practice in which European countries facilitated their implementations under the Urban Waste Water Treatment Directive (UWWTD) in parallel to the US's practices of antimicrobial resistance surveillance in treatment plants by the Centers for Disease Control and Prevention (CDC) (Larsson et al., 2022). For the case of Hong Kong SAR, the Government cooperated with local institutions on a city-wide full-scale interactive application of a tailor-made sewage surveillance programme. They discussed the feasibility and implementation of real-time COVID-19 wave monitoring and combating using such a programme (Ng et al., 2023).

The total volume of flow and seasonality of sewage production are two of the main concerns amid the sewage sampling from sites for surveillance (Gibson et al., 2012), (Rogawski McQuade et al., 2023) because sewage quantity and quality differ by season and even hourly interval. To facilitate sewage sampling in Hong Kong for surveillance, it is necessary to perform a local sewage flow forecast. This article forecasts Hong Kong's monthly sewage flow spanning a one-year interval from June 2023 using the Autoregressive Integrated Moving Average (ARIMA) algorithm and provides insights into sewage sampling.

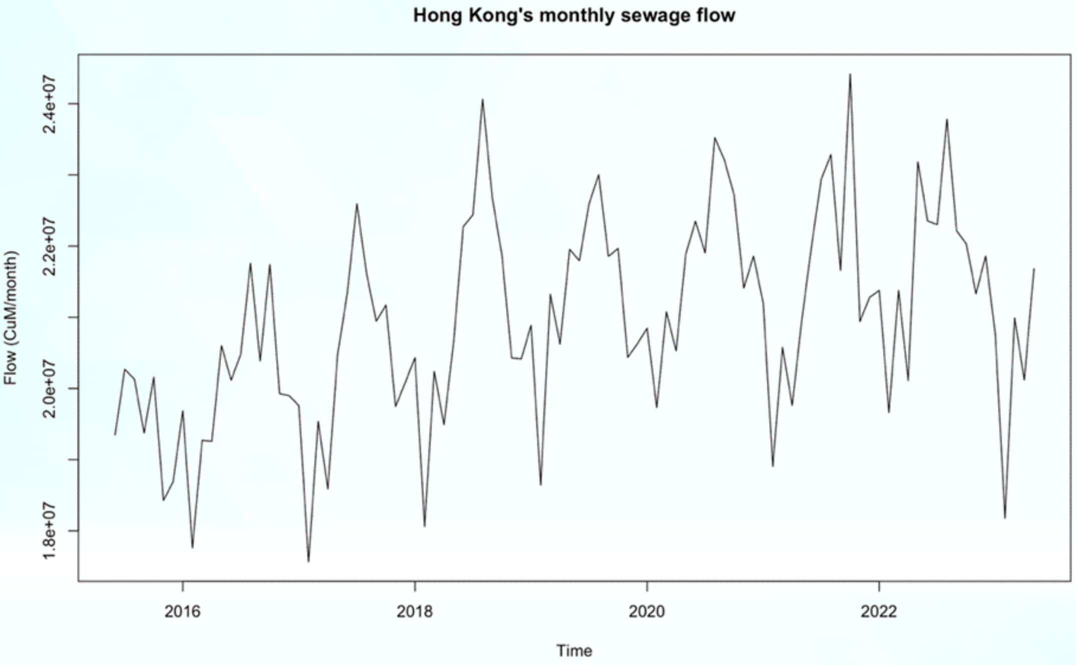
2. Literature review

A couple of researches were conducted regarding the sewage flow forecast. Wąsik and Chmielowski (2016) adopted the Holt-Winters approach in predicting daily inflow in the plant of Nowy Sącz, Poland. Abunama and Othman (2017) utilized the ARIMA model for future sewage inflow in the plant of Selangor, Malaysia. In addition, a sewage-related study conducted by Man et al. on the Chemical Oxygen Demand (COD) load (2019) forecasted the amount of oxygen needed to oxidize organics in sewage. They embraced the use of Autoregressive Moving Average (ARMA) while incorporating Vector Autoregressive (VAR) algorithm. Meanwhile, Man et al. (2019) also studied the Back-Propagation Neural Network (BPNN), Least-Squares Support Vector Machine (LS-SVM), and Genetic Algorithm Back-Propagation Neural Network (GA-BPNN) as contrasting cases.

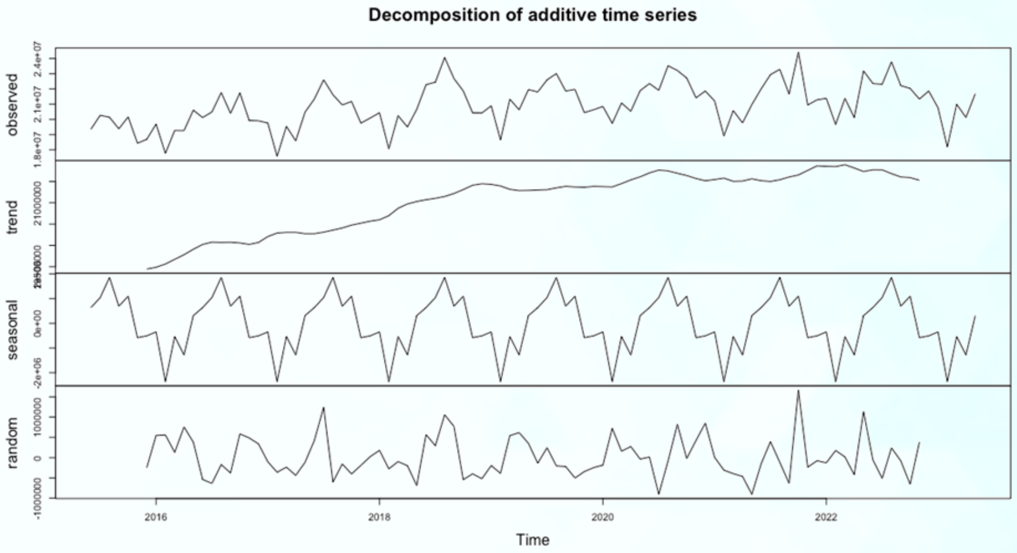
3. Data Processing

This study retrieved the laboratory data and sewage flow dataset released by Hong Kong's Drainage Services Department (DSD). The dataset consists of the daily volume of sewage flow together with the particle contents, etc. from each of the seven sewage treatment plants in Hong Kong spanning from 1st January 2007 to 31st May 2023. Data processing was performed by aggregating individual daily sewage flow from seven plants into daily total sewage flow, followed by a further by-month aggregation.

4. Decomposition



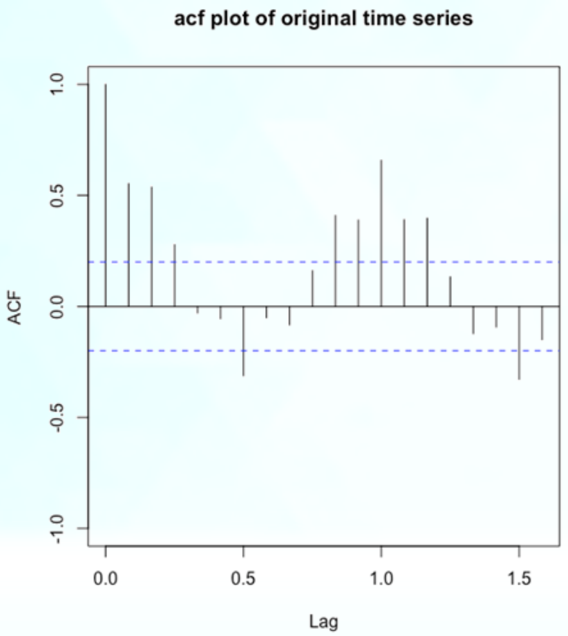
The time series plot shows relative stable fluctuations and a presence of neither exponential nor logarithmic growth tendency. Additive decomposition: $Y_t = T_t + S_t + R_t$, where Y_t is the observation of time series at time t, T is the trend component, S is the seasonal component and R is the residual, will be an ideal way of decomposition in this case and the outcome is presented as follows:

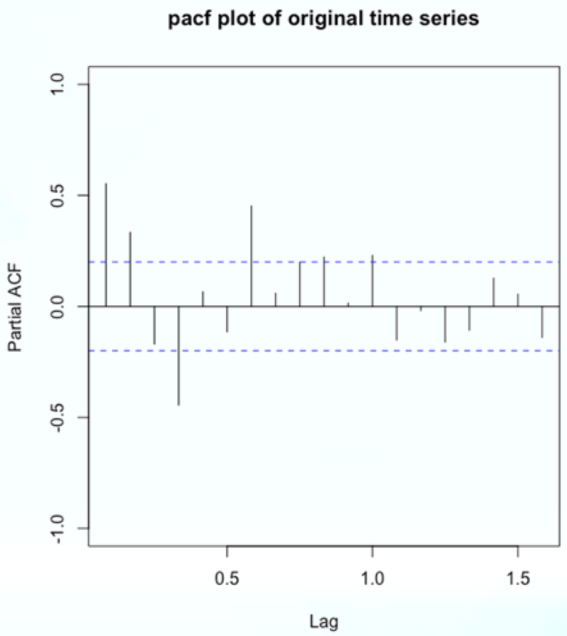


5. Stationarity of time series

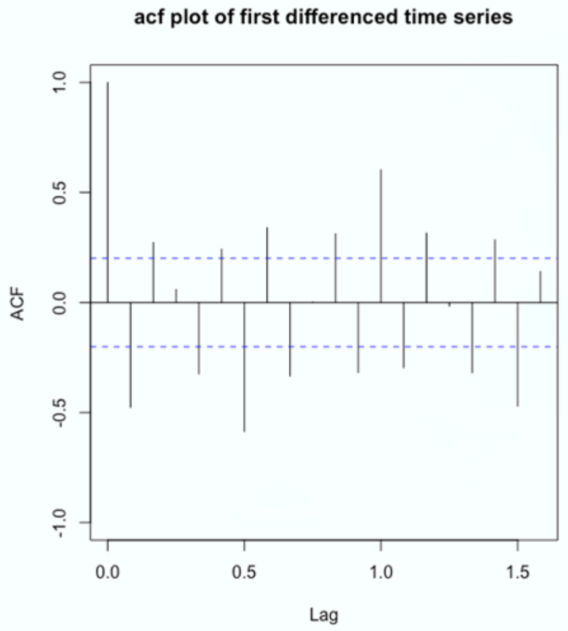
a. ACF, PACF plots and differencing

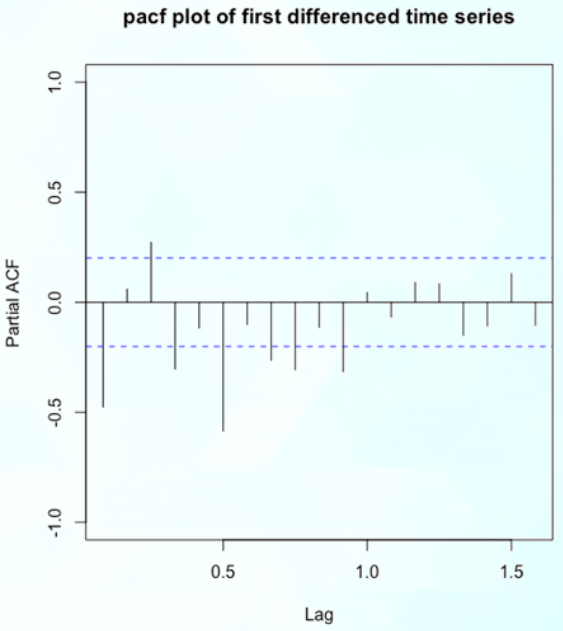
An Autocorrelation Function (ACF) plot reveals the correlation between the values of a time series and its lagged value at different lags. Alike ACF, the Partial Autocorrelation Function (PACF) plot also identifies the nature of the correlation between lags while neglecting the influence of the intervening lag. A stationary time series should generally exhibit ACF and PACF plots with correlations that decay rapidly to 0. Thus, ACF and PACF plots, apart from identifying the Autoregressive (AR) and Moving Average (MA) components respectively, also suggest whether the time series is stationary:


acf plot of original time series


pacf plot of original time series

ACF and PACF plots reveal high correlation at multiple lags are not ideal for implementing the ARIMA algorithm as they imply the presence of non-stationary time series. We thereby perform differencing once (the order of differencing performed corresponds to the value of *d* in ARIMA (*p,d,q*) model) to the time series that involve the subtraction of the current value of the series from the previous one. Then we may assess its ACF and PACF plots on the stationarity again:


acf plot of first differenced time series


pacf plot of first differenced time series

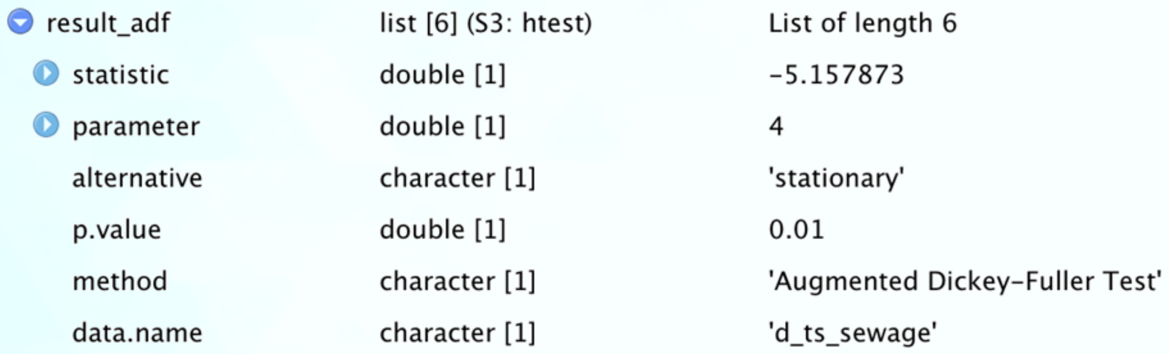
b. ADF and KPSS test

In view of more readily correlated ACF and PACF plots, it should be evident that further differencing the time series can hardly produce a stationary result. We may assess the stationarity of the series in alternative ways, namely by performing Augmented Dickey-Fuller (ADF) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.

An ADF test is a test of the unit root which indicates the behaviour of random walk. By examining the presence of the unit root in the first differenced series, we tell whether the first differenced time series is stationary. The null and alternative hypotheses are:

$$H_0\!: the\ time\ series\ contain\ unit\ root\ and\ is\ not\ stationary$$

$$H_1\!: the\ time\ series\ doesn't\ contain\ a\ unit\ root\ and\ is\ stationary$$



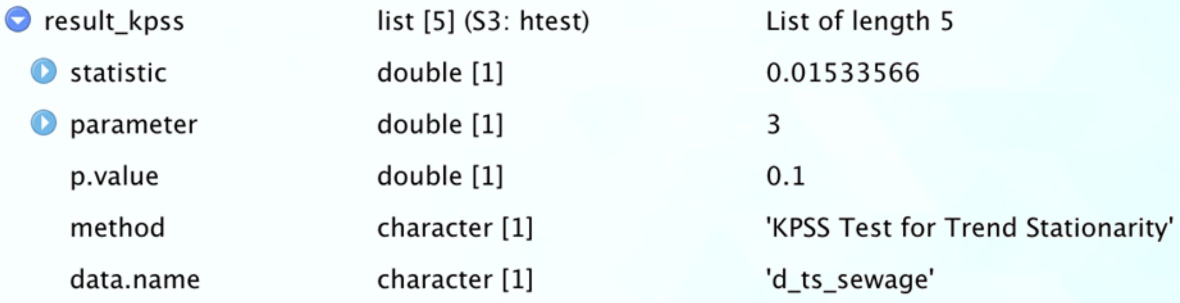
```
result_adf          list [6] (S3: htest)     List of length 6
  statistic         double [1]               -5.157873
  parameter         double [1]               4
  alternative       character [1]            'stationary'
  p.value           double [1]               0.01
  method            character [1]            'Augmented Dickey-Fuller Test'
  data.name         character [1]            'd_ts_sewage'
```

The test statistic of the ADF test (-5.157873) implies the rejection of the null hypothesis. Thus, the first differenced series is claimed to be stationary under the ADF test.

The KPSS test, on the other hand, is also a test of unit root and behaviour of random walk but with reversed null and alternative hypotheses specifying:

$$H_0\!: the\ time\ series\ is\ stationary$$

$$H_1\!: the\ time\ series\ is\ not\ stationary$$



```
result_kpss         list [5] (S3: htest)     List of length 5
  statistic         double [1]               0.01533566
  parameter         double [1]               3
  p.value           double [1]               0.1
  method            character [1]            'KPSS Test for Trend Stationarity'
  data.name         character [1]            'd_ts_sewage'
```

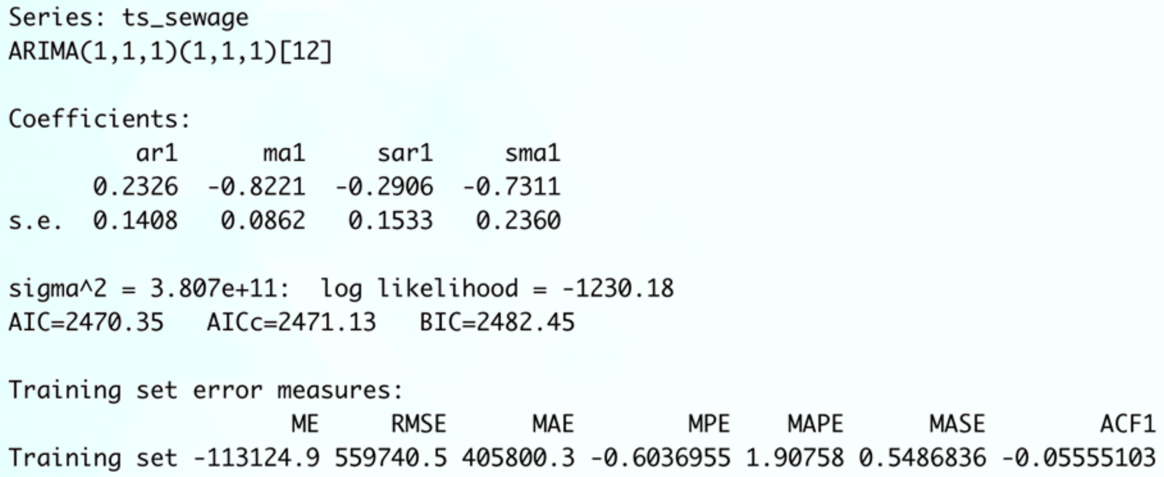
The test statistic of the KPSS test (0.01533566) falls into the non-critical region, so we cannot reject the null hypothesis. The result suggested by the ADF test coincides with that of the KPSS test, indicating the first differenced time series as well as the original series are stationary.

6. ARIMA model selection

Monthly sewage flow undeniably involves seasonality, and we should embrace the use of a seasonal ARIMA model: ARIMA (*p,d,q*) (*P,D,Q*) in which the latter part account for the AR, order of differencing, and MA component in seasonality.

a. Auto ARIMA

The Auto ARIMA function compares the Akaike Information Criterion (AIC) value of the models while continuing to add AR and MA components until the AIC is minimized or reaches the maximum order of the AR and MA components that the algorithm recognised. In the case of a seasonal ARIMA model, it also checks the seasonal pattern and adds seasonal AR and MA if necessary and returns the best model on search, corresponding to the model with the least AIC.

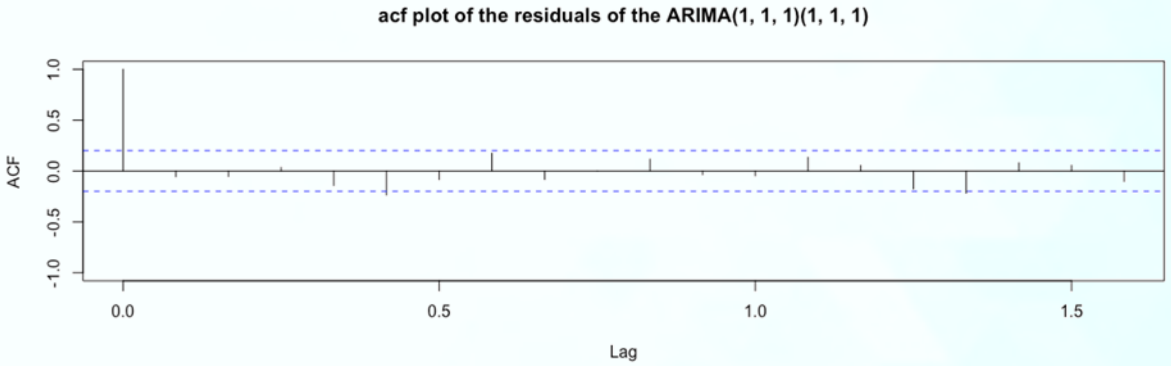


```
Series: ts_sewage
ARIMA(1,1,1)(1,1,1)[12]

Coefficients:
         ar1      ma1     sar1     sma1
      0.2326  -0.8221  -0.2906  -0.7311
s.e.  0.1408   0.0862   0.1533   0.2360

sigma^2 = 3.807e+11:  log likelihood = -1230.18
AIC=2470.35   AICc=2471.13   BIC=2482.45

Training set error measures:
                   ME     RMSE      MAE        MPE    MAPE      MASE        ACF1
Training set -113124.9 559740.5 405800.3 -0.6036955 1.90758 0.5486836 -0.05555103
```

The ARIMA (1,1,1) (1,1,1) model is suggested to be selected.

b. Residuals analysis

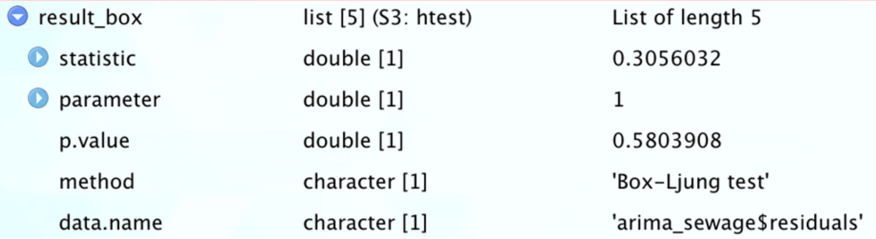
To measure the level of fit, residual analysis is performed and the ACF plot is presented below:



The presence of a high autocorrelation only in lag 0 implies that residuals at time *t* only correlate to the observation at that time. It shows very little influence on observation at other lags. In easier words, the residuals are independently distributed. Besides ACF analysis, a “portmanteau” test called the Ljung-Box Q-test that tests the goodness of fit of the model may also be adopted to analyse the fit of the ARIMA (1,1,1) (1,1,1) with null and alternative hypotheses specified as:

$$H_0\text{:}\, the\ model\ doesn't\ show\ a\ lack\ of\ fit$$

$$H_1\text{:}\, the\ model\ does\ show\ a\ lack\ of\ fit$$

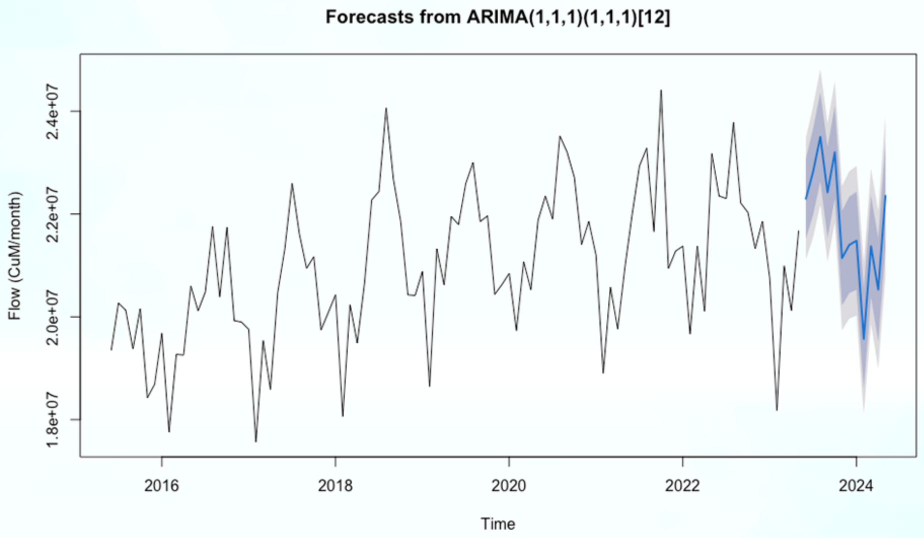


```
result_box        list [5] (S3: htest)    List of length 5
  statistic       double [1]              0.3056032
  parameter       double [1]              1
  p.value         double [1]              0.5803908
  method          character [1]           'Box-Ljung test'
  data.name       character [1]           'arima_sewage$residuals'
```

The underlying distribution for the Ljung-Box Q-test is a Chi-square distribution with the test statistic χ^2 =0.3056032, which falls into the non-critical region. This provides sufficient evidence that the ARIMA (1,1,1) (1,1,1) doesn't show a lack of fit and is, therefore, the best-fit model in predicting the future sewage flow.

7. Forecast

The time series for the prediction is constructed and the forecast value (CuM) regarding each month starting June 2023 and its confidence intervals are presented in the plot below:



8. Conclusion and future work

The one-year forecast starting June 2023 demonstrates a similar pattern as the past sewage flow. The flow (CuM/month) maximizes on summer days and minimizes on winter days. It coincides with the weather changes in Hong Kong. On the other hand, Hong Kong's overall sewage flow exhibits a gentle upward trend since May 2015. This also coincides with the recent growth in population in Hong Kong. As a result, these conclusions and the figures found may provide little help and insights into the sewage situation in Hong Kong, especially to the city-wide tailor-made sewage surveillance programme for the sake of hygiene as well as COVID-19 wave monitoring and combating (Ng et al., 2023).

However, the sewage forecast in monthly intervals is just fundamental preliminary research for sewage sampling as it only describes the flow by month. For a continuous and precise sampling of daily or hourly intervals, one must look into the situation and learn the pattern of such sewage flow to ensure sampling accuracy. Meanwhile, the weather-sewage peak coincidence and the population-sewage trend coincidence are just brief conclusions. There may exist other dominating factors that affect sewage production. In this case, we suggested that dimension reduction techniques (such as the Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), etc.) should be utilized simultaneously with sewage research to investigate the principal factors regarding both long-term and short-term sewage production trends.

9. References

[1] Abunama, T. and Othman, F. (2017) ‘Time series analysis and forecasting of wastewater inflow into Bandar Tun Razak Sewage Treatment Plant in Selangor, Malaysia’, IOP Conference Series: Materials Science and Engineering, 210, p. 012028. doi:10.1088/1757-899x/210/1/012028.

[2] Bar-Or, I. et al. (2022) ‘Regressing sars-COV-2 sewage measurements onto covid-19 burden in the population: A proof-of-concept for quantitative environmental surveillance’, Frontiers in Public Health, 9. doi:10.3389/fpubh.2021.561710.

[3] Gibson, K.E. et al. (2012) ‘Measuring and mitigating inhibition during quantitative real time PCR analysis of viral nucleic acid extracts from large-volume environmental water samples’, Water Research, 46(13), pp. 4281–4291. doi:10.1016/j.watres.2012.04.030.

[4] Hendriksen, R.S. et al. (2019) ‘Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage’, Nature Communications, 10(1). doi:10.1038/s41467-019-08853-3.

[5] Laboratory data and Sewage Flow Data (no date) DATA.GOV.HK. Available at: https://data.gov.hk/en-data/dataset/hk-dsd-dsd_psi_1-stp-sewage-data (Accessed: 30 July 2023).

[6] Larsson, D.G., Flach, C.-F. and Laxminarayan, R. (2022) ‘Sewage surveillance of antibiotic resistance holds both opportunities and challenges’, Nature Reviews Microbiology, 21(4), pp. 213–214. doi:10.1038/s41579-022-00835-5.

[7] Man, Y., Hu, Y. and Ren, J. (2019) ‘Forecasting cod load in municipal sewage based on Arma and var algorithms’, Resources, Conservation and Recycling, 144, pp. 56–64. doi:10.1016/j.resconrec.2019.01.030.

[8] Ng, W. et al. (2023) ‘The city-wide full-scale interactive application of sewage surveillance programme for assisting real-time covid-19 pandemic control – A case study in Hong Kong’, Science of The Total Environment, 875, p. 162661. doi:10.1016/j.scitotenv.2023.162661.

[9] Rogawski McQuade, E.T. et al. (2023) ‘Real-time sewage surveillance for SARS-COV-2 in Dhaka, Bangladesh versus clinical covid-19 surveillance: A Longitudinal Environmental Surveillance Study (December, 2019–December, 2021)’, The Lancet Microbe, 4(6). doi:10.1016/s2666-5247(23)00010-1.

[10] Wąsik, E. and Chmielowski, K. (2016) ‘The use of Holt–Winters Method for forecasting the amount of sewage inflowing into the wastewater treatment plant in Nowy Sącz’, Ochrona Srodowiska i Zasobów Naturalnych, 27(2), pp. 7–12. doi:10.1515/oszn-2016-0009.

Interpretable Machine Learning Techniques on Text Classification Model

Carly LAI Yuk-ling, Jason CHAN Chung-lam, Alex LI Sze-chun,
Proton NG Ka-hei and Hinz SHUM Tsz-hin
Census and Statistics Department

Introduction

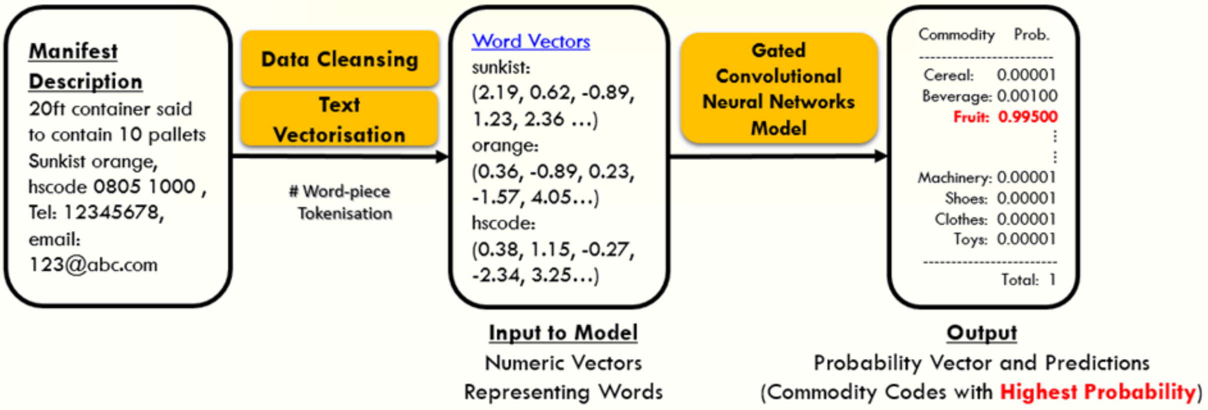
Both traditional statistical modeling and machine learning modeling involve building a mathematical model to link up input and output variables, where their linkages are represented by mathematical equations as well as model parameters. Unlike traditional statistical model, complex machine learning models are generally featured by a large number of equations and parameters and better data fitting but at the same time rendering the model less interpretable by human (also known as 'black box'), posing challenges to designers to carry out model enhancement or justify the reliability of the model in real-life application.

2. Here is where interpretable machine learning techniques kick in. Interpretable machine learning techniques refer to a set of methods that demystify a targeted black-box machine learning model to improve model interpretability and transparency. With the black-box unfolded, one could ensure the reliability of the model through understanding the decision processes of the model.

3. This paper compares two state-of-the-art model-agnostic methods, namely Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), and demonstrates how SHAP can be used in understanding text classification models. Text Analytics Module for Commodity Classification (CC Module), an in-house text classification model maintained by the Census and Statistics Department (C&SD), is taken for demonstration.

4. CC Module automatically classifies consignment into corresponding commodity groups based on commodity descriptions reported on cargo manifests and is being used for auto-coding commodity descriptions in daily work. A summary of the model flow is given in **Figure 1**.

Figure 1 – Model flow of CC Module



5. The first and an important step before building models is data cleansing, i.e. to re-format the raw textual descriptions declared on cargo manifests and remove any inherent noises. Then, word-piece tokenisation and text vectorisation will be performed to transform word tokens into numeric vectors (i.e. Word Vectors) using the pre-trained Word2Vec model. The Word Vectors representing the commodity descriptions will then be fed into the Gated Convolutional Neural Network model and make classification. For each consignment, the model will output prediction probabilities for all commodity groups, with the prediction probabilities sum to 1. The commodity group corresponding to the highest output probability will be the ultimate prediction.

Interpretable Machine Learning Techniques

6. Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) are two popular model-agnostic methods to explain the outputs of any models with some given inputs. Model-agnostic methods can be applied to any machine learning models (even deep learning models) without prior knowledge of the model internals such as model weights or structural information. These methods only require probing machine learning models as black-boxes by passing some inputs into them.

7. Specifically, model-agnostic methods involve training another interpretable machine training model, which is called explanation model, to approximate the targeted black-box model. Given the complexity of modern machine learning models, explanation models usually have only good local approximation. This property is called local fidelity.

Local Interpretable Model-Agnostic Explanations (LIME)

8. Using LIME to interpret a text classification model involves training an explanation model, which is a weighted linear regression model with word tokens as independent variables, with the following steps:

(a) Generate new samples of texts for a given input by randomly removing token(s) or word(s);

(b) Query the black-box to get the prediction of the new samples;

(c) Weight the new samples according to their proximity to the original text;

(d) Train a weighted linear regression on the above new dataset, with each token to be a binary input feature; and

(e) Extract the coefficients of the weighted linear regression to explain the influence of respective token.

9. The use of LIME to interpret a text classification model can be illustrated by the following hypothetical example[1], in which the binary classification model is to detect whether a text message is spam or not. Supposed that there is a message 'For Christmas Song visit my channel! ;)' and the binary classification model predicted this message as spam. Using the LIME method, it will generate new samples by randomly removing words in the message as shown in **Table 1**.

[1] Modified based on https://christophm.github.io/interpretable-ml-book/lime.html

Table 1 – Random samples generated by the LIME method

Sample	Token / Word							Predicted probability	Weight
	For	Christmas	Song	visit	my	channel!	;)		
1	1	0	1	1	0	0	0	0.17	0.43
2	0	1	1	1	1	0	1	0.17	0.71
3	1	0	0	1	1	1	1	0.99	0.71
4	1	0	1	1	1	1	1	0.99	0.86
5	0	1	1	1	0	0	1	0.17	0.57

10. The method generates five randomly truncated pieces from the original message, with '1' indicating the presence of word-piece and '0' indicating the absence (e.g. Sample 1 reads 'For Song visit'). The column 'Predicted probability' shows the output by the binary classification model for the respective samples, and the column 'Weight' reflects how close the sample is with regard to the original message (e.g. weight of sample 1 equals to 3 divided by 7). From these samples, we can observe that whenever the word -piece 'channel!' is present, the prediction probability is very high (0.99), while the prediction probability plummets under its absence (0.17). It suggests that this word-piece is indicative in making classification.

SHapley Additive exPlanations (SHAP)

11. Similar to LIME, SHAP is also a model-agnostic method and creates an explanation model with a linear combination of tokens, but the coefficients are derived from methods based on well-established optimal Shapley values from game theory. Shapley value (slightly different from SHAP value) is a solution to the problem – 'how to assign the contribution or importance of each player under cooperative games, where payoffs are only assigned to the outcome of the game as a whole, not to individual player?'. In the context of machine learning models, Shapley value decomposes the model outputs into a sum of importance values for each input. Lundberg and Lee (2017) showed that Shapley value is the only explanation model satisfying three desirable theoretical properties, namely missingness, consistency and especially local accuracy, where it guarantees the outputs of the explanation model match the outputs of the original black-box model locally (the model by LIME method as illustrated above does not guarantee this).

12. A Shapley value of a token is the weighted average of model outputs without and with the token over all possible combinations of other tokens. In the hypothetical example shown in **Table 1**, for computing the Shapley value of the token 'For', it would require calculating the weighted average of changes in model outputs over the 64 pairs of text without and with the token 'For'. Thus the contribution of 'For' to the model output is assigned in consideration of the presence of all combinations of other tokens.

[2] Lloyd Shapley, who introduced Shapley value in 1951, won the Nobel Memorial Prize in Economic Sciences for it in 2012.

[3] (1) 'For' vs empty text; (2) 'For Christmas' vs 'Christmas'; (3) 'For Song' vs 'Song'; … ,
(64) 'For Christmas Song visit my channel! ;)' vs 'Christmas Song visit my channel! ;)'

[4] Open-source program code at https://github.com/slundberg/shap/blob/master/shap/explainers/_partition.py

13. However, the exact computation of Shapley value is computationally infeasible owing to the large number of possible subsets involved. For a text with *k* tokens, it requires generation model prediction of 2^k different combinations of tokens. The SHAP method solves the computational issues by introducing several feasible ways in approximating Shapley values. For text classification models, the default approximation method in the Python package 'SHAP' computes SHAP values recursively through a hierarchy of features with quadratic exact runtime in terms of number of tokens input. While the computation is relatively more efficient using SHAP, the execution time for computing the Shapley values of all tokens for each manifest being passed through CC Module is around one minute.

14. In gist, the SHAP method gives a SHAP value for each input feature (i.e. each token) when it is applied to a text classification model. The sum of these SHAP values will be equal to the prediction probability of the original black-box model. In other words, the SHAP value of each input feature can be viewed as the contribution of each token to the overall prediction probability.

Comparison of LIME and SHAP

15. As represented by the coefficients in LIME and the SHAP values in SHAP, both LIME and SHAP methods shed light on the decision process of a black-box classification model by showing how a targeted black-box model weights each input token. Yet, LIME results are relatively hard to interpret as the sum of coefficients may not equal to the output prediction probability of the original black-box model. Furthermore, since LIME involves generating random samples, its outputs are stochastic, whereas the outputs of SHAP are stationary (i.e. the same explanation model is built each time). Theoretically speaking, SHAP is superior to LIME.

16. In terms of implementation of these two methods, there are Python packages available for both methods (called 'LIME' and 'SHAP' respectively) and hence technical barriers against implementation should be low. Although SHAP is more computationally intensive, it should be adopted when there are sufficient computer resources.

Table 2 – Summary table on LIME and SHAP

Method	*Model-specific / -agnostic*	*Local / global surrogate model*	*Outputs matched?*	*Stochastic / stationary*	*Runtime per case*
LIME	Model-agnostic	Local	No	Stochastic	Few seconds
SHAP	Model-agnostic	Local	Yes	Stationary	one minute

Application

17. In the ensuing paragraphs, SHAP results are shown to demonstrate the potential usage of Interpretable Machine Learning Techniques on Text Classification Model. The Python package 'SHAP' comes with two visualisation methods, namely force plots and waterfall plots[5]. They visually display the contribution of each token in terms of SHAP values to the model's prediction for each observation. Positive and negative contributions of the respective tokens are shown in red and blue respectively.

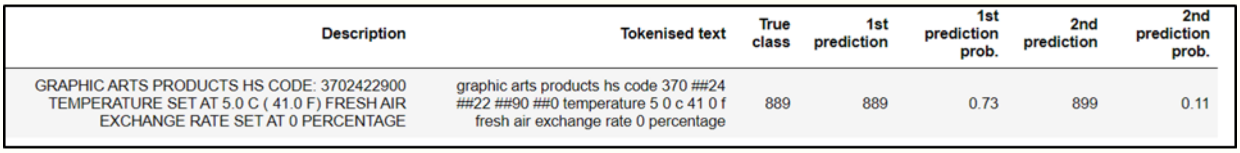
18. SHAP results suggest that CC Module makes proper use of the information in the commodity descriptions reported on cargo manifests for classification. The following insights can be drawn with the use of SHAP:

Insight 1 – Proper utilisation of Harmonized System commodity codes in classification

19. SHAP results suggest that commodity codes under Harmonized System (HS codes)[6], which are commonly included in commodity descriptions of cargo manifests, are properly picked up by CC Module when making classifications. Moreover, CC Module is able to utilise HS codes of different economies by recognising the fact that the first six digits of HS codes in each economy generally comply fully with each other.

20. In **Example 1 (Table 3)** below, 'HS CODE: 3702422900', which is believed to be an HS code of the mainland of China (underlined in red), is reported on the commodity description and is tokenised into 'hs', 'code', '370', '##24', '##22', '##90' and '##0'[7] after pre-processing (underlined in blue). SHAP results reveal that CC Module correctly classifies this consignment under commodity group 889 'Photographic Apparatus, Equipment and Supplies and Optical Goods; Watches and Clocks' (i.e. the output '1st prediction' is equal to 'true class') with a prediction probability of 73%. CC Module also suggests that the prediction probability for the second-highest prediction, commodity group 899 'Other Manufactured Articles', is 11%.

Table 3 – Output of CC Module (Example 1)



Description	Tokenised text	True class	1st prediction	1st prediction prob.	2nd prediction	2nd prediction prob.
GRAPHIC ARTS PRODUCTS HS CODE: 3702422900 TEMPERATURE SET AT 5.0 C (41.0 F) FRESH AIR EXCHANGE RATE SET AT 0 PERCENTAGE	graphic arts products hs code 370 ##24 ##22 ##90 ##0 temperature 5 0 c 41 0 f fresh air exchange rate 0 percentage	889	889	0.73	899	0.11

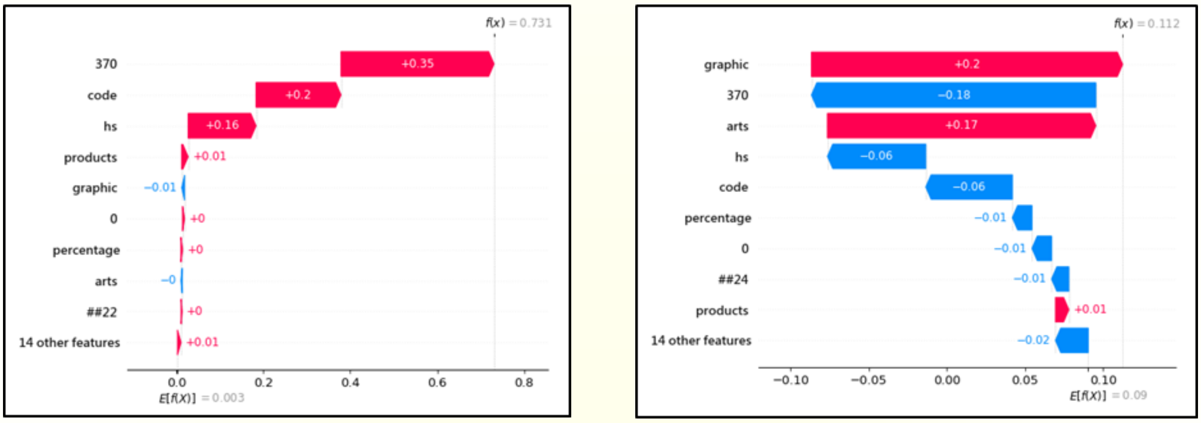
[5] The use of these plots will be demonstrated in later paragraphs.

[6] Harmonized System (HS) is designed by the World Customs Organization (WCO) and is a hierarchical system of commodity classification and coding. In Hong Kong, the Hong Kong Harmonized System (HKHS) is an 8-digit classification system with the first 6 digits complying fully with the HS designed by the WCO; and the additional 7th and 8th digits meeting requirements for more detailed commodity classification in Hong Kong.

[7] The "##" sign before sub-words indicates that they are not the first sub-word of the word. For instance, '##24' may cover '37024' or 'hs24'.

21. The SHAP values of each token in respect of the first and second predictions are shown in the waterfall plots in **Figure 2**. Waterfall plot is applied to provide more detailed explanations for an individual prediction. The bottom of a waterfall plot starts as the expected value of the model output *E[f(x)]* which refers to the prior expectation under the background data distribution. Each row of waterfall plot shows how the positive (red) or negative (blue) contribution of each feature (token) moves the model output from prior expectation to the final prediction probability given the evidence of all the features. The final prediction probability is denoted as f(x) in the top right-hand corner in the charts.

Figure 2 – Waterfall plots of SHAP values in respect of Example 1



Commodity group 889 (correct prediction)

Commodity group 899 (incorrect prediction)

22. It is observed that the tokens ‘370’, ‘code’ and ‘hs’ contribute the most in making the correct classification of commodity group 889 (together contributed 71% points). Besides, the text ‘370’ also helps lower down the probability in making incorrect prediction of commodity group 899, by 18% points. For the subsequent digits of the HS code, their contributions are negligible in both the correct and incorrect predictions, suggesting that HS codes at refined levels may not be essential in making correct predictions.

Insight 2 – Room for improvement in standardisation of common Chinese phrases

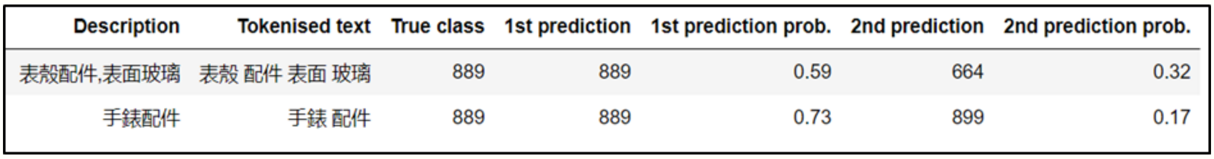
23. Another finding is that CC Module’s performance can be enhanced by considering common Chinese phrases when performing tokenisation. This will ensure commodity descriptions reported on cargo manifests are correctly split for model fitting and classification.

Polysemic phrases

24. Polysemic phrases are common in Chinese and should be more properly considered before tokenisation. In Example 2 (Table 4) below, both consignments contain the parts of watches and are indisputably of commodity group 889. Although CC Module correctly classifies both consignments, the confidence (as represented by the prediction probabilities) of such predictions is not as high as we expected, and the prediction probability is notably lower for the first consignment even both descriptions are similar. Our study shows that this is mainly attributable to the interchangeable use of ‘表’ and ‘錶’ in the Chinese community. Moreover, ‘表面’ is polysemic in Simplified Chinese, where ‘表面’ may refer to

'surface' or 'dial' of a watch. All these have brought additional noise to CC Module in making classification.

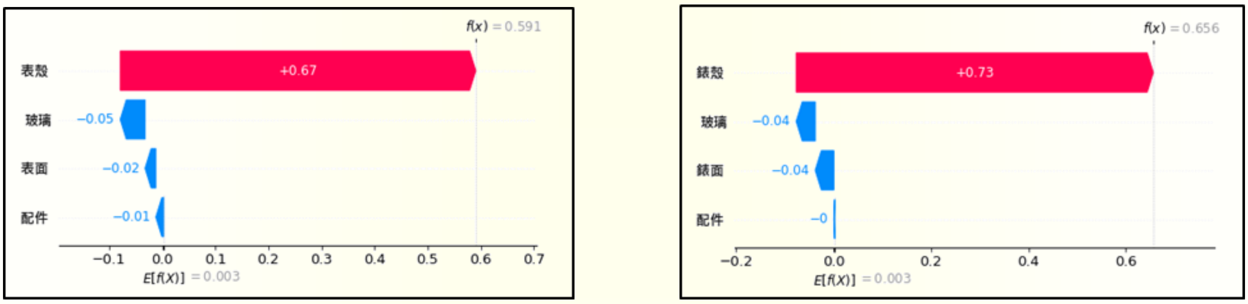
Table 4 – Output of CC Module (Example 2)



Description	Tokenised text	True class	1st prediction	1st prediction prob.	2nd prediction	2nd prediction prob.
表殼配件,表面玻璃	表殼 配件 表面 玻璃	889	889	0.59	664	0.32
手錶配件	手錶 配件	889	889	0.73	899	0.17

25. **Figure 3** shows the SHAP values of the original texts and the modified texts (i.e. '錶殼', '配件', '錶面' and '玻璃'). It demonstrates the gain in terms of prediction probability without re-training CC Module after modifying the input tokens (65.6% - 59.1% = 6.5% points), particularly due to the increased contribution by '錶殼' upon the modification. In addition, as evidenced by its low contribution, CC Module somehow still ignores '錶面' when making classification based on the modified text. It further adds on that CC Module may be confused about '表' and '錶' when making classification related to commodity group 889.

Figure 3 – Waterfall plots of SHAP values in respect of Example 2



Original texts

Modified texts (after converting '表' to '錶')

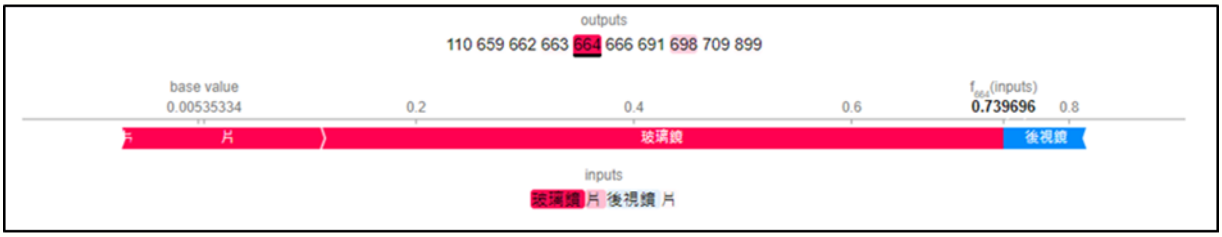
Tokenisation of Chinese word

26. Given an input text string, there are different ways to conduct text splitting, thus generating different sets of input tokens. As there is no blank space in Chinese language for word delimitation, a more detailed rules in performing Chinese word tokensiation will be desirable.

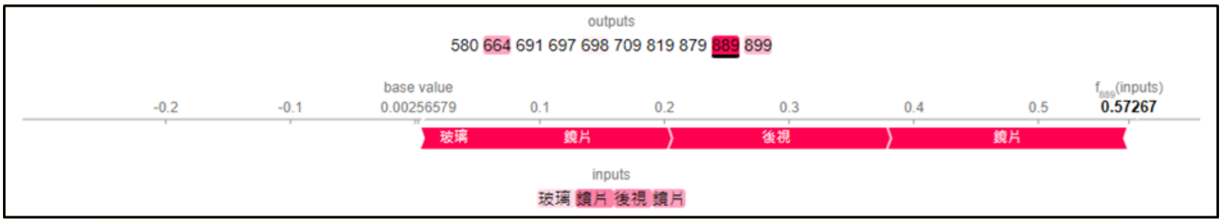
27. Consider an item with the Chinese description '玻璃鏡片 後視鏡片' as shown in **Example 3 (Figure 4)**. CC Module tokenises the original text into four sub-words (namely '玻璃鏡', '片', '後視鏡', '片') and it incorrectly classifies the item under commodity group 664 'Glass and Glassware' with a prediction probability of 74%. Yet, the actual commodity group should be 889 'Photographic Apparatus, Equipment and Supplies and Optical Goods; Watches and Clocks'.

Figure 4 – Force plots of SHAP values in respect of Example 3

Contribution of each token to commodity group 664 (incorrect prediction)



Contribution of each token to commodity group 889 (correct prediction)



28. We can see that the token ‘玻璃鏡’ plays a very significant role in predicting the item as commodity group 664. If another way of tokenisation is applied by dividing the commodity decryption into ‘玻璃’, ‘鏡片’, ‘後視’ and ‘鏡片’, CC Module can classify this consignment correctly with a prediction probability of 57%.

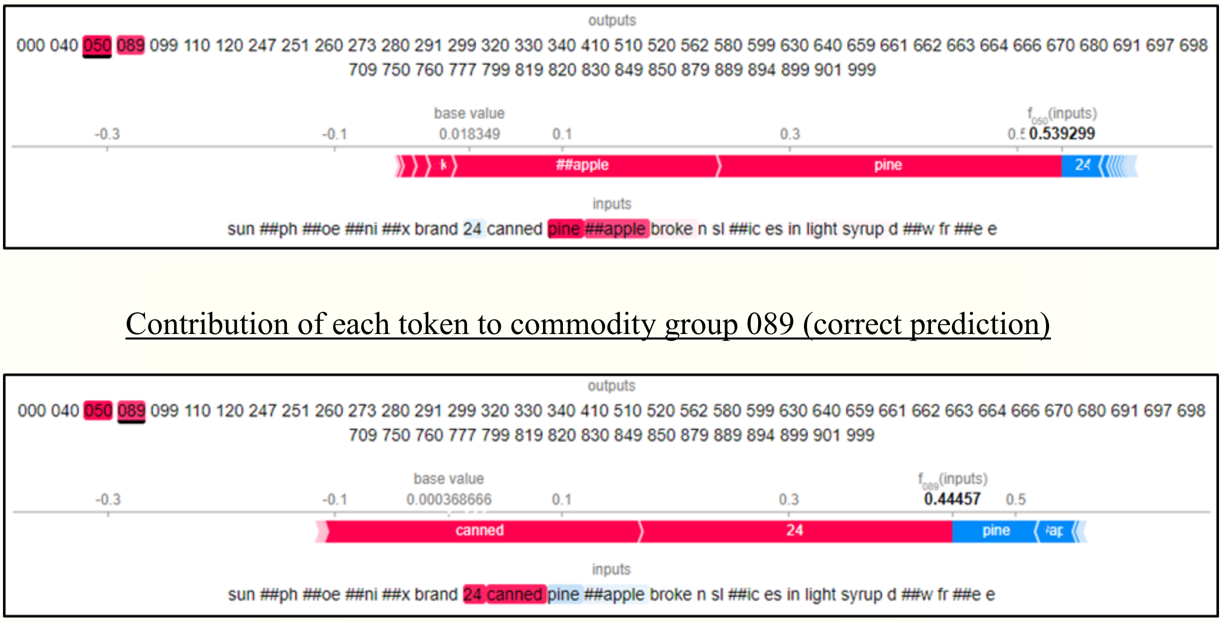
29. Therefore, compiling a comprehensive list with more common phrases of Chinese texts for standardisation should further boost CC’s performance. It can reduce the number of entries in the dictionary of predefined terms as well as noise in model training and making classifications.

Insight 3 – Identification of keywords crucial for classification

30. SHAP results also suggest that particular keywords in the commodity descriptions are crucial or even decisive for classification. In **Example 4 (Figures 5 and 6)** below, CC Module classifies this canned pineapple slices consignment incorrectly under commodity group 050 ‘Vegetables and Fruit’ but the actual commodity group should be 089 ‘Canned Food’. It is noted that the token ‘canned’ plays a decisive role in making classification as it determines whether a food item should be classified under canned foods or other commodity groups related to fresh food. In fact, CC Module also recognises that the canned pineapple slices may belong to commodity group 089 by giving it the second highest prediction probability of 45%, but the value is still lower than that of commodity group 050.

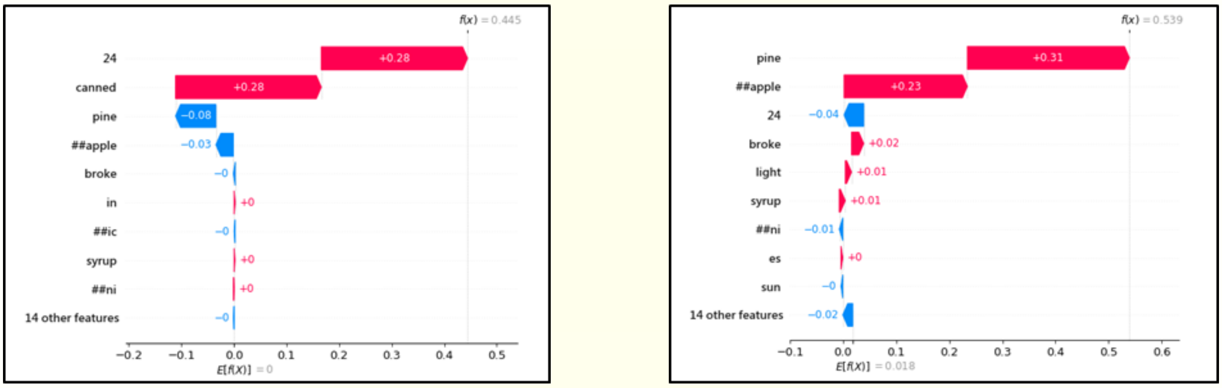
Figure 5 – Force plots of SHAP values in respect of Example 4

Contribution of each token to commodity group 050 (incorrect prediction)



Contribution of each token to commodity group 089 (correct prediction)

Figure 6 – Waterfall plots of SHAP values in respect of Example 4



Commodity group 089 (correct prediction)

Commodity group 050 (incorrect prediction)

31. **Figure 5** shows the SHAP values of the canned pineapple slices consignment and **Figure 6** shows the waterfall plots for the top two tokens. The figures reveal that CC Module classifies the 'canned pineapple slices' under incorrect commodity group 050 with a prediction probability of 54%, of which tokens 'pine' and 'apple' contributed 31% points and 23% points respectively. Other tokens in the commodity description have negligible impact for commodity group 050.

32. On the other hand, CC Module classifies the 'canned pineapple slices' under commodity group 089 with a prediction probability of 45%, of which both sub-words '24' and 'Canned' each contributed 28% points. The combined prediction probability of 56% is even higher than that of the incorrect commodity group 050. However, CC module recognises that sub-words 'pine' and 'apple' in the commodity description are more related to fruits instead of canned foods. As such, negative SHAP values are assigned to these sub-words and reduce the prediction probability for the correct commodity group 089.

33. With SHAP, tokens which are decisive in commodity group classification can be identified systematically. With this information, a set of rules for rule-based classification can be developed to those commodity groups during the phase of pre-processing to further improve the accuracy of CC Module.

Concluding remarks

34. Interpretable machine learning techniques, especially SHAP, are found to be very conducive to the enhancement of text analytics models as it enables us to understand the decision process of a given model. Apart from refining the model itself through modifying its pre-processing procedures and adjusting the weighting of inputs through oversampling in model training, some rule-based algorithm may also be mined according to the output of SHAP to further improve the efficiency and effectiveness of a model alongside the use of text analytics algorithms. As machine learning models become more complex, the application of interpretable machine learning techniques will become increasingly important.

References

Miller, Tim. (2017). *"Explanation in artificial intelligence: Insights from the social sciences."*, arXiv:1706.07269.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you? Explaining the predictions of any classifier."*, *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, arXiv:1602.04938

Lundberg, S. M., & Lee, S. I. (2017). *"A unified approach to interpreting model predictions"*, *Advances in Neural Information Processing Systems*, *30*.

Latest development of accreditation and mutual recognition of Graduate Statistician professional qualification between the Hong Kong Statistical Society and the Royal Statistical Society

Introduction

The Hong Kong Statistical Society (HKSS) has been offering the professional membership of Graduate Statistician (GradStat) and Certified Statistician. The Accreditation Board (AB) of HKSS has also accredited a number of statistical programmes offered by universities in Hong Kong up to the standard of GradStat. In regard to the professional GradStat qualification, there is mutual recognition between the HKSS and the Royal Statistical Society (RSS) in U.K. Indeed, the HKSS and the RSS have long history of close collaboration. In August 2001, the RSS and the HKSS signed an Agreement under which the examination held many years by the RSS in Hong Kong was replaced by the HKSS examination as from the May 2002 round of the examination. Under the agreement, the two societies have also established mutual recognition of the professional qualifications granted by each other, and the accreditation granted by the HKSS is fully endorsed by the RSS. The recognition by the RSS is a strong proof of the professional standard of both the HKSS and the statistical programmes of universities in Hong Kong.

In 2015, the RSS informed the HKSS of the decision that the RSS would discontinue its professional examinations from September 2017, after a strategy review to take forward the work of the RSS. Instead of conducting professional examinations, the RSS had decided to move to an accreditation model after 2017. Following the decision of the RSS, the HKSS ceased to offer its professional examination after the 2017 round of the examination. Opportunity was also taken to explore with the RSS for setting up a similar accreditation model in Hong Kong. As the deliberation between the two societies took time, mutual agreement was reached between the two societies on a transitional arrangement for candidates with partial completion of the past Graduate Diploma examination and undergraduates of accredited courses to continue to acquire the GradStat qualification by taking university programmes accredited by the HKSS or the RSS. The mutual agreement was extended due to the outbreak of COVID-19 which deferred the deliberation process. In between, exchange of views was made mainly through online meetings and emails.

As the COVID-19 pandemic started to subdue, the discussion of joint accreditation and mutual recognition of the GratStat qualification has resumed in the first half of 2023. The RSS has come up with updated guidelines of its accreditation model and introduced the new professional qualifications of Data Analyst, Data Science Professional and Advanced Data Science Professional. To enable a more thorough discussion with the RSS on the new accreditation model and the new professional qualifications, the AB in collaboration with the HKSS Council and the University of Hong Kong, the Chinese University of Hong Kong and the City University of Hong Kong, invited RSS representatives to meet in Hong Kong.

Visit to HKSS by RSS representatives

Under HKSS's invitation, Professor Rachel Hilliam, Vice President of Professional Affairs and Chair of the Alliance for Data Science Professionals, Ms Nicola Emmerson, Director of Membership and Professional Affairs and Mr Ricky McGowan, Head of Standards and Corporate Relations of the RSS visited Hong Kong during 6 - 11 November 2023. The aim of the visit of the RSS team was threefold: (1) to accredit statistical programmes of the three universities in Hong Kong in accordance with the latest criteria and standards specified in the accreditation model of the RSS; (2) to share with the HKSS the methods and procedures adopted in applying the latest accreditation standards and (3) to introduce the new Data Science Professional and Data Analyst membership to the HKSS.

During the visit, RSS representatives visited the three universities and processed the accreditation applications of 14 statistical programmes. Representatives of the AB of the HKSS participated in the entire accreditation exercise with a view of understudying the use of the RSS standards and guidelines in accrediting statistical programmes based on the latest accreditation models of the RSS. During the process, it was noted that some additional information of the statistical programmes was to be provided to the RSS for completing the accreditation. As at the time of preparing this article, six programmes have been accredited on an unconditional basis and two on a conditional basis, while remaining programmes are being processed after additional information has been provided to the RSS.

New professional qualifications

During the RSS's visit, Prof Rachel Hilliam delivered a talk on "The changing landscape of UK statistics and the road to professionalisation of Data Science" to members of the HKSS. The talk explored the rise of Data Science and how this shaped the role of a statistician in the UK. In response to the multidisciplinary nature of data science and the need to provide new industry-wide standards covering a wide range of issues from good data management to effective data problem solving skills and ethnical use of data, the RSS in collaboration with six prestigious societies of the UK (viz. the British Computer Society, the Chartered Institute for IT, the Operational Research Society, the Institute of Mathematics & its Applications, the National Physical Laboratory and the Alan Turing Institute) formed the Alliance of Data Science Professionals in 2020. The initial chairperson of the Alliance is Prof Rachel Hilliam. The Alliance aims to promote the professional standard of data science through introduction of the new Data Science Professional pathway. The work quickly gained support from the UK Government, where it was referenced in the National Data Strategy of the UK in November 2020 and more recently in the UK Parliament POST report "Data Science Skills in the UK Workforce" in June 2023.

The new Data Science Professional pathway offered by the RSS and the Alliance has three levels (see also the Chart below):

Data Analyst

Introduced by the RSS in May 2021, data Analyst is a registered grade of professional membership of the RSS. It provides formal recognition of a member's statistical training. One may be eligible for the RSS registered Data Analyst award if he/she meets all of the criteria listed below:

- has successfully completed an RSS Quality Marked course or a course that meets this standard;
- has at least one year's work experience within a statistical role; and
- is able to supply evidence of 30 hours of continued professional development in the last 12 months.

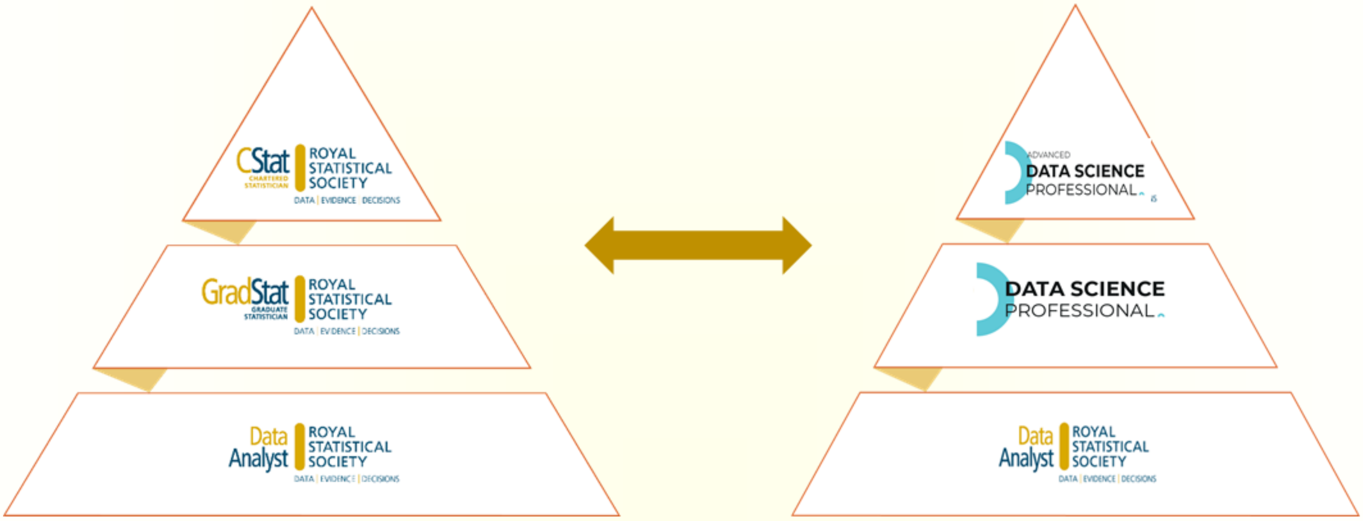
Data Science Professional

Data Science Professional is a new certificate for recent graduates with at least two years of full-time work experience within data science, providing formal recognition of qualifications, professional training and experience. This certificate is the work of the RSS with the Alliance for Data Science Professionals of which the RSS is a founding member, creating a competency framework and pathway for all of those working within data science.

Advanced Data Science Professional

Advanced Data Science Professional is a new certificate for experienced data scientists with at least five years of full-time work experience within data science, providing formal recognition of qualifications, professional training and experience. Same as the Data Science Professional, the certificate is the work of the RSS with the Alliance for Data Science Professionals.

Chart: The RSS Professional Pathways


CStat CHARTERED STATISTICIAN ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
GradStat GRADUATE STATISTICIAN ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Data Analyst ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
ADVANCED DATA SCIENCE PROFESSIONAL
DATA SCIENCE PROFESSIONAL
Data Analyst ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS

Source: Seminar on the changing landscape of UK statistics and the road to professionalisation of Data Science

Transitional agreement

On the basis of the week-long visit, both the HKSS and the RSS agreed to continue to work towards the goal of reaching an agreement in which the RSS will license the HKSS to use the RSS accreditation standards and procedures to accredit programmes of universities in Hong Kong. The benefits of the agreement will include mutual recognition of accredited qualifications and the RSS GradStat membership and the proposed Hong Kong equivalent professional membership. A new agreement is planned to be drawn up in 2024. In the meantime, a transitional agreement was signed by both societies on 10 November 2023.

Way forward

The AB of the HKSS will continue to work with the RSS with a view to introducing the accreditation model in Hong Kong and renewing the mutual recognition of the GradStat qualification granted by the two societies in the near future. We would also like to explore the opportunity to invite more academic staff of universities in Hong Kong to join the AB as board members or assessors to strengthen the capability of AB in processing future accreditation applications. We will also work towards the target of granting Hong Kong equivalent membership of Data Analyst, Data Science Professional and Advanced Data Science Professional as the next steps.

Prof. Alan WAN, President of the Hong Kong Statistical Society,
Mr HW FUNG, chairperson of the HKSS Accreditation Board,
and members of the HKSS Accreditation Board dined with the representatives from the RSS.

The talk delivered by Prof Rachel HILLIAM, the RSS Vice-president for Professional Affairs.

A transitional agreement was signed by both the HKSS and the RSS on 10 November 2023.

2022/23 Statistical Project Competition for Secondary School Students

Organising Committee* of the 2022/23 Statistical Project Competition

The 2022/23 Statistical Project Competition (SPC) for Secondary School Students, the 37th round of the Competition since 1986/87, was successfully completed. The SPC was jointly organised by the Hong Kong Statistical Society (HKSS) and the Education Bureau. The objective of the 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 with their own choices of themes, 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 "45th Anniversary of Hong Kong Statistical Society Prize for the Best Thematic Project" and the "Department of Management Sciences, the City University of Hong Kong Prize for the Best Graphical Presentation of Statistics".

To help interested participants prepare for the Competition, a Briefing Seminar for 2022/23 SPC was held on 22 October 2022. Representatives from the Census and Statistics Department and Hang Seng Indexes Company Limited, sponsor of 2022/23 SPC, had introduced the use of official statistics and shared with participants the topic on "Hang Seng Index and Statistics" respectively. The winners of the last round of Competition were also invited to share their experiences. In addition, an online exhibition of past winning projects was also carried out from 1 November 2022 to 30 November 2022.

Encouraging number of entries

In 2022/23 SPC, 151 entries and 86 entries were submitted for the Junior Section and the Senior Section respectively from 943 students of 64 secondary schools. The number of entries for both Junior and Senior Sections were moderately higher than the previous round. Demonstrating the diversity of topics, the entries covered various socio-economic issues of Hong Kong.

Adjudication panel led by Professor CHEUNG Ka-chun

An adjudication panel, led by the Chief Adjudicator, Professor CHEUNG Ka-chun of The University of Hong Kong, and comprised some 53 academics from local tertiary institutions as well as statisticians and research managers 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. The Organising Committee would like to express our special thanks to Professor CHEUNG Ka-chun, and Professor Keith CHAN Kin-wai of The Chinese University of Hong Kong for serving as the Chairpersons of the interview panel for Junior and Senior Sections respectively.

Prize Presentation Ceremony

The Prize Presentation Ceremony for the 2022/23 SPC took place on 3 June 2023 at the Auditorium of the Hong Kong Federation of Youth Groups Building. Professor Alan WAN Tze-kin, President of HKSS, Mr Leo YU Chun-keung, Commissioner for Census and Statistics, Ms Teresa CHAN Mo-ngan, Deputy Secretary for Education, Ms Candy LAM, Chief Strategy Officer of Hang Seng Indexes Company Limited, and Professor CHEUNG Ka-chun, Chief Adjudicator, were invited to the Ceremony to present prizes and trophies to the winning teams.

Group photo taken in the Prize Presentation Ceremony

Regarding the results of the Competition, students of La Salle College, who used official statistics to study the feasibility to achieve conductive environment for popularisation of electric vehicles in Hong Kong, won the First Prize for the Junior Section. Students of St. Paul's Co-educational College won the Second Prize, while students of Sheng Kung Hui Tang Shiu Kin Secondary School won the Third Prize. The Prize for the Best Thematic Project and the Prize for the Best Graphical Presentation of Statistics were both won by La Salle College.

2022/23年度
中學生統計習作比賽頒獎典禮
Prize Presentation Ceremony of
2022/23 Statistical Project Competition
for Secondary School Students
教育局
Education Bureau
恒生指數

Mr Leo YU presented the First Prize for the Junior Section to students of La Salle College

As for the Senior Section, the statistical report from students of Victoria Shanghai Academy was appraised as the best among all the projects. They applied official statistics to analyse the appropriateness of selected predictors of Hong Kong's housing market. Students of Delia Memorial School (Hip Wo) won the Second Prize, while students of Diocesan Girls' School won the Third Prize. The Prize for the Best Thematic Project and the Prize for the Best Graphical Presentation of Statistics were both won by Delia Memorial School (Hip Wo).

2022/23年度
中學生統計習作比賽頒獎典禮
Prize Presentation Ceremony of
2022/23 Statistical Project Competition
for Secondary School Students
教育局
Education Bureau
恒生指數

Ms Candy LAM presented the First Prize for the Senior Section to students of Victoria Shanghai Academy

2022/23年度
中學生統計習作比賽頒獎典禮
Prize Presentation Ceremony of
2022/23 Statistical Project Competition
for Secondary School Students

Professor Alan WAN, Ms Teresa CHAN and Professor CHEUNG Ka-chun presented prizes to winning teams

Social Media

As social media platforms have become integral to people's daily lives, a Facebook page and an Instagram for SPC have been set up for promoting the SPC starting from 2022/23 round. Interested parties can scan the QR codes below to follow our latest news.

Gratitude

The Organising Committee would like to express sincere gratitude to the patrons of the Competition, Mr Leo YU Chun-keung, Commissioner for Census and Statistics, and Ms Teresa CHAN Mo-ngan, Deputy Secretary for Education, for their support to the event. Special thanks to the Adjudication Panel and helpers who had paid tremendous contribution to the success of the 2022/23 SPC.

*Organising Committee for the 2022/23 SPC:

Ms Carly LAI	Census and Statistics Department
Mr CHAN Sau-tang	Education Bureau
Ms Sue CHENG	Census and Statistics Department
Mr Michael CHU	Census and Statistics Department
Miss Carmen LO	Census and Statistics Department
Mr Ian NG	Census and Statistics Department
Mr Hinz SHUM	Census and Statistics Department
Mr Stanley TSANG	Census and Statistics Department
Mr Wilson YUNG	Census and Statistics Department

News Section

◆ Personnel Changes (New Appointments, Promotions and Retirements)

Department of Statistics of the Chinese University of Hong Kong (CUHK)

※ Prof LI Gen has joined the Department of Statistics of CUHK as Assistant Professor with effective from August 2023.

※ Dr WRIGHT, John Alexander of Department of Statistics of CUHK has been promoted to Senior Lecturer with effective from August 2023.

Department of Management Sciences of City University of Hong Kong (CityU)

※ Dr Lilun DU joined the department as an Associate Professor with effect from September 2023.

※ Dr Baojun DOU joined the department as an Assistant Professor with effect from September 2023.

Department of Mathematics of the Hong Kong Baptist University (HKBU)

※ Dr Shunan YAO joined the department as an Assistant Professor with effect from September 2023.

※ Dr Heng PENG has been promoted to Professor with effect from September 2023.

News Section

◆ Personnel Changes (New Appointments, Promotions and Retirements) (Cont')

Department of Mathematics and Information Technology (MIT) of the Education University of Hong Kong (EdUHK)

※ Dr. SUEN Chun Kit Anthony has been promoted to Associate Professor with effect from July 2023.

※ Dr. SO Chi Fuk Henry has been promoted to Senior Lecturer I with effect from July 2023.

※ Dr. CHENG Kell Hiu Fai has been promoted to Senior Lecturer II with effect from July 2023.

※ Dr. WANG Dichen has joined MIT of EdUHK as Lecture I with effect from August 2023.

※ Ms. XIE Yishan Ellen has joined MIT of EdUHK as Lecture II with effect from August 2023.

◆ Result of the HKSS - John Aitchison Prize 2024

By the deadline of November 17, 2023, high quality submissions were received for the prize. After careful deliberations, the panel decided to award the HKSS - John Aitchison Prize in Statistics 2024 to Dr Guohao SHEN (2022 PhD Graduate at the Chinese University of Hong Kong) for his paper "Deep Nonparametric Regression on Approximate Manifolds: Non-asymptotic Error Bounds with Polynomial Prefactors" published in the *Annals of Statistics*, 2023, Vol. 51, No. 2, 691–716. This paper is a joint work with Yuling JIAO, Yuanyuan LIN, and Jian HUANG.

Dr Guohao SHEN

The research work has three notable contributions: (a) the error bounds of the estimator based on deep neural networks achieve the minimax optimal rate and improve upon existing results by depending polynomially on the predictor's dimension, rather than exponentially; (b) the study demonstrates that the neural regression estimator can overcome the curse of dimensionality under the assumption that the predictor is supported on an exact or approximate manifold; and (c) the work derives a novel approximation error bound for Hölder smooth functions using ReLU activated neural networks, which is of independent interest and has applications in various other deep learning problems. The research can be placed in the broader context of statistical machine learning, particularly deep learning, which has demonstrated remarkable empirical successes in various applications.

We would like to express our gratitude for the adjudication panel of the Prize. The Organising Committee for the HKSS - John Aitchison Prize in Statistics 2024 is as follows:

Chairperson	
Prof Alan WAN Tze-kin	City University of Hong Kong
Adjudication Panel	
Prof CHAN Ngai-hang	City University of Hong Kong (Chair of Adjudication Panel)
Prof HUANG Jian	Hong Kong Polytechnic University
Prof LU Zudi	University of Southampton
Executive Committee Members	
Dr Thomas LO Kar-kei	Census and Statistics Department
Mr Jason CHAN Chin-tang	Census and Statistics Department
Mr Michael LAU Tsz-ho	Census and Statistics Department