香港統計學會
Hong Kong Statistical Society
http://www.hkss.org.hk

Bulletin
Volume 45 No.1
April 2023

Editor's Foreword

Dear Members,

Welcome to the 2023 issue of the HKSS Bulletin. It's my honour to serve the HKSS as the Publication Secretary. I would like to express my gratitude to the predecessor, Dr Benson LAM, for his support and help in preparing this bulletin. I would also want to thank the members of the Editorial Board, Dr Billy LI and Mr Michael LAU for their valuable contributions to the Bulletin.

In this issue, we have the President's Forum. This issue of the Bulletin highlights three articles: one from Professor Ben DAI on his work on "Statistical Consistency in Ranking-Based Recommender Systems"; one from me to discuss an issue about an analysis on Carbon Trading in China Market; and one from Mr Matthew WONG and Mr Alex YIU on the calibration of 2021 Population Census results. Dr Wilson KWAN has arranged a workshop of data visualisation through R to some secondary school teachers. The Organising Committee of the 2021/22 Statistical Project Competition briefs us the successful completion of the Competition.

We would like to use this opportunity to express our special thanks to all contributors to this Bulletin and members of the Editorial Board.

Edmond CHAN

		Phone	Fax	Email
Editor	: Dr. CHAN, Edmond Chun-man, CUHK	3943 7935	-	chunmanchan@cuhk.edu.hk
Secretary	: Dr. LI, Billy Yeuk-goat, C&SD	3150 8941	3150 8993	-
Member	: Mr. LAU, Michael Tsz-ho, C&SD	3903 7042	2116 0370	thlau@censtatd.gov.hk

CONTENTS

(Vol. 45/No.1, April 2023)

President's Forum

Professor Alan WAN Tze-kin

The COVID-19 pandemic has posed unprecedented challenges to the world, and Hong Kong has not been immune to its effects. However, the Census and Statistics Department (C&SD) and the Hong Kong Statistical Society (HKSS) have remained steadfast in their commitment to promoting statistical knowledge and excellence.

Commemorating the 55th Anniversary of the C&SD and the 45th Anniversary of the HKSS, a seminar was held on November 11, 2022, at the Chiang Chen Studio Theatre of the Hong Kong Polytechnic University. The seminar aimed to provide insights into various statistical topics and issues related to census and survey data. Distinguished speakers, including Professor Ian McKEAGUE of City University of Hong Kong, Professor Tarani CHANDOLA of University of Hong Kong, and young professional staff of C&SD, presented on the theme of "Perspectives on Uncertainty and Error of Statistics".

Furthermore, a workshop on "Data Visualisation with Power BI" was held on June 17, 2022, at PolyU's Hung Hom Bay Campus. Co-organized by HKSS and EDB, the workshop was initially planned for January 2022 but was postponed due to the pandemic. The interactive workshop was well-received by the twelve teachers who attended, with many expressing appreciation for the knowledge and materials covered and found the experience to be highly beneficial.

To commemorate the 45th Anniversary of the HKSS, we enlisted the help of a company to design an Anniversary Logo for promotional purposes.

The Inaugural HKSS-John Aitchison Prize in Statistics 2023 received high quality submissions by the October 31, 2022 deadline. The panel, after careful deliberations, awarded the prize to Dr. Jingming WANG (2021 PhD Graduate at the Hong Kong University of Science and Technology) for the co-authored paper titled "Statistical inference for principal components of spiked covariance matrices".

Besides, I am happy to share with you the news that I was recently invited for an interview by the "Significance" Magazine, an international statistical magazine produced jointly by the Royal Statistical Society, American Statistical Association and the Australian Statistical Society. During the interview, I had the opportunity to discuss some of the activities of the HKSS. You may refer to the interview content through this link.

The C&SD and the HKSS remain committed to promoting statistical knowledge and excellence in Hong Kong, despite the challenges posed by the pandemic. These events and activities are a testament to their unwavering commitment to the local statistical community.

Thank you for your continued support of our mission.

Statistical Consistency in Ranking-based Recommender Systems

Dr Ben DAI

The Chinese University of Hong Kong

1. Introduction

Due to the extraordinary development of big data, recommender systems are proposed to predict users' preferences over various items by borrowing similar information from other users or items, thus facilitate users to search the best needed item among massive options. It has become a crucial part of e-commerce, with applications in restaurant guides (Entree; [1]), movie rentals (MovieLens; [2]), personalized e-news (Daily learner; [3]) and book recommendations (Amazon; [4]).

The performance of recommender systems highly depends on how to pool the information from similar users and items. In this note, recommender systems predict a user's preference for a large number of items through user-item specific information, and a relatively small number of observed preference feedbacks. Many machine learning and statistical methods emerge for formulating recommender systems, to predict unknown ratings by averaging over similar users' ratings with weights; such as the matrix/tensor factorization approach [5], regularized singular value decomposition (regularized SVD; [6]), probabilistic latent semantic analysis (pLSA; [7]), latent Dirichlet allocation (LDA; [8]), and restricted Boltzmann machines (RBM; [9]). In particularly, the regularized matrix factorization has become popular due to its peak performance and scalability in computation in real applications.

Despite the success of the existing recommender systems, in practice, recommendations are more preferred for providing a short or Top-K list of top-preferred items as opposed to a complete list of items with estimated preference scores. Therefore, a ranking approach has become more relevant than classification or regression in recommender systems.

2. Background

In ranking-based recommender systems, we consider a training set with ranking data of $(\mathbf{x}_{il}, y_{il})$; $i = 1, \ldots, n$; $l = 1, \ldots, m$, where $x_{il} \in X \subset \mathbb{R}^p$ is the joint features of the i-th query $\mathbf{q}_i$ and the l-th document $\mathbf{d}_l$. Specifically, x_{il} can be joint textual raw tokens of the query-document pair, or the preprocessed handcrafted features, including text match, document statistics, and topical matching [10]. Moreover, $y_{il} \in R$ is the relevance score between the i-th query and the l-th document, and n is the number of queries, and m is the number of documents. The primary goal is to construct a ranking function that can provide an appropriate ranking of all the documents for each query.

It is sensible to assume that samples $\mathbf{z}_i = ((\mathbf{x}_{i1}, y_{i1}), \ldots, (\mathbf{x}_{im}, y_{im}))$; $(i = 1, \ldots, n)$ are independent and identically distributed samples following an unknown joint probability measure on $\mathbf{Z} = (\mathbf{X}_{1:m}, \mathbf{Y}_{1:m})$, where $\mathbf{X}_{1:m} = (\mathbf{X}_1, \cdots, \mathbf{X}_m)^T$ and $\mathbf{Y}_{1:m} = (Y_1, \cdots, Y_m)^T$. Note that $(\mathbf{X}_l, Y_l)$; $(l = 1, \ldots, m)$ are dependent, since they share the same query information, yet after conditional on query information $(\mathbf{X}_l, Y_l)|\mathbf{Q}_l$; $(l = 1, \ldots, m)$ are iid samples.

3. Statistical consistency

A number of metrics have been proposed to access the ranking performance. We mainly focus on the pairwise zero-one loss. Specifically, given a (listwise) ranking function $\mathbf{f}(\mathbf{X}_{1:m}) : \mathbb{R}^{m\times p} \rightarrow \mathbb{R}^m$, the pairwise zero-one loss is defined as:

$$\text{PairLoss}\left(\mathbf{Y}_{1:m}, \mathbf{f}(\mathbf{X}_{1:m})\right) = \sum_{1\le l<l'\le m} \mathbf{1}\left((Y_l - Y_{l'})\left(f_l(\mathbf{X}_{1:m}) - f_{l'}(\mathbf{X}_{1:m})\right) \le 0\right), \tag{1}$$

where $\mathbf{1}(.)$ is an indicator function. The pairwise loss evaluates the relative order of any two items. Indeed, PairLoss is non-convex, which indicates a source of the computation difficulty. To address this issue, a surrogate loss $\phi(\mathbf{Y}, \mathbf{f}(\mathbf{X}_{1:m}))$ is usually introduced to enable a trackable estimation procedure. Taken together, the estimated (listwise) ranking function is obtained from:

$$\widehat{\mathbf{f}} = \underset{\mathbf{f}\in\mathscr{F}}{\text{argmax}} \frac{1}{n}\sum_{i=1}^{n} \phi\left(\mathbf{y}_{1:m}, \mathbf{f}(\mathbf{x}_{1:m})\right), \tag{2}$$

where $\mathscr{F}$ is a class of candidate ranking functions. On this ground, it is rather important to check if the estimator based on a surrogate loss coincides with the best ranker with respect to PairLoss. Therefore, we present the statistical consistency in ranking-based recommender systems. To proceed, we give the definition of Bayes ranker.

Lemma 1 (Bayes ranker) $\mathbf{f}^*$ *is the best ranker (a global minimizer) w.r.t.* $\mathbf{E}PairLoss(\mathbf{Y}_{1:m}, \mathbf{f}(\mathbf{X}_{1:m}))$ *if and only if* $\mathbf{f}^*$ *satisfies that*

$$\left(\mathbf{f}_l^*(\mathbf{X}_{1:m}) - \mathbf{f}_{l'}^*(\mathbf{X}_{1:m})\right)\left(\mathbb{P}(Y_l > Y_{l'}|\mathbf{X}_l, \mathbf{X}_{l'}) - 1/2\right) \ge 0, \quad \textit{for } 1 \le l < l' \le m; \textit{ almost surely.} \tag{3}$$

Statistical consistency provides reassurance that optimizing a surrogate does not hinder the search for a function that achieves the optimal Bayes ranking risk, and thus admit such a search to proceed within the scope of computationally scalable algorithms. To carry this agenda, three kinds of statistics consistency are introduced in ranking problem to measure the quality of the surrogate loss ϕ.

Definition 1 (Fisher consistency) *A surrogate loss* ϕ *is Fisher consistency with PairLoss if and only if*

$$\mathbf{f}^* \in \underset{\mathbf{f}}{\text{argmin}}\, \mathbb{E}\phi\left(\mathbf{Y}_{1:m}, \mathbf{f}(\mathbf{X}_{1:m})\right),$$

where $\mathbf{f}^*$ *is defined in Lemma 1.*

Fisher consistency is the weakest possible condition on φ: the best ranker on φ should coincide with the Bayes ranker defined in Lemma 1.

Definition 2 (Risk consistency) *If for every sequence of ranking functions* $\mathbf{f}^{(k)}$*:*

$$\mathbb{E}\phi\left(\mathbf{Y}_{1:m},\mathbf{f}^{(k)}(\mathbf{X}_{1:m})\right)\to\mathbb{E}\phi\left(\mathbf{Y}_{1:m},\mathbf{f}^{*}(\mathbf{X}_{1:m})\right),\quad implies\quad \mathbb{E}PairLoss\left(\mathbf{Y}_{1:m},\mathbf{f}^{(k)}(\mathbf{X}_{1:m})\right)\to\mathbb{E}PairLoss\left(\mathbf{Y}_{1:m},\mathbf{f}^{*}(\mathbf{X}_{1:m})\right),$$

then the surrogate loss $\phi(.)$ *is ranking-consistency.*

Note that [11] suggests that if ϕ is continuous and Fisher consistency, then risk consistency is hold for binary classification. However, in ranking problem, it is not valid in general.

Definition 3 (Excess risk bounds) *If for every ranking function* $\mathbf{f}$*, there exists* $c>0$ *and* $\alpha>0$*,*

$$\left|\mathbb{E}PairLoss\left(\mathbf{Y}_{1:m},\mathbf{f}(\mathbf{X}_{1:m})\right)-\mathbb{E}PairLoss\left(\mathbf{Y}_{1:m},\mathbf{f}^{*}(\mathbf{X}_{1:m})\right)\right|\le c\left|\mathbb{E}\phi\left(\mathbf{Y}_{1:m},\mathbf{f}(\mathbf{X}_{1:m})\right)-\mathbb{E}\phi\left(\mathbf{Y}_{1:m},\mathbf{f}^{*}(\mathbf{X}_{1:m})\right)\right|^{\alpha},\quad (4)$$

then the surrogate loss $\phi(.)$ *is excess risk consistency.*

Definition 3 indicates that the excess risk on PairLoss can be upper bounded by the surrogate loss ϕ. For example, if the convergence rate of the regret of $\hat{f}$ on ϕ is $O_P(1/n)$, then based on Definition 3, its convergence rate on PairLoss is $O_P(1/n^{\alpha})$. Therefore, $\alpha>0$ is the parameter indicating the quality of ϕ, and a large α is preferred.

4. References

[1] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[2] Bradley N Miller, Istvan Albert, Shyong K Lam, Joseph A Konstan, and John Riedl. Movielens unplugged: Experiences with an occasionally connected recommender system. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pages 263–266. ACM, 2003.

[3] Daniel Billsus and Michael J Pazzani. User modeling for adaptive news access. *User Modeling and User-Adapted Interaction*, 10(2-3):147–180, 2000.

[4] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, (1):76–80, 2003.

[5] Andrey Feuerverger, Yu He, and Shashi Khatri. Statistical significance of the Netflix challenge. *Statistical Science*, 27(2):202–231, 2012.

[6] Arkadiusz Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD Cup and Workshop*, pages 5–8, 2007.

[7] Thomas Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, 22(1):89–115, 2004.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. Journal *of Machine Learning Research*, 3(Jan):993–1022, 2003.

[9] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning*, pages 791–798. ACM, 2007.

[10] Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. In Yahoo*! Learning to Rank Challenge*, pages 1–24, 2011.

[11] Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE transactions on information theory*, 51(1):128–142, 2005.

A time-series analysis and approach on Carbon Trading

T. Z. WONG, Department of Statistics, CUHK
H. T. TSANG, Department of Systems Engineering and Engineering Management, CUHK
C. M. CHAN, Department of Statistics, CUHK

The paper will be organized as follows. We will firstly start with the discussions of the background about carbon trading. Secondly, we would like to review the existing literatures to understand what other scholars had done successfully. Thirdly, we will propose a Time Series (TS) approach and methodology for the analysis of the central contribution of this paper in relation to carbon trading. Then we follow by the concluding statements and discussions of future works.

1. Background about carbon trading

Global warming is a worldwide issue which deteriorated in recent years. To reduce the impact, Paris Agreement was signed by 196 parties and committed to controlling the temperature rising below 2 degrees Celsius (UN,2022). China which is the largest carbon emission country (World Bank, 2020) announced that will strive to peak carbon dioxide emissions before 2030 and achieve carbon neutrality before 2060 (UN, 2021).

The carbon emission trading scheme (ETS) is one of the measures to lessen greenhouse gas. Guangzhou Emissions Exchange (CEEX, 2019) was launched in 2012 and is the first Emissions Exchange in China whose total volume of spot exceeded 100 million tons and a total turnover surpassed 2 billion yuan (CEEX,2019). The trading price reflected part of the effort of emission reduction. Since it would be considered an internal cost of the high-polluting industries or a profit of the green-tech companies. The rest of the paper will discuss the existing methods of forecasting the trading price, using the time series (TS) method to analyze the CEEX trading price trend and summarize the future tendency.

2. Literature review

To predict the trading price, and trading volume more accurately, Lu and his team used 6 different machine learning models. (Lu et al., 2020).The six machine learning models include radial basis function neural network (RBFNN) (Dhanalakshmi et al., 2009), support vector machine using particle swarm optimizer (PSO-SVM) (Wang and Li, 2019), support vector machine using simulated annealing and fruit fly optimization algorithm (SA-FFOA-SVM) (Lu et al. 2019), random forest (RF)(Breiman, 2001), extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016), kernel-based nonlinear extension of the Arps decline model optimized by grey wolf optimizer (GWO-KENA)(Ma and Liu, 2018).

3. Analysis by TS method

3.1 Data analysis

This study constructs a data set of CEEX market information from 2016 to 2021. According to the data shown by CEEX (2022), it reveals the following information.

Statistical characteristics					
Number of data	Number of test data	Maximum	Minimum	Mean	Standard deviation
1134	126	30.84	8.10	19.12	6.41

Table 1. Statistical characteristics of the data

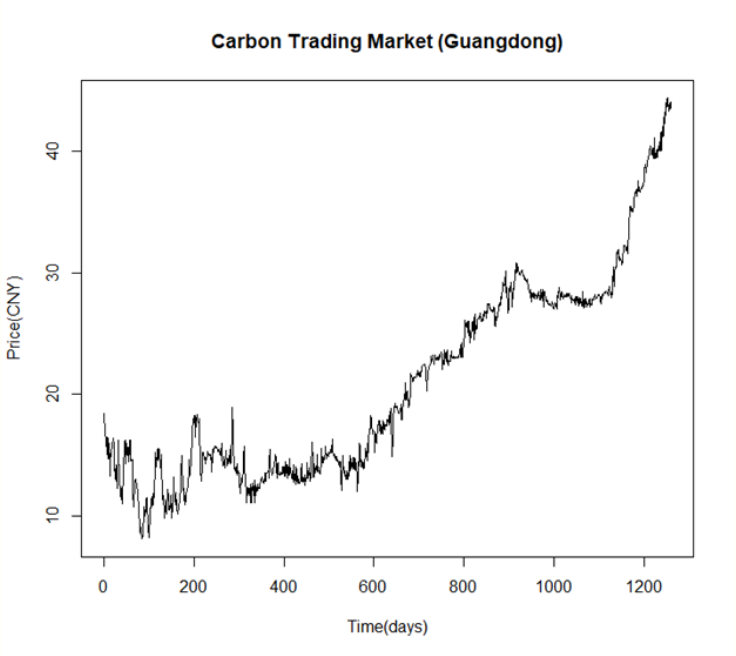


Figure 1. Time Series plot of the carbon market (Guangdong)

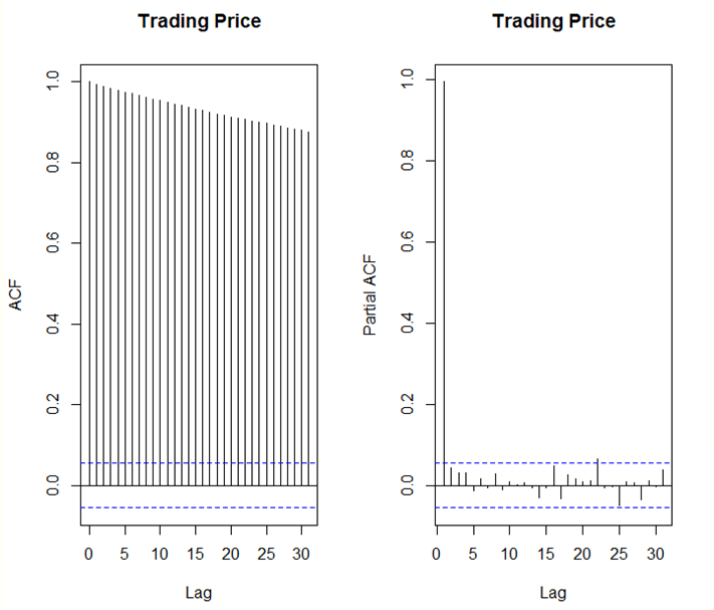


Figure 2. ACF & PACF of the data

The time series plot in Figure 1 shows that the variance and mean are fluctuating over the time. Hence the data is nonstationary.

Since the data consisted by $Xt = T_t + S_t + N_t$, where X is the trading price, T is the trend, S is the seasonal effect and N is the noise (residual). To obtain a stationary data, the trend and seasonality will be removed.

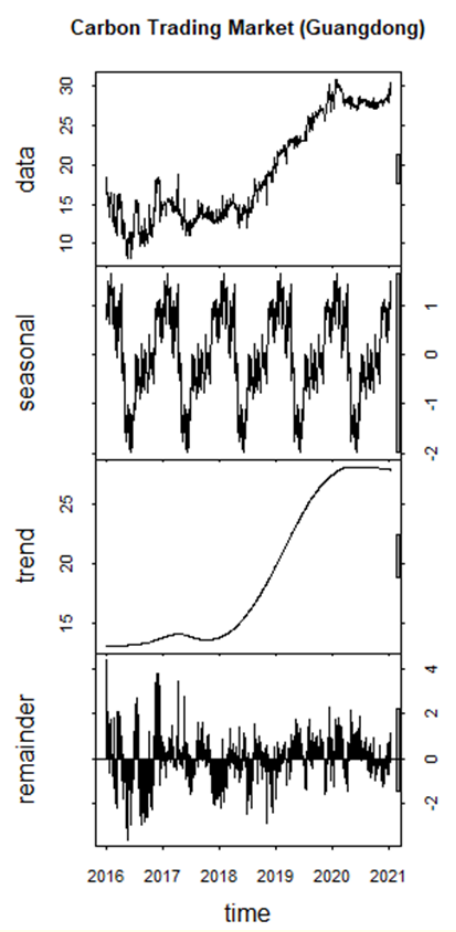


Figure 3. Data decomposition

Moving average is one of the methods to obtain the trend and seasonal information. The data is filtered by moving average method. The trading price, trend and the residual are as follows:

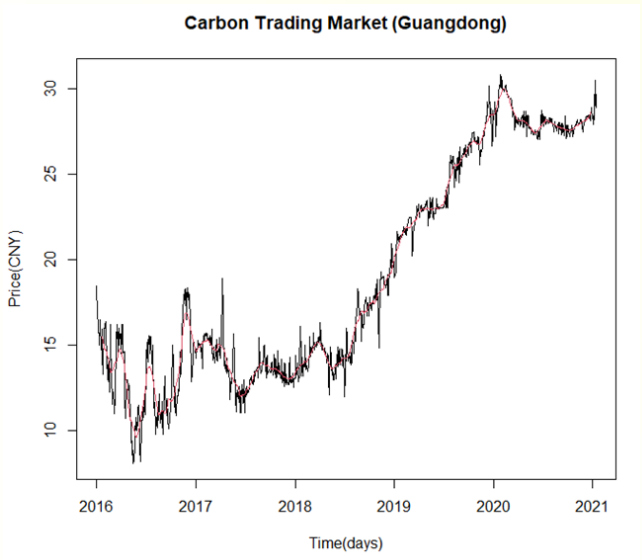


Figure 4. Trend estimation

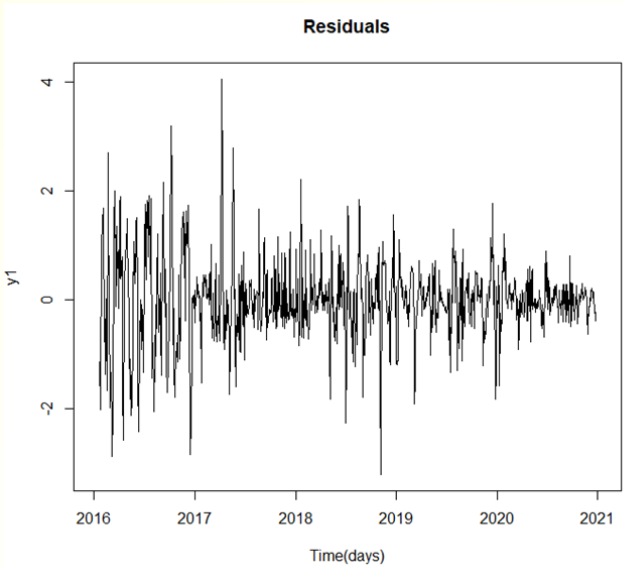


Figure 5. Residuals without trend

3.2 Model identification

After removing the trend and seasonal effect, the residuals and related ACF and PACF plots are as follows:

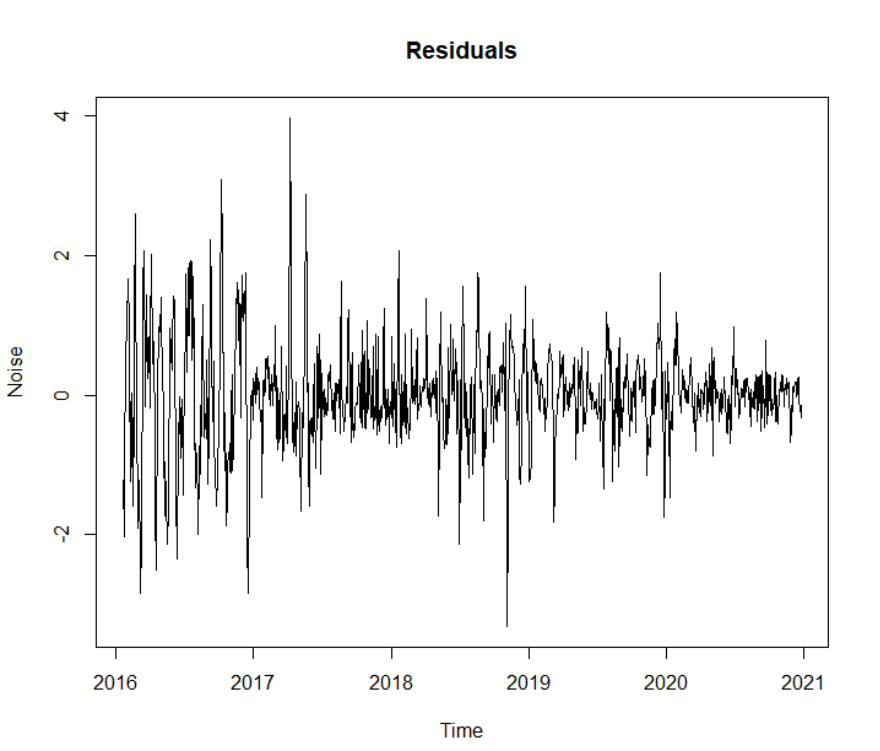


Figure 6. Residuals without trend and seasonality

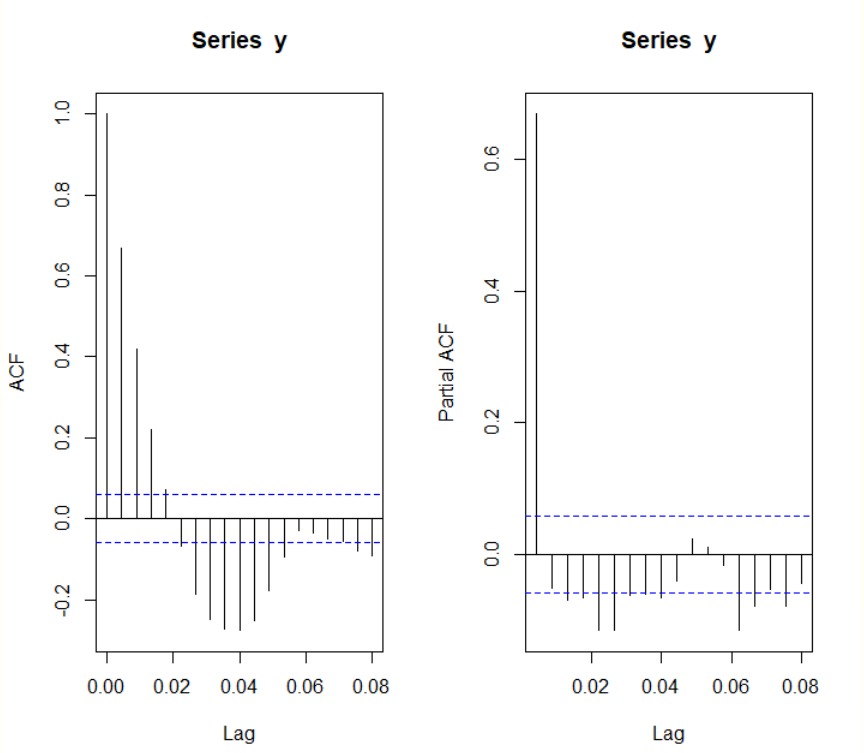


Figure 7. ACF & PACF of residuals

The ACF and PACF plots indicate that there is a strong relationship at certain lags, making model identification easier. It is possible that AR(1), MA(1), ARMA (5,4) may fit the model.

3.3 Model fitting

The data will be fitted into each candidate, and their Akaike's Information Criterion (AIC) will be obtained for comparison.

After comparison, the ARMA(5,4) model has the smallest AIC value, which will be selected.

The model is:

$$(1 - 2.1953B + 0.9232B_2 + 1.6089B_3 - 1.8853\,B_4 + 0.5900B_5)Y_t = (1 + 1.6129B + 0.0455B_2 - 1.6010B_3 + 0.9425B_4)Z_t$$

where $Z_t \sim N(0, 0.2629)$

3.4 Model diagnostics

To check how well the model fits the data, residual analysis using ACF test and Ljung-Box test are chosen. The plots are as follows:

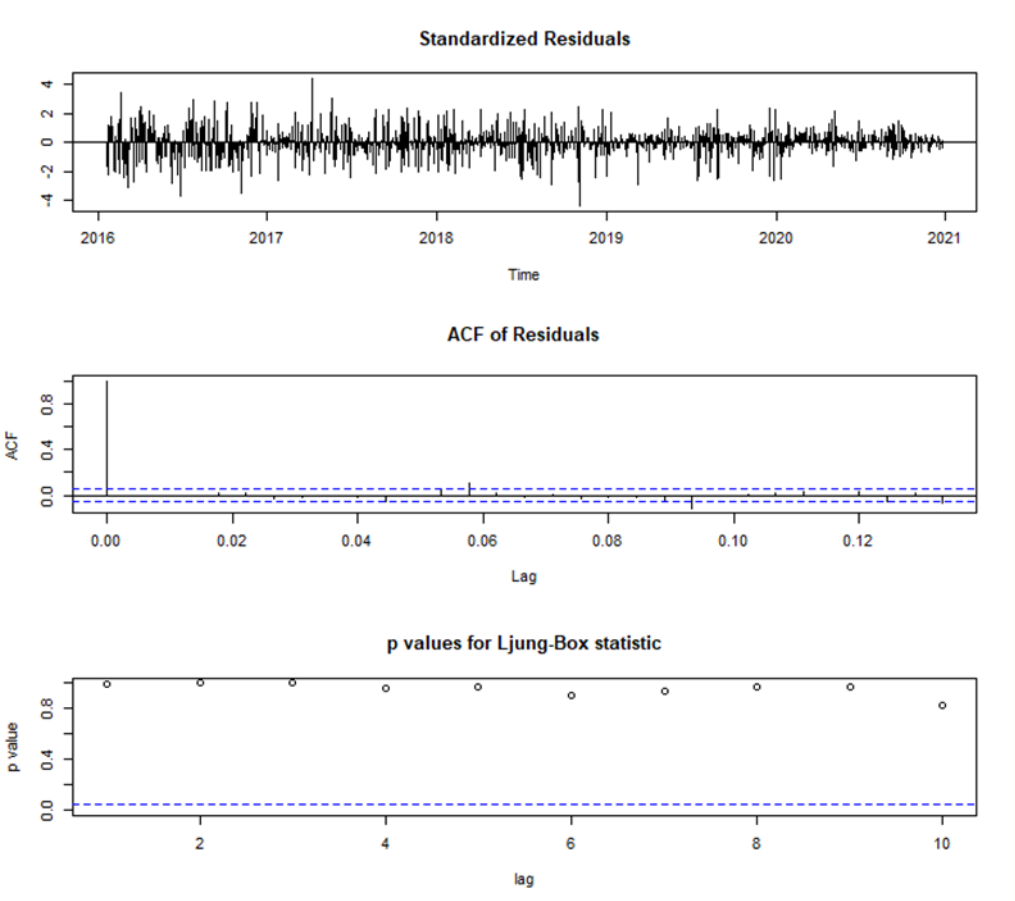


Figure 8. Ljung-box test

From the plots, it is obvious that all the p-values are larger than 0.05 which means it passes the Ljung-Box statistic and fit the model.

Portmanteau test statistic is another method to check the goodness of fit the model. The Portmanteau test statistic is given by:

$$Q(h) = n(n+2)\sum_{j=1}^{h}\frac{\hat{r}_z^2(j)}{n-j}$$

which follows a Chi-squared distribution with degrees of freedom h-p-q.

If Q(h) is bigger than the 95% percentile of the χ^2(h-p-q) distribution, we reject $H_0: \hat{z}_t \sim WN$

Since the Portmanteau test does not reject H_0, implying that the model is a good fit to the data.

3.5 Prediction

Prediction by Holt-Winters approach.

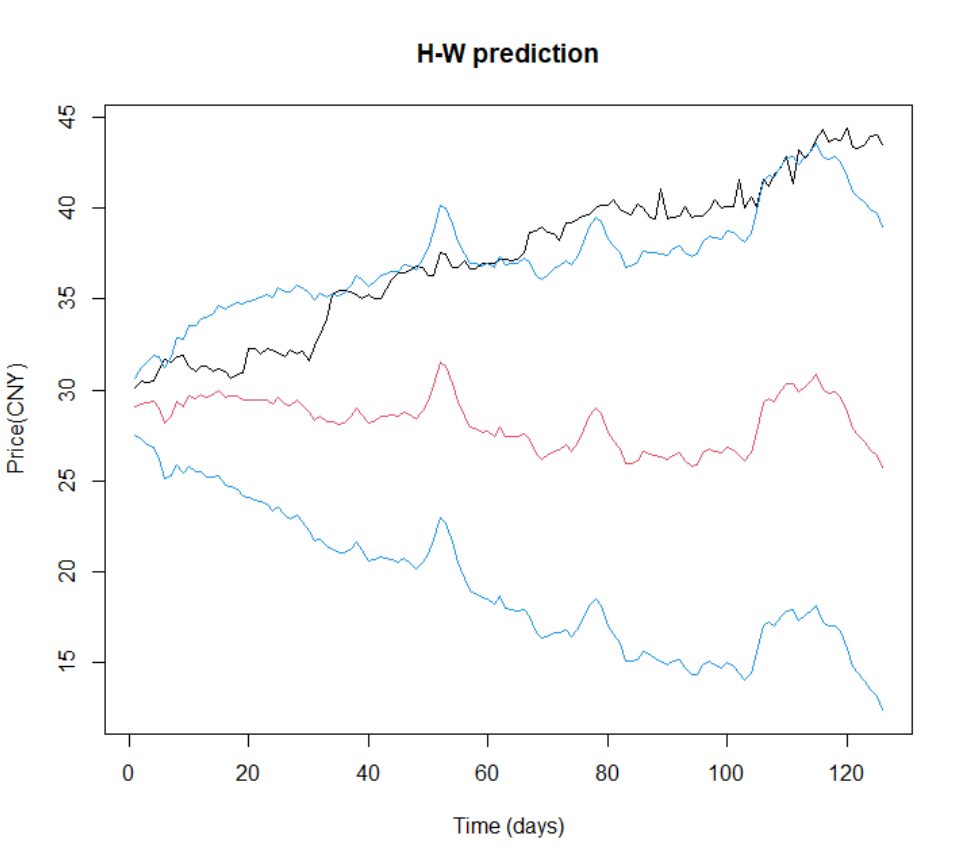


Figure 9. Holt-Winters prediction, prediction for the trading price

The black line is the actual price. The red line represents the predicted value of each day. The blue lines are the upper and lower bounds of a 95% prediction interval on the daily trading price.

The graph shows that at the beginning of the time, the actual price is within the boundaries. After a period of time, there is a difference between the actual price and the predicted price. It is because of the other unexpected factors, like the policy, economic environment.

4. Conclusions and future works

This study shows that the tendency of a carbon price has been rising in recent years. To reduce carbon emissions, carbon trading is one of the big directions. The higher trading price may enforce the carbon-intensive industries to close and encourage the companies to improve their energy efficiency. It may benefit the development of green industries by selling the carbon emission allowance.

China is the largest emitter of carbon dioxide, which means it has the potential largest trading volume in the carbon market around the world. In order to build a global carbon trading market and attract the international investor, the trading price may be correlated to the European Union Emission Trading System which is a developed market. Hence there are some future research directions. First direction is the tendency between the EU market and China market. The second one is to consider more factors such as the policy and economic indicators, that are expected to have obvious impacts on the carbon price in the forecasting model for a higher accuracy of the prediction.

5. References

[1] Breiman, L., 2001. Random forests. Machine Learning. 45 (1), 5-32.

[2] Chen, T., Guestrin, C., 2016, August. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785-794.

[3] Dhanalakshmi, P., Palanivel, S., Ramalingam, V., 2009. Classification of audio signals using SVM and RBFNN. Expert Syst. Appl. 36 (3), 6069-6075.

[4] Guangzhou Emissions Exchange (CEEX), 2019. About us. http://www.cnemission.com/article/gywm/201907/20190700001675.shtml (Accessed 1 August 2022).

[5] Guangzhou Emissions Exchange Historical data (CEEX), 2022. https://www.cnemission.com/article/hqxx/ (Accessed 1 August 2022).

[6] Lu, H., Azimi, M., Iseley, T., 2019a. Short-term load forecasting of urban gas using a hybrid model based on improved fruit fly optimization algorithm and support vector machine. Energy Rep. 5, 666-677.

[7] Lu, H., Ma, X., Huang, K., & Azimi, M., 2020. Carbon Trading Volume and price forecasting in China using multiple machine learning models. *Journal of Cleaner Production*, *249*, 119386. https://doi.org/10.1016/j.jclepro.2019.119386

[8] Ma, X., Liu, Z., 2018. Predicting the oil production using the novel multivariate nonlinear model based on Arps decline model and kernel method. Neural Comput. Appl. 29 (2), 579-591.

[9] United Nations (UN), 2022. Paris agreement. https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement (Accessed 1 August 2022).

[10] United Nations (UN), 2021. China headed towards carbon neutrality by 2060; President Xi Jinping vows to halt new coal plants abroad. https://news.un.org/en/story/2021/09/1100642 (Accessed 1 August 2022).

Calibration of 2021 Population Census Results Using Generalised Regression Estimation Method

Matthew WONG Tsz-lim and Alex YIU Cheuk-wing
Census and Statistics Department

Background

It is an established practice from 1961 for Hong Kong to conduct a population census once every 10 years and a by-census in the middle of the intercensal period. The 2021 Population Census (21C) was conducted in June to August 2021. Population Censuses provide up-to-date benchmark information on the socio-economic characteristics of the population of Hong Kong and on its geographical distribution. Such statistics are vital to the planning and policy formulation of the Government. In the 21C, about nine-tenths of the households were subject to simple enumeration with the "Short Form" to provide basic demographic information of their household members, while the remaining one-tenth of the households were subject to more detailed enquiry with the "Long Form" on a broad range of demographic and socio-economic characteristics of their household members. The "Short Form" covered only some basic questions. On the other hand, the "Long Form" included not only those questions covered in the "Short Form", but also additional ones relating to the socio-economic characteristics of the population and the characteristics of households and quarters. Taking together the common information collected in both the "Short Form" and "Long Form" gave the complete enumeration results on basic characteristics of the population, e.g. the number of persons by sex and age. Meanwhile, the complete enumeration results aforesaid were also used as auxiliary variables, forming the basis for estimation of the detailed socio-economic characteristics collected through the "Long Form".

2. In the estimation process of the 21C, C&SD used the Generalised Regression Estimation (GRE) to conduct calibration. If the traditional method were used, i.e. reciprocals of inclusion probabilities as grossing-up factors (GUFs), inconsistency would occur between the "Long Form" estimates and the corresponding control totals obtained from the complete enumeration results on auxiliary variables. Thus, the GRE was used instead. The GRE adjusted initial GUFs to yield calibrated GUFs with aim of reconciling the "Long Form" estimates to the corresponding control totals on auxiliary variables, while minimising the adjustment distances. This enabled better utilisation of the common information collected in both the "Short Form" and "Long Form" and thus improved the precision and consistency of small areas and population sub-groups estimates. Furthermore, the GRE exploited the linear relationship between the auxiliary variables and the detailed socio-economic characteristics collected through the "Long Form", thereby improving the precision of the estimates relating to the detailed socio-economic characteristics. Therefore, the GRE had indeed been widely used internationally.

3. Consider a finite population of N quarters, a "Long Form" sample S of n quarters and p control totals. Define:

X_k: the control total for the k^{th} auxiliary variable;

d_i: the initial GUF for the i^{th} quarter, which is the reciprocal of the inclusion probability;

w_i: the calibrated GUF for the i^{th} quarter;

x_{ki}: the count of the k^{th} auxiliary variable for the i^{th} quarter;

$\hat{X}_{HTk}$: the estimator of the total for the k^{th} auxiliary variable derived using the initial GUFs, i.e. the Horvitz-Thompson estimator;

$\hat{X}_{GREGk}$: the estimator of the total for the k^{th} auxiliary variable derived using the calibrated GUFs; and

q_i: the chosen constant for defining the importance of the i^{th} quarter to the distance function (expression (i) below), and

$q_i = \sqrt{\sum_{j=1}^{202} RP_{j,i}^2}$, where

$RP_{1,i}$ is the count of male residents aged 0 living in the i^{th} quarter; …;

$RP_{101,i}$ is the count of male residents aged 100+ living in the i^{th} quarter;

$RP_{102,i}$ is the count of female residents aged 0 living in the i^{th} quarter; …;

$RP_{202,i}$ is the count of female residents aged 100+ living in the i^{th} quarter.

And also define the following vectors:

$$\vec{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ \vdots \\ x_{pi} \end{bmatrix}, \vec{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ \vdots \\ X_p \end{bmatrix}, \vec{\hat{X}}_{HT} = \begin{bmatrix} \hat{X}_{HT_1} \\ \hat{X}_{HT_2} \\ \vdots \\ \vdots \\ \hat{X}_{HT_p} \end{bmatrix} = \sum_{i \in S} d_i \vec{x}_i \text{ and } \vec{\hat{X}}_{GREG} = \begin{bmatrix} \hat{X}_{GREG_1} \\ \hat{X}_{GREG_2} \\ \vdots \\ \vdots \\ \hat{X}_{GREG_p} \end{bmatrix} = \sum_{i \in S} w_i \vec{x}_i \Bigg) (i),$$

4. The GRE aims to adjust the initial GUFs d_i (i = 1, …, n) to yield the calibrated GUFs w_i (i = 1, …, n) to reconcile the "Long Form" estimates $\hat{X}_{GREGk}$ (k = 1, …, p) to the corresponding control totals X_k (k = 1, …, p) on auxiliary variables, while minimising the adjustment distances between the calibrated GUFs w_i (i = 1, …, n) and the initial GUFs d_i (i = 1, …, n). This represents a classical constrained minimisation problem and the calibrated GUFs w_i (i = 1, …, n) can be obtained using the method of Lagrange multipliers by differentiating expression (i) below with respect to w_i (i = 1, …, n) and λ_k (k = 1, …, p) (dummy variables called Lagrange multipliers) and equating it to zero:

$$L = \sum_{i \in S} q_i d_i G\left(\frac{w_i}{d_i}\right) + \sum_{k=1}^{p} \lambda_k \left(X_k - \sum_{i \in S} w_i x_{ki} \right) (i),$$

where $G\left(\frac{w_i}{d_i}\right)$ is a strictly convex function to measure the distance of $\frac{w_i}{d_i}$ from 1, with a natural choice being $\frac{1}{2}\left(\frac{w_i}{d_i} - 1\right)^2$ $(i = ,1, \ldots, n)$.

5. It can be shown that the closed-form solutions for Lagrange multipliers $\vec{\lambda}$ and the calibrated GUFs w_i ($i = 1, \ldots, n$) are:

$$\vec{\lambda} = \left(\sum_{i \in s} \frac{d_i}{q_i} \vec{x}_i \vec{x}_i'\right)^{-1} \left(\vec{X} - \hat{\vec{X}}_{HT}\right) \quad (ii)$$

$$w_i = d_i \left(1 + \frac{\vec{\lambda}' \vec{x}_i}{q_i}\right) = d_i \left[1 + \frac{\vec{x}_i'}{q_i} \left(\sum_{i \in s} \frac{d_i}{q_i} \vec{x}_i \vec{x}_i'\right)^{-1} \left(\vec{X} - \hat{\vec{X}}_{HT}\right)\right] \quad (iii)$$

6. Define y_i the count of the detailed socio-economic characteristics of the i^{th} quarter. The GRE estimator of the total for the detailed characteristics $\hat{Y}_{GREG}$ was:

$$\hat{Y}_{GREG} = \sum_{i \in s} w_i y_i = \sum_{i \in s} d_i \left[1 + \frac{\vec{x}_i'}{q_i} \left(\sum_{i \in s} \frac{d_i}{q_i} \vec{x}_i \vec{x}_i'\right)^{-1} \left(\vec{X} - \hat{\vec{X}}_{HT}\right)\right] y_i$$

$$\hat{Y}_{GREG} = \sum_{i \in s} d_i y_i + \left[\left(\sum_{i \in s} \frac{d_i}{q_i} \vec{x}_i \vec{x}_i'\right)^{-1} \sum_{i \in s} \frac{d_i}{q_i} \vec{x}_i y_i\right]' \left(\vec{X} - \hat{\vec{X}}_{HT}\right)$$

$$\hat{Y}_{GREG} = \sum_{i \in s} d_i y_i + \vec{\hat{\beta}}' \left(\vec{X} - \hat{\vec{X}}_{HT}\right), \quad \text{where} \quad \vec{\hat{\beta}} = \left(\sum_{i \in s} \frac{d_i}{q_i} \vec{x}_i \vec{x}_i'\right)^{-1} \sum_{i \in s} \frac{d_i}{q_i} \vec{x}_i y_i$$

7. The neat, closed-form solution *(iii)* reveals the required inputs for the GRE, namely the initial GUFs d_i ($i = 1, \ldots, n$) , control totals $\vec{X}$, and a "Long Form" quarters dataset containing $\vec{x}_i$ $(i = 1, \ldots, n)$ with a dimension of $n \times p$.

8. A major drawback of the GRE estimator above is that negative or extremely large/small calibrated GUFs can be resulted. This problem has been tackled by truncating the ratios of the calibrated GUFs to the initial GUFs, i.e. w_i/d_i ($i = 1, \ldots, n$), to fall within an upper limit (U) and lower limit (L). With the application of truncation, the GRE involves solving of a system of nonlinear equations with the truncated linear function, which is complicated in computation. Thus, the calibrated GUFs are usually approximated using numerical methods. In the 21C, the truncation upper limit and lower limit were set to be 3.0 and 0.3 respectively. R package "nleqslv", a useful package for solving a system of nonlinear equations using Broyden's method, was used to approximate the calibrated GUFs. In other words, Broyden's method was used to search the numerical solution of the calibrated GUFs ($i = 1, \ldots, n$) that minimised the distance function (expression (iv)):

$$\sum_{i \in S} q_i d_i G\left(\frac{w_i}{d_i}\right) \quad (iv), \quad \text{subject to}$$

$$X_k = \sum_{i \in S} w_i x_{ki} \quad (k = 1, \ldots, p) \quad \text{and}$$

$$L \leq \frac{w_i}{d_i} \leq U \quad (i = 1, \ldots, n)$$

Application of the GRE in the 21C

9. The operation process of the GRE in the 21C involved two key procedures as follows:

Procedure 1: Determination of the choice of control totals

10. The 21C covered all quarters, households and population in Hong Kong were covered. Complete enumeration results obtained by taking together the common information collected in both the "Short Form" and "Long Form" could provide control totals of basic characteristics for quarters, households and population at refined geographical level, e.g. control totals for the number of occupied quarters by type of quarters and District Council district (DCD). Table 1 summarises the control totals adopted in the 21C estimation process.

Table 1: Control totals adopted in the 21C

Type of control totals	x Geographical demarcation	x Basic characteristics
Occupied quarters	18 DCDs, 452 District Council Constituency Areas (DCCAs), selected housing estates and building groups	Type of quarters
Household		Household size
Population		Sex, age, ethnicity and resident status

11. In determining the choice of control totals, due consideration was given to the following aspects:

I. *Consistency between the benchmark data and the complete enumeration results* – The results of population censuses provide important benchmark data for the compilation of a wide range of other population figures. These benchmark data are taken together with statistical data generated from administrative systems and sample surveys to form a population statistical database. In view of the importance of these benchmark data, C&SD included the relevant complete enumeration results as control totals in the estimation process, especially those relating to small areas, to ensure that the benchmark data were consistent to the complete enumeration results at the refined geographical level.

II. *Small areas and population sub-groups – To produce more reliable statistics for smaller areas and population sub-groups, control totals at refined geographical level (e.g. housing estates and building groups) and for population sub-groups (e.g. ethnicities) were adopted. Empirical results of the GRE in the 21C revealed that when totals of small population aforesaid were not* controlled, the GRE might yield estimated totals of the small population that deviated from both the complete enumeration results and design-based estimates (reciprocals of inclusion probabilities as GUFs) in some degree. As such, depending on the actual outcome of the GRE, relevant control totals of small areas and population sub-groups were included in the estimation process when necessary.

III. *Precision of estimates of detailed characteristics* – The GRE exploited the linear relationship between the auxiliary variables and detailed characteristics to produce more precise estimates of detailed characteristics. Since many detailed characteristics of quarters, households and population were highly correlated with the basic characteristics such as type of quarters, household size, sex, age, ethnicity and resident status, control totals relating to these basic characteristics were included in the 21C.

12. On the other hand, in order to avoid excessive GUF adjustments, dramatic changes in distributions of GUFs and extremely calibrated GUFs / highly-skewed distributions of the calibrated GUFs, the number of the relevant "Long Form" records must be sufficiently large for each of the control totals concerned and thus the number of control totals in the GRE shall be kept reasonable. In this connection, the following processes were taken to refine the final sets of control totals used in the 21C:

I. *Dropping of control totals with insufficient number of "Long Form" records*– For example, it had been planned to include the control totals of population for each of the top 10 ethnic groups in each DCD. However, in conducting quality assurance for the GRE, it was found that controlling the totals of population of the ethnic groups with very few "Long Form" records could cause a few extremely calibrated GUFs and the poor convergence between the rest of the "Long Form" estimates and their corresponding control totals on auxiliary variables. Thus, it was necessary to drop the control totals with too few "Long Form" records until the extreme skewness of the distributions of the calibrated GUFs was corrected and the overall convergence in the GRE was achieved.

II. *Grouping of control totals* – Some control totals were not appropriate to be included at the 452 DCCAs level because of insufficient number of the corresponding "Long Form" records. As a result, the 452 DCCAs were grouped into 47 broad areas (called "DCCA groups") to include the control totals aforesaid in the GRE. Similarly, some population control totals were not feasible to be included for each age, and these population control totals were included for quinquennial age groups instead.

III. *Inclusion of small area control totals in two runs* – The GRE was performed in two runs to include additional control totals at the housing estates and building groups level with a view to achieving more precise small area estimates. In the first run, the GRE covered the control totals at the DCD, DCCA and DCCA group level, which the number of the corresponding "Long Form" records was large. Then, the number of occupied quarters, households and population of housing estates and building groups were scrutinised and compared with the complete enumeration results, design-based estimates (reciprocals of inclusion probabilities as GUFs) and administrative records (if available). Thereafter, if the derivations were large, the relevant control totals for the housing estates and building groups concerned were added in the second run, together with the control totals already covered in the first run to form one single set of system of nonlinear equations in the GRE. Including control totals in two runs was able to improve the precision of small areas estimates while keeping the number of control totals reasonable.

13. A total of around 10 000 control totals were finally included in the second run of the GRE in the 21C for some 0.3 million "Long Form" records, representing an average of 30 "Long Form" records per control total, as similar to the case of the GRE in the 2016 Canadian Census of Population with 130 000 control totals and 3.5 million "Long Form" records.

Procedure *2: Construction of the quarters dataset*

14. As mentioned in para. 7, a quarters dataset of dimension $n \times p$ containing $\vec{x}_i \ (i = 1, \ldots, n)$ was required for performing the GRE. For instance, to control the number of male residents aged 59 living in a particular DCD, the quarters dataset shall contain the number of male residents aged 59 living in that DCD for each quarters record. The more control totals adopted, the more counts of auxiliary variables had to be created (see Table 2).

Table 2: Demonstration for the data structure of the quarters dataset

Identifier of occupied quarters	x_{1i}	x_{2i}	…	Number of male residents aged 59 living in a particular DCD	…	x_{pi}
1	1	2	…	1	…	x_{p1}
2	0	0	…	2	…	x_{p2}
3	3	5	…	0	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
n	3	4	…	1	…	x_{pn}

15. In the 21C, the GRE was developed using SAS with R package "nleqslv" embedded. The GRE was performed at the DCD level independently and simultaneously. This involved 18 quarters datasets for the 18 DCD GREs, each of which had a dimension of roughly around 15 000 "Long Form" records × 500 counts of auxiliary variables. Compared with performing one single GRE at the territorial level, performing 18 DCD GREs in parallel markedly reduced the system workload and runtime of the whole estimation process, and thus allowed noticeable increase in the number of control totals in the GRE, flexible addition/deletion of control totals to accommodate the local differences among DCDs, and prompt bug fixing and quality assurance work. On the contrary, if the GRE were performed at the territorial level, this would involve a large quarters dataset of dimension of roughly 300 000 "Long Form" records × 10 000 counts of auxiliary variables and increased the runtime dramatically.

Quality assurance of the estimation process of the 21C

16. In the 21C, the GRE successfully reconciled about 10 000 "Long Form" estimates on auxiliary variables with the corresponding control totals. In the quality assurance of the estimation process, the following three sets of indicators were closely monitored to ensure that the GUF adjustments made by the GRE were confined to the acceptable range:

I. the ratios of the calibrated GUFs to the initial GUFs
II. the absolute distances between the calibrated GUFs and the initial GUFs
III. the means, C.V.s, skewnesses and kurtoses of the calibrated GUFs under different combinations of truncation upper and lower limits.

17. For the ratios of the calibrated GUFs to the initial GUFs, the distribution of the ratios was bell-shaped, with the ratios mainly concentrated in the interval of 1.0 - 1.1. Almost all ratios fell within the interval of the truncation lower limit of 0.3 and the truncation upper limit of 3.0. Only very few ratios were out of the interval of 0.3 – 3.0 because of the subsequent GUF rounding process or the introduction of white noise to the GUFs during the statistical data disclosure control with a view to protecting data confidentiality.

18. As regards the absolute distances between the calibrated GUFs and the initial GUFs, most of the absolute distances were equal to or less than 3, suggesting the GUF adjustments were acceptable.

19. A sensitivity analysis was conducted for the GRE. The GRE was performed under different combinations of truncation upper and lower limits, and the changes on the corresponding means, C.V.s, skewnesses and kurtoses of the calibrated GUFs were recorded and closely monitored. It was revealed that the changes on the moments aforesaid and the calibrated GUFs were not notable under the different combinations of truncation upper and lower limits, suggesting that the GRE estimators were stable.

Concluding remarks

20. This paper detailed both the methodological framework and practical application procedures of the GRE to obtain more precise and consistent estimates for the 21C.

21. In recent years, quite a number of countries/economies have been establishing comprehensive administrative registers on population, households and employment. Given their readiness, comprehensiveness and quality, these registers are able to provide benchmark information to supplement control totals for censuses or even are potential alternatives to complete enumeration in future censuses. The GRE allows the inclusion of a great number of control totals of different kinds in the estimation process. With the increasing availability of administrative records, the application of the GRE will be more and more important.

References

Census and Statistics Department. (2002). *"Hong Kong 2001 Population Census – Main Report –Volume II"*.

Census and Statistics Department. (2007). *"The Use of the Calibration Estimation Method in the General Household Survey"*, Research Bulletin Issue No .4.

Census and Statistics Department. (2012). *"Hong Kong 2011 Population Census – Main Report : Volume II"*.

Census and Statistics Department. (2017). *"Hong Kong 2016 Population By-census – Technical Report"*.

Deville, J.C., Särndal, C.E. (1992). *"Calibration Estimators in Survey Sampling", Journal of the American Statistical Association*.

Hasselman, B. (2018). *"Package 'nleqslv'"*.

Lam, J., Cheng, J. (2014). *"Imputation and Estimation Methods for the 2011 Population Census"*, Research Bulletin Issue No .22.

Lohr, S. (2021). *"Sampling Design and Analysis"*, Chapman and Hall/CRC.

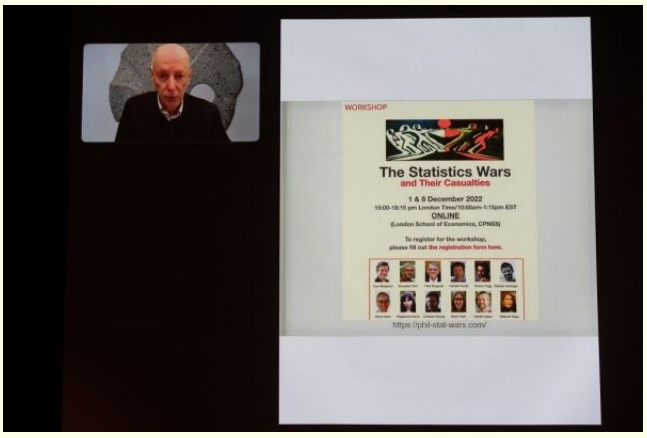
Statistics Canada. (2018). *"Sampling and Weighting Technical Report, Census of Population, 2016"*.

Seminar in Commemoration of 55th Anniversary of the Census and Statistics Department and 45th Anniversary of the Hong Kong Statistical Society

On November 11, 2022, a seminar was held at the Chiang Chen Studio Theatre in the Hong Kong Polytechnic University to commemorate the 55th anniversary of the Census and Statistics Department (C&SD) and the 45th anniversary of the Hong Kong Statistical Society. The seminar aimed to provide valuable insights into various statistical topics and issues related to census and survey data.

Mr Leo YU, Commissioner for Census and Statistics (left) and Professor Alan WAN, President (right) delivered opening speech when the seminar began.

The event was graced by distinguished speakers who shared their views on different aspects of statistics. The first speaker was Professor Ian McKEAGUE, who is the Head and Chair Professor of the Department of Biostatistics at City University of Hong Kong. He delivered a presentation on the topic of "Statistical Uncertainty and the Humor of Groucho Marx," highlighting the importance of understanding the sources of statistical uncertainty and the ways in which humor can be used to communicate statistical concepts.



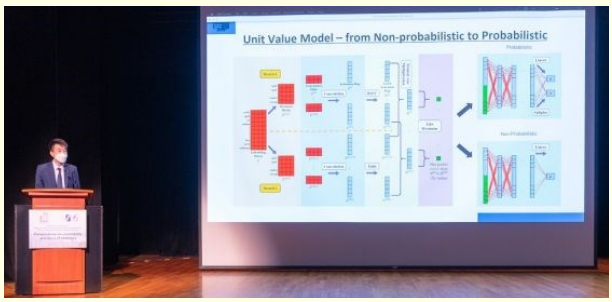
Professor McKEAGUE

The second speaker was Professor Tarani CHANDOLA, who is the Professor of Medical Sociology and Director of the Methods Hubs at the Faculty of Social Sciences of The University of Hong Kong. Professor CHANDOLA discussed "Non-response and Missing Data Methods in Census and Survey Data," emphasizing the need for appropriate methods to handle missing or incomplete data in survey research and the implications of such methods for statistical analysis.

Professor CHANDOLA

Miss Natalie CHUNG

The third session was a panel discussion featuring Miss Natalie CHUNG, Mr Ian NG, and Mr Benjamin CHAN, who are Statisticians and Research Manager at the Census and Statistics Department. They discussed "Anomaly Detection in Merchandise Trade Data: From Rule-based to Deep Learning Approach," presenting a range of approaches to detecting anomalies in merchandise trade data and highlighting the potential of deep learning methods to improve anomaly detection accuracy.



Mr Ian NG

Mr Benjamin CHAN

The seminar provided a valuable opportunity for professionals in the field of statistics to share their expertise and insights on various topics related to census and survey data. The event was well-attended and received positive feedback from the participants. The Census and Statistics Department and the Hong Kong Statistical Society are commended for organizing such an informative and engaging seminar that contributed to the advancement of knowledge in the field of statistics.

Mr Tim CHAU, Deputy Commissioner for Census and Statistics, Prof Alan WAN and speakers

2021/22 Statistical Project Competition for Secondary School Students

Organising Committee* of the 2021/22 Statistical Project Competition

The 2021/22 Statistical Project Competition (SPC) for Secondary School Students, the 36th 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 on one of the following themes: population, education or environment and health, 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 Thematic Project and Prize for the Best Graphical Presentation of Statistics.

To help interested participants prepare for the Competition, an online Briefing Seminar for SPC for 2021/22 was held on 23 October 2021. Representatives from the Census and Statistics Department and the Datality Lab Limited, sponsor of 2021/22 SPC, had introduced the official statistics and the application of data science 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 during 29 October 2021 to 12 November 2021.

Encouraging number of entries

Affected by COVID-19, the face-to-face lessons of secondary schools had been partially suspended in 2022, posting great difficulties for school teachers and participating students in discussing and preparing of their statistical projects. Despite the situation, 125 entries and 83 entries were submitted for the Junior Section and the Senior Section respectively from 63 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, Professor CHEUNG Ka-chun of The University of Hong Kong, and comprised some 46 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 Michael WONG Kwok-yee of the Hong Kong University of Science and Technology for serving as the Chairpersons of the interview panel for Senior and Junior Section respectively. To reduce the risk of infection, this round of panel interviews was conducted online through Zoom.

Prize Presentation Ceremony

The Prize Presentation Ceremony for the 2021/22 SPC took place on 22 October 2022 at the Lecture Theatre of the Kowloon Tong Education Services Centre. Professor Alan WAN Tze-kin, President of HKSS, Mr Leo YU Chun-keung, Commissioner for Census and Statistics, Miss Yvonne LAM Si-hang, Principal Education Officer (Curriculum Development) of the Education Bureau, Mr. Roland LEUNG, Managing Director of Datality Lab 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 Diocesan Girls' School, who used official statistics to study the poverty rate of ethnic minorities in Hong Kong, won the First Prize of the Junior Section. Students of Stewards Pooi Kei College won the Second Prize, while students of Pui Ching Middle School won the Third Prize. The Prize for the Best Thematic Project and the Prize for the Best Graphical Presentation of Statistics were won by the First and the Third teams respectively.

Professor WAN presented the First Prize for the Junior Section to students of Diocesan Girls' School

As for the Senior Section, the statistical report from students of Stewards Pooi Kei College was appraised as the best among all the projects. They applied official statistics to analyse the phenomenon of "silver tsunami" and its impacts on Hong Kong's healthcare system. Students of Holy Family Canossian College won the Second Prize, while students of Diocesan Girls' School won the Third Prize. Students of another team of Diocesan Girls' School won the Prize for the Best Thematic Project. Meanwhile, the Prize for the Best Graphical Presentation of Statistics was won by the First team.

Mr. Roland LEUNG presented the First Prize for the Senior Section to students of Stewards Pooi Kei College

Mr Leo YU, Miss Yvonne LAM and Professor CHEUNG Ka-chun presented prizes to winning teams

Gratitude

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

*Organising Committee for the 2021/22 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 Hinz SHUM	Census and Statistics Department
Mr Michael CHU	Census and Statistics Department
Mr Stanley TSANG	Census and Statistics Department

LATEST NEWS

News Section

◆ Personnel Changes (New Appointments, Promotions and Retirements)

The Chinese University of Hong Kong (CUHK)

※ Prof WANG Junhui has joined the Department of Statistics of CUHK as Professor with effect from August 2022.

※ Dr LEUNG Sze-him Isaac and Dr LIU Kin-yat have joined the Department of Statistics of CUHK as Lecturer with effect from September and August 2022 respectively.

※ Prof YAU Chun-yip and Prof FANG Xiao of the Department of Statistics of CUHK have been promoted to Professor and Associate Professor respectively .

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

※ Dr DOU Baojun has joined the Department of Management Sciences of CityU as Assistant Professor with effect from November 2022.

※ Dr LI Hanwei has joined the Department of Management Sciences of CityU as Assistant Professor with effect from July 2022.

※ Prof SHOU Biying of Department of Management Sciences of CityU has been promoted to Professor with effect from July 2022.

※ Dr Sun Zhankun of Department of Management Sciences of CityU has been promoted to Associate Professor with effect from July 2022.

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

※ Dr ZHOU Le has joined the Department of Mathematics of HKBU as Assistant Professor with effect from May 2022.

LATEST NEWS

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 BAI Shurui Tiffany has joined MIT of EdUHK as Assistant Professor with effect from July 2022.

※ Dr CHAN Tse-tin David has joined MIT of EdUHK as Assistant Professor with effect from August 2022.

※ Dr LI Xin Stephen has joined MIT of EdUHK as Research Assistant Professor with effect from August 2022.

※ Dr CHEUNG Ho-yin Haoran has joined MIT of EdUHK as Lecturer I with effect from August 2022.

※ Dr SINGH Manpreet has joined MIT of EdUHK as Lecturer I with effect from August 2022.

※ Professor KONG Siu-cheung has been appointed as Research Chair Professor of E-Learning and Digital Competency of MIT of EdUHK.

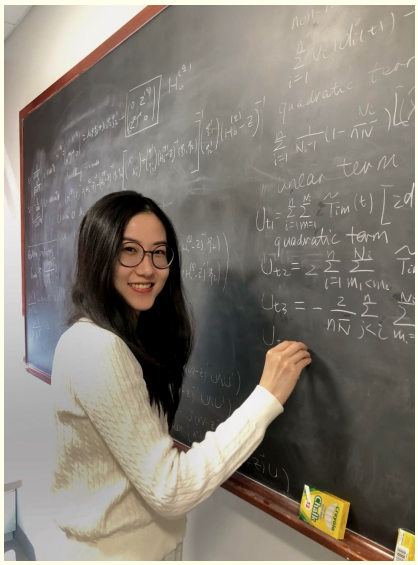
Department of Applied Mathematics of the Hong Kong Polytechnic University (PolyU)

※ Professor HUANG Jian has joined Department of Applied Mathematics of PolyU as Chair Professor under the University Strategic Hiring Scheme with effect from June 2022.

※ Dr HAN Ruijian has joined Department of Applied Mathematics of PolyU as Assistant Professor with effect from September 2022.

LATEST NEWS

Result of the first HKSS-John Aitchison Prize 2023



We are pleased to present the report of the Adjudication Panel of the HKSS-John Aitchison Prize in Statistics 2023. By the deadline of October 31, 2022, high quality submissions were received for the prize. The panel was faced with the challenging task of selecting a winner from such a high-calibre pool with exceptional quality.

After careful deliberations, the panel has decided to award the HKSS-John Aitchison Prize in Statistics 2023 to Dr. Jingming WANG (2021 PhD Graduate at the Hong Kong University of Science and Technology). Dr. WANG is a co-author of the paper titled "Statistical inference for principal components of spiked covariance matrices", which was jointly written with Z.G. BAO, X.C. DING, and K. WANG and published in Annals of Statistics, 50(2): 1144-1169 (2022).

The theory of spiked covariance matrix has become an important area of research in statistics in recent years, and has found applications in fields such as genetics and finance. Dr. WANG's work has filled a longstanding gap in the literature by obtaining, under very general conditions, the joint limiting distribution of outlying eigenvalues and their associated eigenvectors in the super-critical region. This provides fundamental understanding of high dimensional principal component analysis.

We believe that Dr. WANG's outstanding contributions to the field of statistics make her a deserving recipient of this prestigious award. HKSS-John Aitchison Prize in Statistics is awarded once a year, with its value currently being HK$12,000. For more details for the Prize, please refer to its website (https://www.hkss.org.hk/index.php/events/john-aitchison-prize-in-statistics).