香港統計學會 HONG KONG STATISTICAL SOCIETY

HKSS Bulletin - Volume 48 No.1
April 2026

EDITOR'S FOREWORD

Dear Members,

Welcome to the 2026 issue of HKSS Bulletin. It is with great pleasure that I present to you this issue of the HKSS Bulletin. First, I would like to thank Dr. Marjorie CHIU who served as the Editor last year, as well as her invaluable guidance to me during the transition period.

There are three featured articles covered in this issue. From the academic field, Dr. LAM Ming-long shares his insights on Shapley Values for Linear Regression Model, offering valuable methodological perspectives for statistical analysis. From the Census and Statistics Department, Alex YIU, Timothy CHU, and Sam TSE elaborate on the re-engineering of the 2026 Population Census, introducing a pioneering approach that balances cost-effectiveness and data quality. Additionally, the Organising Committee provides a summary of our annual event, the 2024/25 Statistical Project Competition for Secondary School Students, showcasing the achievements of young statisticians and the vitality of statistical education in Hong Kong.

Finally, I would like to express my heartfelt gratitude to all the contributors for their valuable submissions, as well as the editorial board for their efforts. Their commitment has made this issue possible.

With warm regards,

Wai-tung HO

			Email
Editor	:	Dr. HO, Wai-tung, PolyU CPCE	wai.tung.ho@cpce-polyu.edu.hk
Secretary	:	Ms. YEUNG, Angela Yuen-ting, C&SD	aytyeung@censtatd.gov.hk
Member	:	Mr. LAU, Michael Tsz-ho, C&SD	-

CONTENTS

(Vol. 48/No.1, April 2026)

PRESIDENT'S FORUM

Professor CHAN Ngai-hang

As we embrace the opportunities of 2026, I extend my warmest wishes to all members of the Hong Kong Statistical Society (HKSS) for a rewarding year ahead.

For nearly five decades, HKSS has served as a cornerstone in promoting statistical excellence and collaboration. Throughout this journey, our Society has remained steadfast in its mission to offer a platform where researchers, educators, students, government officials, and industrial practitioners converge to exchange ideas, stimulate methodological advancements, and advocate for sound statistical practices across diverse sectors. Collectively, our members strive to harness the power of data to generate meaningful insights that contribute to the betterment of our community.

Reflecting on the past year, our engagement with regional partners has been particularly fruitful. A delegation of HKSS members actively participated in the Guangdong-Hong Kong-Macao Greater Bay Area Statistical Forum held in September 2025, further strengthening cross-boundary cooperation and knowledge sharing. Such collaborative platforms remain essential to our shared growth, and HKSS looks forward to deepening these ties with statistical organisations near and far.

In October 2025 the Society commemorated the World Statistics Day with a public lecture on "Quality Statistics and Data for Everyone," underscoring our

commitment to enhancing statistical literacy across the wider community. Additionally, academic exchanges were enriched through events such as the CityUHK Biostatistics Workshops in June, which provided valuable opportunities for scholarly discourse.

I am also pleased to announce the successful conclusion of the selection process for the HKSS–John Aitchison Prize in Statistics 2025/26. The quality of this year's nominees was extremely outstanding, and I wish to thank the selection committee for their difficult and meticulous deliberations. I offer my sincere congratulations to the two prize recipients, whose valuable contributions to statistics have advanced the field and reflected strong scholarly excellence. We look forward to celebrating their achievements with all members of the Society.

Looking ahead, HKSS will continue to champion innovation and professional development in response to emerging challenges and trends. The seminar entitled "Integration of AI in Statistical Work: Challenges and Opportunities," given by Professor Xuming HE, the former President of ISI held earlier this year at CityUHK, exemplified our dedication to staying at the forefront of technological advancements. I encourage all members to remain engaged with forthcoming seminars, networking gatherings, and other related activities designed to support our vibrant membership. My sincere thanks go to each of you for your continued dedication and enthusiasm, which remain the bedrock of our Society's enduring mission.

CALCULATING SHAPLEY VALUES FOR LINEAR REGRESSION MODEL

Dr. LAM Ming-Long

Data Science Institute, The University of Chicago

Introduction

Fairness is a multifaceted concept encompassing equity, equal opportunity, impartiality, and non-discrimination. It is a virtue deeply rooted in human behavior and societal norms. Fairness does not mean that everyone gets what they want, but that each person receives what they deserve. In a utopian society, the share a person merits should be proportional to their contributions to the community's well-being. It seems to us that we have settled the question of fairness when, in fact, we just opened Pandora's Box. It leads to more questions than answers. For instance, how should we define community well-being? Can we measure a person's contribution to a community? Do we determine a person's contribution based on the impacts on the receivers or the motivations of the givers?

In addition to being a social issue, data scientists face another aspect of fairness in model assessment. They may be required by law to attest that their models do not deliberately or disparately discriminate against people in protected classes: race, color, religion, sex (including transgender status, sexual orientation, and pregnancy), national origin, age (40 or older), disability, or genetic information. For data scientists whose work is used for personnel management (e.g., hiring, promotion, termination, and compensation), retail finance (e.g., loan applications and insurance premiums), law enforcement (e.g., crime prevention and suspect profiling), and college admissions (e.g., ranking of applicants), their work must first be scrutinized by legal teams. Although most seasoned data scientists avoid predictors that are clearly associated with protected classes, the

resulting models may still produce outcomes that could legally jeopardize them. To prevent their models from being sidelined or discarded, data scientists must know how to modify them. If the predictors that place us in a predicament are less critical to the model outcomes, we may drop them from the model without hesitation. Otherwise, we must identify alternative predictors or, reluctantly, abandon the model to circumvent the legal concern.

A fundamental task in ensuring fairness is to assess how each party's contribution affects the outcome. Suppose the result is an extra-large round pizza. Everyone at the table argues that they deserve their fair slice. Should the person who brings the most ingredients or spends the most time making the pizza receive the largest slice? Does this imply that the person who gets no pizza has contributed nothing? What about the cook who just stepped away from the table?

Fair Share of Uber Fare

Let us consider a practical example to get some ideas. The Gleacher Center is the University of Chicago's downtown campus in Chicago. It is located next to the Chicago River and offers convenient access to the Magnificent Mile, an upscale shopping district in Chicago. After their afternoon classes at the Gleacher Center, Amy, Betty, and Cathy share an Uber ride to their destinations. They agree to pay their fair share of the fare. Amy alights at Water Tower Place to watch a theatrical performance with her boyfriend. Betty gets off at Lincoln Park Zoo to attend her family's dinner. Finally, Cathy ends the Uber ride at Wrigley Field to join her coworkers to watch a baseball game. The final fare, including tips, is $25.90 in United States dollars. Cathy pays the full fare with her credit card. So, how much should Cathy receive from Amy and Betty? It is not an equal share of the fare, as each woman travels a different distance. Water Tower Place is 0.8

miles north of the Gleacher Center, Lincoln Park Zoo is 3.9 miles away, and Wrigley Field is 6.7 miles away. Should their shares of the fare be proportional to the distances they traveled? Not exactly! All three ladies are in the car for the initial 0.8 miles. After Amy departs, Betty and Cathy remain in the vehicle for the following 3.1 miles. Finally, Cathy rides alone for the last 2.8 miles. Without additional information, we cannot help these three ladies calculate their fair shares.

What information do we need? It is the Uber fares from the Gleacher Center to each destination. These amounts allow us to calculate each lady's share for each leg of the trip. We already know that the Uber trip from the Gleacher Center to Wrigley Field costs $25.90. Therefore, we need only the fares from the Gleacher Center to Lincoln Park Zoo and Wrigley Field. Here is the additional information. The fare from the Gleacher Center to Water Tower Place is $8.50; to Lincoln Park Zoo, $18.34. Amy, Betty, and Cathy will equally share the $8.50 fare for the first leg of the trip. Each lady will pay one-third of $8.50, or $2.83. Amy's share will not change after she leaves the party. The second leg of the trip is from Water Tower Place to Lincoln Park Zoo. It costs $18.34, minus $8.50, leaving $9.84. Betty and Cathy will split this amount equally. Including her share of the first leg of the trip, Betty's share is $2.83 plus half of $9.84, totaling $7.75. Finally, Cathy will be solely responsible for the final leg of the trip, which will cost her $25.90, minus $18.34, leaving $7.56. In summary, Cathy should pay $2.83 for the first leg, $4.92 for the second leg, and $7.56 for the final leg. In total, Cathy's share is $15.31. To verify that our calculation is correct, we add up the ladies' shares. Amy's share is $2.83, Betty's share is $7.75, and Cathy's share is $15.31. The sum is $25.89, which is one penny short of the actual fare of $25.90. In practice, our calculation is correct

within the rounding errors. Amy may voluntarily contribute $0.01, as she pays the least.

The above manual calculations are valid because the destinations are approximately on a straight line on a Chicago map, starting at the Gleacher Center and ending at Wrigley Field. Are the calculations still valid if the destinations are not on a straight line? Is there a more systematic way to handle the calculations? The short answer is to treat the problem as a coalition game. The longer answer is to calculate the Shapley Values.

Coalition Game

A team sport, such as baseball, basketball, football, or volleyball, is an athletic activity in which players organize into teams. Instead of competing as individuals, teams compete against other teams. Their common goal is to score as many points as possible while avoiding penalties and injuries. Teams enter tournaments (e.g., the World Cup, the World Series, or the Super Bowl) to compete against other teams. In addition to trophies or medals, the winning team may also receive monetary awards. How does the team's management distribute the financial awards to each player? How many bragging rights can players claim? Are reserve players included, too?

Although a coalition game (also known as a cooperative game) is a mathematical model studied in Game Theory, it operates like a team sport. In a coalition game, the players are the features. Subsets of these features form coalitions. An empty coalition has no features. The Grand Coalition is the superset that contains all features. When a model is trained on a subset of features along

with default terms (e.g., Intercept and Offset), it will deliver a specific level of performance, as measured by assessment metrics. Higher metric values indicate better model performance. Conventionally, a model without any features should provide a metric value of zero, and a model with more features should yield a higher metric value than one with fewer features. After we have settled on a particular model, we will have questions similar to those of a sports team's management. How do we define the contribution of each feature to model performance? What portion of model performance is attributed to each feature? What about features that are still under consideration but not yet included in the model?

Shapley Value Algorithm

Shapley Values provide an absolute measure of the importance of each feature considered in a model. Shapley Values account for all possible sequences in which the features can enter the model. Each sequence generates one realization of importance, and Shapley values are the average of all realizations. Even when a feature does not enter the model (due to linear dependence), its Shapley value can still indicate its importance if the model includes it.

The Shapley Value is named after Dr. Lloyd Stowell Shapley (1923 – 2016) in recognition of his 1953 paper on the value of an N-person game. In addition to this contribution, Dr. Shapley and Dr. Alvin Eliot Roth (b.1951) were jointly awarded the 2012 Nobel Prize in Economic Sciences for their work on the theory of stable allocations and the practice of market design. Lipovetsky and Conklin (2001) first applied the Shapley Value, a tool from cooperative game theory, as a measure of feature importance in a linear regression model. Conklin, Powaga, and Lipovetsky

(2004) further demonstrated the advantage of the Shapley Value in identifying key drivers of customer satisfaction.

Suppose $x_1, \dots, x_m$ are features being considered for a model. We represent the Grand Coalition by $T = \{x_1, \dots, x_m\}$ which is the universal set of features. Let S denotes a general coalition with s features. When we train a model with features in S, it will yield an assessment metric, denoted by $v(S)$. The Shapley Value of the i^{th} feature will be

$$\phi_i(v) = \frac{1}{m} \sum_{S \subseteq (T/\{x_i\})} \big(v(S \cup \{x_i\}) - v(S)\big) \Big/ \binom{m-1}{s}$$

The notation $S \subseteq (T/\{x_i\})$ implies that S is a subset of T, but the feature x_i cannot be contained in S.

To make the formula easier to interpret, we will enter the features into the model one at a time. Since there are m features, the feature x_i can enter the model at m different steps. At each step, we will compare two sub-models. Except for the feature x_i, both sub-models have the same features contained in the set S. Since the model with the feature x_i and all the features in S will yield a higher assessment metric than the model with just the features in S, the difference of $v(S \cup \{x_i\}) - v(S)$ is a non-negative value. Heuristically, the difference is the marginal contribution of the feature x_i to the model performance, provided the features in S are already in the model. The combinatorial value $\binom{m-1}{s}$ represents the number of ways we can choose s features from $m-1$ features in $T/\{x_i\}$. Finally, the Shapley Value is the average

of all marginal contributions of the feature x_i over all possible ways to choose S and all possible steps for the feature x_i can enter the model.

Shapley Values for Shared Uber Ride

Let us calculate the Shapley Values for the example in which Amy, Betty, and Cathy share an Uber ride. The Grand Coalition consists of the three ladies. Although we do not know how Uber calculates the fares, let us assume the fares are the assessment metric for Uber's model. Below are the trip fares by number of riders.

1. Zero Rider: $\nu(\varnothing)$ = \$0 where $\varnothing$ represents the empty set
2. Single Rider: $\nu(\{\text{Amy}\})$ = \$8.50, $\nu(\{\text{Betty}\})$ = \$18.34, and $\nu(\{\text{Cathy}\})$ = \$25.90
3. Two Riders: $\nu(\{\text{Amy, Betty}\})$ = \$18.34, $\nu(\{\text{Amy, Cathy}\})$ = \$25.90, and $\nu(\{\text{Betty, Cathy}\})$ = \$25.90
4. Three Riders: $\nu(\{\text{Amy, Betty, Cathy}\})$ = \$25.90

Let me imagine each lady getting into the Uber one by one. How many ways can the three ladies enter the vehicle? The answer is six, which equals the factorial of three. Let us look at one of these six ways. Amy enters the car first, followed by Betty, and finally Cathy. When the Uber vehicle arrives, but nobody enters, the trip fare is $\nu(\varnothing)$ = \$0. Once Amy enters the car, the trip fare jumps to $\nu(\{\text{Amy}\})$ = \$8.50. Betty's entrance increases the trip fare to $\nu(\{\text{Amy, Betty}\})$ = \$18.34. The final trip fare is $\nu(\{\text{Amy, Betty, Cathy}\})$ = \$25.90 after Cathy enters the vehicle. The table below displays the current trip fare when a specific rider enters the car.

Grand Coalition Formation Sequence	Current Trip Fare		
	First Rider Enters	Second Rider Enters	Third Rider Enters
{Amy, Betty, Cathy}	$8.50	$18.34	$25.90
{Amy, Cathy, Betty}	$8.50	$25.90	$25.90
{Betty, Amy, Cathy}	$18.34	$18.34	$25.90
{Betty, Cathy, Amy}	$18.34	$25.90	$25.90
{Cathy, Amy, Betty}	$25.90	$25.90	$25.90
{Cathy, Betty, Amy}	$25.90	$25.90	$25.90

Next, we calculate each lady's contribution (i.e., the increase in trip fare) when she enters the vehicle. Let us use the sequence {Betty, Amy, Cathy} to step through the calculation. When Betty enters the vehicle, the trip fare jumps from $0 to $18.34. We determine that Betty's contribution to this sequence is $18.34. After Amy enters the vehicle, the trip fare remains $18.34 because Betty will ride a longer distance than Amy. Therefore, Amy's contribution is $0. After Cathy enters the vehicle, the trip fare increases to $25.90. This indicates Cathy's contribution is $25.90 minus $18.34, leaving $7.56. The following table shows the calculations for all six sequences. The integer in parentheses indicates the lady's position in the sequence. The contribution is the difference, calculated by subtracting the preceding position's trip fare from the lady's current position's trip fare.

Grand Coalition Formation Sequence	Current Trip Fare			Individual Contribution		
	First Rider Enters	Second Rider Enters	Third Rider Enters	Amy	Betty	Cathy
{Amy, Betty, Cathy}	$8.50	$18.34	$25.90	(1) $8.50	(2) $18.34 – $8.50 = $9.84	(3) $25.90 – $18.34 = $7.56
{Amy, Cathy, Betty}	$8.50	$25.90	$25.90	(1) $8.50	(3) $25.90 – $25.90 = $0	(2) $25.90 – $8.50 = $17.40
{Betty, Amy, Cathy}	$18.34	$18.34	$25.90	(2) $18.34 – $18.34 = $0	(1) $18.34	(3) $25.90 – $18.34 = $7.56
{Betty, Cathy, Amy}	$18.34	$25.90	$25.90	(3) $25.90 – $25.90 = $0	(1) $18.34	(2) $25.90 – $18.34 = $7.56
{Cathy, Amy, Betty}	$25.90	$25.90	$25.90	(2) $25.90 – $25.90 = $0	(3) $25.90 – $25.90 = $0	(1) $25.90
{Cathy, Betty, Amy}	$25.90	$25.90	$25.90	(3) $25.90 – $25.90 = $0	(2) $25.90 – $25.90 = $0	(1) $25.90

Finally, we calculate the average of the six individual contributions for each lady. For Amy, her average is ($8.50 + $8.50 + $0 + $0 + $0 + $0) / 6 = $17.00 / 6 = $2.83. For Betty, her average is ($9.84 + $0 + $18.34 + $18.34 + $0 + $0) / 6 = $46.52 / 6 = $7.75. For Cathy, her average is ($7.56 + $17.40 + $7.56 + $7.56 + $25.90 + $25.90) / 6 = $91.88 / 6 = $15.31. The sum of the three averages is $2.83 + $7.75 + $15.31 = $25.89. These three averages are the Shapley Values for each lady. The following table displays each lady's Shapley Value. To easily understand the relative magnitudes of their contributions, we routinely calculate the percentage of the Shapley Values relative to the total fare, which is $25.90.

Rider	Shapley Value	Percentage of Total
Amy	$2.83	10.94%
Betty	$7.75	29.94%
Cathy	$15.31	59.12%
Total	$25.90	100.00%

We can conclude that Cathy's contribution is almost twice that of Betty, and 5.4 times that of Amy. Betty's contribution is 2.7 times that of Amy.

Linear Regression with Aliased Features

An aliased feature is linearly dependent on other features. In other words, the values of an aliased feature can be expressed as a linear combination of the values of different features. In training a linear regression model, aliased features are often excluded to avoid the unwanted singularities during the estimation. Even though there are no regression coefficients for the aliased features, do they still have their Shapley Values? The answer is Yes. We will explain using the following example.

ID	factor1	cont1	cont2	y
1	Negative	-3	3.3	-1.65
2	Negative	-2	2.2	0.03
3	Negative	-1	1.1	0.77
4	Zero	0	0	1.84
5	Positive	1	-1.1	1.87
6	Positive	2	-2.2	3.65
7	Positive	3	-3.3	3.99

The variable factor1 is a categorical feature with three categories. The variables cont1 and cont2 are continuous features. Finally, the variable y is our dependent variable. As with all regression models, we will use the Coefficient of Determination as the assessment metric. At this point, acute observers may already have noticed that cont2 equals cont1 multiplied by -1.1. In short, cont1 and cont2 are linearly dependent or aliased. Depending on the order in which these two variables enter the model, the latter variable will be found to be redundant by most, if not all, regression applications. By default, that variable will have a regression coefficient of zero.

Having this piece of information in mind, we will proceed as usual. The first step is to train all eight possible linear regression models that can be generated from these three features. The character "1" represents the Intercept term, which is always in our regression models.

Model	Coefficient of Determination
1	0.0
1 + factor1	0.760745812123
1 + cont1	0.961873446543
1 + cont2	0.961873446543
1 + factor1 + cont1	0.978246677416
1 + factor1 + cont2	0.978246677416
1 + cont1 + cont2	0.961873446543
1 + factor1 + cont1 + cont2	0.978246677416

The second step is to express each feature's contribution during the formation of a Grand Coalition sequence in terms of the Coefficients of Determination. To avoid overcrowding the table, we list only the model specifications from which the Coefficients of Determination are generated.

Grand Coalition Formation Sequence	Contribution in terms of the Coefficient of Determination		
	factor1	cont1	cont2
{factor1, cont1, cont2}	{1 + factor1}	{1 + factor1 + cont1} – {1 + factor1}	{1 + factor1 + cont1 + cont2} – {1 + factor1 + cont1}
{factor1, cont2, cont1}	{1 + factor1}	{1 + factor1 + cont2 + cont1} – {1 + factor1 + cont2}	{1 + factor1 + cont2} – {1 + factor1}
{cont1, factor1, cont2}	{1 + cont1 + factor1} – {1 + cont1}	{1 + cont1}	{1 + cont1 + factor1 + cont2} – {1 + cont1 + factor1}
{cont1, cont2, factor1}	{1 + cont1 + cont2 + factor1} – {1 + cont1 + cont2}	{1 + cont1}	{1 + cont1 + cont2} – {1 + cont1}

Grand Coalition Formation Sequence	Contribution in terms of the Coefficient of Determination		
	factor1	cont1	cont2
{cont2, factor1, cont1}	{1 + cont2 + factor1} – {1 + cont2}	{1 + cont2 + factor1 + cont1} – {1 + cont2 + factor1}	{1 + cont2}
{cont2, cont1, factor1}	{1 + cont2 + cont1 + factor1} – {1 + cont2 + cont1}	{1 + cont2 + cont1} – {1 + cont2}	{1 + cont2}

By plugging in the correct Coefficient of Determination, we produce the following table.

Grand Coalition Formation Sequence	Contribution in terms of the Coefficient of Determination		
	factor1	cont1	cont2
{factor1, cont1, cont2}	0.760745812123	0.217500865293	0.000000000000
{factor1, cont2, cont1}	0.760745812123	0.000000000000	0.217500865293
{cont1, factor1, cont2}	0.016373230873	0.961873446543	0.000000000000
{cont1, cont2, factor1}	0.016373230873	0.961873446543	0.000000000000
{cont2, factor1, cont1}	0.016373230873	0.000000000000	0.961873446543
{cont2, cont1, factor1}	0.016373230873	0.000000000000	0.961873446543

Finally, we calculate the Shapley Values for the features by averaging their contributions across the six formation sequences. The following table shows the features' Shapley Values and their percentages relative to the sum of their Shapley Values.

Feature	Shapley Value	Percentage of Total
factor1	0.264497424623	27.04%
cont1	0.356874626397	36.48%
cont2	0.356874626397	36.48%
Total	0.978246677416	100.00%

As you may have already found, the sum of the Shapley Values is equal to the Coefficient of Determination of the full model: {1 + factor1 + cont1 + cont2}. This

is not an accidental result. In fact, the sum of the Shapley Values is always equal to the model's assessment metrics when all features are in the Grand Coalition.

This example shows that although aliased features may not be included in a regression model, their Shapley Values are not necessarily zero. In this case, cont1 and cont2 have identical Shapley Values. This implies they are equally effective in predicting the dependent variable, and either one suffices.

Essential Features in Predicting House Sale Price

When the number of features exceeds three, calculations become more tedious, time-consuming, and error-prone. Using software libraries to obtain results is the way to go. Liang (2021) developed the ShapleyValue library for R. Please visit https://cran.r-project.org/web/packages/ShapleyValue/index.html for documentation on this R library. For Python users, there is none, technically speaking. The SHAP library does not calculate Shapley Values but instead produces Shapley Additive Explanations values. The SHAP values approximate Shapley Values adequately when the number of features is exceedingly large, making Shapley Values infeasible to compute. The author has written Python code that calculates Shapley Values from first principles. Please contact the author by email to obtain the code.

Our final example examines the features that predict the sale prices of 2,496 single-family homes in Hyde Park Township in 2021. A township is a local administrative division of a county. The Hyde Park township is part of Cook County, the State of Illinois. Except for the purpose of assessing county property tax, the township does not have its own government because it is entirely within

the legal city limit of Chicago. The data is part of the larger Residential Modeling Characteristics (Chicago) dataset, which can be retrieved from the Cook County Data Portal at https://datacatalog.cookcountyil.gov/Property-Taxation/Assessor-Archived-05-11-2022-Residential-Modeling-/8f9d-wy2d/about_data. To avoid discriminating against homeowners, we use only the following features related to house attributes and neighborhood characteristics.

Categorical Features

1. Attic Type: *Full*, *Partial*, or *None*.
2. Basement: *Full*, *Partial*, or *None*.
3. Central Air Conditioning: *No* or *Yes*.
4. Central Heating: *Electric Heater*, *Hot Water Steam*, or *Warm Air Furnace*.
5. Residence Type: *Individual Townhome*, *One-story House*, *Split-level House*, or *Two-or-more story House*.
6. Road Proximity: *< 100 Feet*, *101 – 300 Feet*, or *> 300 Feet* from a heavy traffic road
7. Roof Material: *Shake*, *Shingle + Asphalt*, *Slate*, *Tar + Gravel*, *Tile*, or *Other*

Continuous Features

1. Number of Bedrooms
2. Building Square Feet, in thousands of square feet (1,000 square feet = 92.9 square meters)
3. Number of Fireplaces
4. Number of Full Baths
5. Garage Size, in number of cars
6. Number of Half Baths
7. Land Square Feet, in thousands of square feet

8. Tract Median Income, in ten-thousands of dollars (USD 10,000 is approximately HKD 77,980)

Dependent Variable

1. Natural Logarithm of Sale Price in United States dollars

After training all 32,768 (i.e., 2^{15}) possible linear regression models in Python 3.13 for 1,312.50 seconds on a commodity laptop (1.7 GHz Intel CPU, 12 GB RAM, and Microsoft Windows 11 Home), we obtained the Shapley Values for 15 predictors. The table below displays the fifteen Shapley Values in descending order.

Feature	Shapley Value	Percentage
Building Square Feet	0.1177	24.18%
Full Baths	0.0928	19.08%
Tract Median Income	0.0826	16.97%
Central Air Conditioning	0.0583	11.98%
Residence Type	0.0440	9.04%
Half Baths	0.0239	4.92%
Fireplaces	0.0212	4.35%
Roof Material	0.0162	3.33%
Garage Size	0.0130	2.67%
Attic Type	0.0064	1.32%
Land Square Feet	0.0030	0.63%
Central Heating	0.0027	0.55%
Road Proximity	0.0026	0.53%
Basement	0.0022	0.45%
Total	0.4865	100.00%

At first glance, the Coefficient of Determination is only 0.4865, which is mediocre given that the model contains fifteen predictors. We would argue that the Coefficient of Determination is already quite decent for practical purposes. These fifteen predictors reflect a home buyer's logical and quantitative reasoning. Attributes such as floorplan layout, appliance quality, backyard availability, and the academic standing of local schools, which are not recorded in the data, may account for the model's lack of fit.

If we can accept this model's Coefficient of Determination, we can proceed to examine the Shapley Values. Like home buyers anywhere, their top criterion is the size of the living space. It accounts for almost a quarter of the model's Coefficient of Determination. The number of full bathrooms comes in second, accounting for 19% of the model's Coefficient of Determination. A full bathroom is not only a place to meet our sanitary needs but also a personal sanctuary for overwhelmed adults seeking temporary refuge. Every adult in the household wants their personal sanctuary, don't they? It is no surprise that the number of full baths ranks second in terms of importance to the model. Tract Median Income ranks third, indirectly indicating the neighborhood's quality and affordability. The last predictor that accounts for at least 10% of the model's Coefficient of Determination is whether central air conditioning is available. Although Chicago is known for its harsh winter, the summer heat can be brutal, too! Therefore, having central air conditioning is a sign of a modern or recently renovated residence.

Conclusion

We cannot use Shapley Values directly to assess a model's fairness. However, Shapley Values provide a quantitative measure of a predictor's contribution to a model's goodness-of-fit. Although we still need to determine whether a predictor will be a legal liability when deploying the model, its Shapley Value can indicate the adverse impact on the model's goodness-of-fit should we be asked to remove it.

For readers with a keen interest in the Shapley Value, Algaba, Fragnelli, and Sánchez-Soriano (2019) may meet their needs. This book dedicates 24 chapters, written by various authors, to the Shapley Value.

Reference

Algaba, Encarnación, Fragnelli, Vito, and Sánchez-Soriano, Joaquín (editors) (2019). *Handbook of the Shapley Value (First Edition)*. Chapman and Hall/CRC.

Conklin, Michael, Powaga, Ken, and Lipovetsky, Stan (2004). Customer Satisfaction Analysis: Identification of Key Drivers. *European Journal of Operational Research*, Volume 154, Pages 819 – 827.

Liang, Jingyi (2021). ShapleyValue 0.2.0: Shapley Value Regression for Relative Importance of Attributes. https://CRAN.R-project.org/package=ShapleyValue.

Lipovetsky, Stan and Conklin, Michael (2001). Analysis of Regression in Game Theory Approach. *Applied Stochastic Models in Business and Industry*, Volume 17, Pages 319 – 330.

Shapley, Lloyd Stowell (1953). A Value for N-Person Games. *Contributions to the Theory of Games (AM-28), Volume II*. Pages 307-317. Princeton University Press.

RE-ENGINEERING THE 2026 POPULATION CENSUS - A PIONEERING APPROACH TO COST-EFFECTIVENESS AND DATA QUALITY

Mr. Alex YIU Cheuk-wing, Mr. Timothy CHU Tin-chung, Mr. Sam TSE Chun-sum
Census and Statistics Department

I. Introduction

1. The 2026 Population Census (26C) is underway, with the data collection having commenced in January. 26C marks a major milestone in the modernisation of Hong Kong's population censuses. The Census and Statistics Department (C&SD) has substantially re-engineered the census methodology and operations. 26C represents a methodological breakthrough, under which C&SD makes far more extensive use of government administrative data to produce benchmark population figures and supplement census data. This shall enhance data quality and significantly reduce respondent burden. Drawing on lessons learned from uncontrollable events such as the COVID-19 pandemic, 26C extends the data collection period from 1.5 months to a full year running throughout 2026, thereby reducing operational risks and enhancing cost-effectiveness. This unprecedented full-year data collection strategy is supported by an innovative, risk-based statistical mechanism to ensure data collected throughout the year better reflect the mid-year position as at 30 June 2026 and to maintain comparability with past population censuses/by-censuses.

2. Taken together, the above re-engineering initiatives constitute a pioneering approach to advancing the population census process in Hong Kong and transitioning from a traditional census approach to a combined census approach. With more extensive use of administrative data and the adoption of a full-year data collection strategy, the total costs incurred for the 2026 and 2031 Population Censuses will be reduced by 40%, translating into savings of around HK$680 million at current prices. The manpower requirement for interviewers is also

substantially reduced, from thousands in past population censuses to about 200 for 26C. This paper sets out the various re-engineering initiatives of 26C to facilitate readers' understanding of the latest developments in Hong Kong's population censuses. The proposed implementation procedures of 26C were tested in a Small-scale Test Survey conducted from April to July 2025, from which the experience gained is also shared in this paper.

II. Review of the overall design of Hong Kong's population census

3. Population censuses are conducted in Hong Kong regularly to provide up-to-date benchmark information on the demographic and socio-economic characteristics of the Hong Kong population and on its geographical distribution. Census data are vital references for government planning and policy formulation, as well as for business and academic research.

4. From 1961 to 2021, Hong Kong conducted a population census every ten years (i.e. in years ending in "1"), with a population by-census in between (i.e. in years ending in "6"). The population census involved a comprehensive enumeration of the entire population. In the 2011 and 2021 rounds, it covered about nine-tenths of households using a "short form" questionnaire for simple enumeration and the remaining one-tenth using a "long form" questionnaire for detailed enquiry. The large-scale short form enumeration was designed to produce accurate benchmark population figures on size and age-sex structure. However, such a large-scale undertaking was resource-intensive and exposed to considerable operational risks. By contrast, the population by-census selected about one-tenth of households (since 2006) for detailed enquiry using the long form only.

III. Re-engineering initiatives of 26C

5. C&SD conducted a Business Process Re-engineering (BPR) Study to formulate strategic initiatives to drive census modernisation. C&SD conducted

an in-depth review on the latest international practices, thoroughly examined Hong Kong's local circumstances and the needs of data users, and engaged with experts extensively. Specifically, as a major step to modernise population censuses, and to reduce respondent burden, operational risks and costs, the following initiatives are implemented from 2026 onwards:

◆◆◆

Initiative 1: adopting a sampling scheme the same as population by-censuses to enumerate only about one-tenth of households for detailed enquiry every five years, while the remaining nine-tenths of households are not required to fill out any questionnaire

6. Initiative 1 introduces a register-based approach for compiling benchmark population figures using the administrative data from the Immigration Department (ImmD). This approach enhances the accuracy of the benchmark compilation and removes the need to conduct the large-scale, resources-intensive short form enumeration, thereby delivering significant cost-effectiveness and reducing respondent burden dramatically.

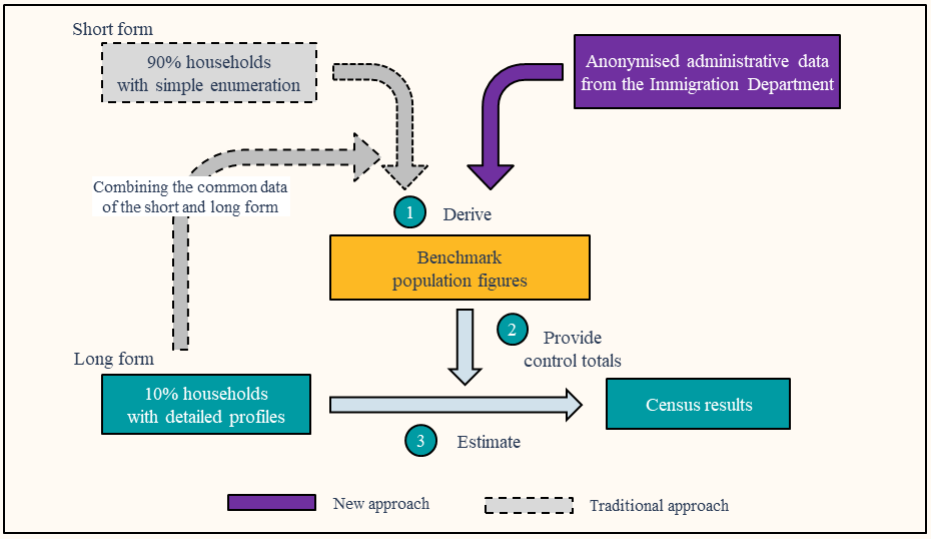
7. In recent years, C&SD has been exploring the feasibility of compiling benchmark population figures by a new approach with more comprehensive use of anonymised administrative data from ImmD. ImmD provides immigration and vital registration services to Hong Kong residents and persons intending to come to Hong Kong, including the issuance of identity documents, visas and entry permits, and the registration of births and deaths. By integrating all relevant registration data furnished by ImmD upon anonymisation, a comprehensive database of eligible residents, comprising Hong Kong Permanent Residents and Hong Kong Non-permanent Residents, can be established by C&SD to serve as an extended register to facilitate the counting of the population at specific reference time-points.

8. Moreover, the arrivals and departures of every person through the border control points of Hong Kong are also comprehensively administered by ImmD. By applying record linkage techniques to each person on the register of eligible residents, C&SD can retrieve passenger movement records (i.e. arrival and departure records at individual level) and calculate their duration of stay in specific periods. Hence, each individual on the register can be classified into Usual Residents, Mobile Residents, and Non-Residents, thereby establishing the benchmark population[1] at time-points as desired.

9. Building on C&SD statisticians' technical expertise in data integration and their domain knowledge of population-related administrative systems, the integration and management of ImmD's administrative data have been enhanced in recent years. C&SD has recently secured the supply of anonymised administrative data from ImmD with markedly enhanced coverage and frequency to support the derivation of benchmark population figures. This register-based approach enhances the accuracy of benchmark compilation by leveraging comprehensive population coverage in the administrative systems and avoiding errors in data collection that are inevitably inherent in the short form enumeration (e.g. nonresponse error). As a result, starting from 26C, C&SD adopts the new compilation approach, and the short form enumeration is abolished. Figure 1 presents an overview of the new approach.

[1] The concept of the Hong Kong Resident Population, which comprises "Usual Residents" and "Mobile Residents", has been adopted to compile population estimates since August 2000. "Usual Residents" cover: (a) Hong Kong Permanent Residents who have stayed in Hong Kong for at least three months during the six months before or after the reference time-point; and (b) Hong Kong Non-permanent Residents who are in Hong Kong at the reference time-point. "Mobile Residents" are Hong Kong Permanent Residents who are similarly defined as in (a) above, except that their length of stay in Hong Kong is at least one month but less than three months.

Figure 1 Overview of the new approach to derive benchmark population figures



◆◆◆

Initiative 2: Using government administrative data more extensively to supplement census data

Markedly more extensive and granular application

10. Initiative 2 transforms administrative data into valuable statistical data and integrates them as a core component of the compilation of census statistics. By directly supplementing selected census data items with administrative data, the initiative reduces the need to collect certain information through the questionnaire, thereby simplifying the questionnaire and easing respondent burden. In addition, through the development of more sophisticated statistical models that incorporate richer auxiliary information, data quality can be enhanced. These enhancements are underpinned by C&SD's cross-departmental data collaboration and the development of pioneering AI innovations to enable large volume data integration.

11. C&SD has long used administrative data in the compilation of census results. In 26C, C&SD has strengthened data collaboration with more government departments and obtained administrative data with markedly greater coverage and granularity. Key sources include quarters data from the Rating and Valuation

Department; household profiles from the Housing Department and the Housing Society; water usage data from the Water Supplies Department; data on recipients of government assistance schemes from the Social Welfare Department and the Working Family and Student Financial Assistance Agency; and education profiles of students enrolled in student health services from the Department of Health.

12. Integrating administrative data into census data is a complex undertaking. To ensure the successful application of administrative data, C&SD has established the Social Data Development Branch to drive cross-departmental data collaboration and has devoted substantial effort and technical expertise to handling the integration processes, including conceptual alignment, record linkage, data pre-processing, and statistical method development. With these sustained endeavours, C&SD is able to harness administrative data and more fully unleash their value. Some notable applications of administrative data in 26C are discussed below.

13. Simplify the questionnaire. With access to more administrative data sources with greater granularity, 26C has trimmed some census questions to ease respondent burden while enhancing data accuracy. These include questions on the subsidy amounts of various government assistance schemes (e.g. Comprehensive Social Security Assistance, Social Security Allowance, Working Family Allowance, and various student finance), rent of public rental housing units, and floor area of accommodation (except subdivided units and rural settlements). C&SD conducted research to compare data from another regular household survey with the corresponding administrative records. The findings indicated that reported and administrative data are complementary in nature. While reported data better reflect the latest profiles of households and persons at the time of enumeration, administrative data pertaining to monetary or numerical data items tend to provide more accurate information.

14. Enhance small area estimation and facilitate mid-year adjustments. In data processing, C&SD makes use of administrative data as at the mid-year position to conduct statistical adjustments, ensuring that data collected throughout the year can better reflect the mid-year occupancy situation and population distribution at small area level. By combining administrative data from various sources that are indicative of the presence of households and persons, C&SD is able to conduct a "signs of life" analysis. Such administrative data as water consumption, date of commencement of current tenancy, household profiles of public rental housing, and data on recipients of government assistance schemes, including their changes over time, can serve as informative indicators that enable C&SD to assess the mid-year occupancy status and household size at quarters level for non-contact or vacant cases, thereby improving the accuracy of small area statistics.

15. Enhance statistical imputation. More sophisticated imputation models are developed to further enhance data accuracy. For example, the age, sex, and other characteristics of household members retrieved from household profiles of public rental housing and data on recipients of government assistance schemes can be used as matching variables to identify more similar donor records for imputation. In addition, education profiles of students enrolled in student health services can also be used to resolve missing data and inconsistent values across related data items. As a result, imputation quality is enhanced, thereby improving overall data quality while reducing the need for manual data editing.

Self-develop AI-based text analytics model for linking census data with large volumes of administrative data

16. To safeguard data confidentiality, only anonymised administrative data are obtained from government departments. In linking census data with administrative data, addresses are used as the primary matching key. However, the address matching process is challenging, as it involves up to ten million

records and a common address standard has yet to be fully adopted across government departments. Variations in address representation, inconsistent formatting, and incomplete addresses are frequently encountered.

17. To ensure accurate and scalable record linkage, C&SD has self-developed an AI-based text analytics model for address matching. The model comprises two key stages. In the first stage (parsing stage), the model parses full addresses from administrative records into distinct address components, including street name and number, estate and building name, floor number, and flat number, based on the address structure defined in C&SD's address database "Frame of Quarters"[2]. In the second stage (matching stage), the model performs probabilistic matching by comparing the parsed address components with the address records in the Frame of Quarters, and automatically identifies the most probable matches based on computed similarity scores, thereby linking census data with administrative data. To strengthen accuracy at scale, C&SD has leveraged "AI-versus-AI strategies" to conduct automated quality assurance on the linkage results. At present, the model has demonstrated a high level of linkage accuracy and satisfactory performance. The AI-enabled automated parsing and probabilistic matching workflow provides a cost-effective and accurate solution, enabling high volume address matching at scale with minimal manual review.

18. While address matching is accurate, households recorded in administrative data may not necessarily correspond to those in census data. This may occur because residential addresses in administrative records might not be updated in a timely manner, or because respondents might provide a correspondence address that differs from their current accommodation when applying for certain

[2] C&SD maintains an address database "Frame of Quarters", which covers the land area of Hong Kong. It stores a list of addresses of permanent quarters in built-up areas and area segments in non-built-up areas to serve as a sampling frame for census taking and household surveys.

government services. Therefore, it is necessary to further match records based on the age and sex of household members.

◆◆◆

Initiative 3: extending the data collection period from 1.5 months to one year

19. Initiative 3 is an unprecedented data collection strategy that extends the data collection period of 26C from 1.5 months (usually in July and August) to a full year from January to December 2026, enabling a more resilient and cost-effective census operation. The extended data collection period creates a sufficient buffer to substantially mitigate the operational risks arising from uncontrollable events, such as pandemics (e.g. COVID-19) and extreme weather conditions, which may delay the release of statistics or at worst suspend the entire data collection work if the period is confined to 1.5 months. Spreading out the resource requirements over a longer period, particularly computing capacity for internal and external use, also avoids costly investments that would otherwise be required to meet inflated peak demands.

20. With a longer data collection period, a much smaller, better-trained and more efficient team of interviewers can be deployed, which is conducive to collecting high quality data. About 200 Survey Interviewers are employed in 26C, replacing the need to hire thousands of short-term and inexperienced Temporary Field Workers as in the past. The overheads in recruitment and staff training, together with the investment in office accommodation, equipment, and related logistics support, are greatly reduced.

21. While 26C data collection work spans a year, the reference moment of 26C data remains at mid-year, i.e. 30 June 2026. An innovative, risk-based statistical mechanism has been devised to ensure that the data collected throughout the year better reflect the mid-year position and that 26C results remain comparable with those of past population censuses/by-censuses conducted near mid-year. The methodology is discussed in detail in paragraphs 22 to 30 below.

IV. Reflect mid-year position

22. Under the full-year data collection strategy, data collected in the first five months of 2026 refer to the reporting date (e.g. respondents sampled in January 2026 are asked to report data as at January 2026). There is a possibility that the characteristics of the reported households and persons may change between the reporting date and mid-year. Specifically, while most demographic characteristics (e.g. year of birth, sex, ethnicity, place of birth, use of languages, and highest level of educational attainment) and housing characteristics (e.g. tenure of accommodation and number of rooms in the residence) are expected to remain largely stable or unchanged over time, such economic characteristics as employment and income may vary, and some households may move in or out. Therefore, statistical adjustments are required to bring the data to the census reference moment on 30 June 2026. For data collected in the remaining months of 2026, the information refers to the census reference moment (respondents sampled in June and later months are asked to report data as at the census reference moment). Different statistical adjustments will be applied to data collected in the first five months and the remaining months of 2026, as discussed below.

Data collected in the first five months of 2026
(before the census reference moment)

23. C&SD has devised a quality assurance mechanism to ensure that the data collected in the first five months of 2026 are coherent with the mid-year position in a cost-effective manner. The mechanism identifies households with a likelihood of change in socio-economic characteristics for targeted confirmation and updating in the second half of 2026 (termed “confirmation checks”) through the telephone interview. The mechanism comprises two strategies: sample allocation and risk-based identification.

Sample allocation

24. C&SD has designed sample allocation strategies to prioritise the enumeration of households expected to exhibit more stable profiles in the first five months of 2026, thereby reducing the number of cases requiring subsequent confirmation checks.

25. All along, the Frame of Quarters has been adopted as the sampling frame for census taking, with the unit of quarters being the sampling unit. To ensure a complete and up-to-date sampling frame covering all units of quarters as at the mid-year position, C&SD has been conducting regular updates using administrative data from various sources, including occupation permits and demolition consents, supplemented by on-site updates undertaken by experienced field officers. These administrative data are advance records in that the registration of occupation permits and demolition consents usually takes place well before residents actually move in or out. Therefore, C&SD is able to identify small areas in advance with a higher degree of variability in occupancy, such as those with newly completed buildings or buildings pending demolition. Enumeration in these areas is scheduled around mid-2026 to accurately capture the occupancy situation as at the mid-year position.

26. In addition, the sample allocation is specially designed such that the sampled quarters for the first five months of 2026 predominantly consist of public rental housing units and subsidised housing units[3]. Based on comparator data from household surveys regularly conducted by C&SD, it is observed that residents of these types of housing units generally exhibit relatively stable socio-economic characteristics over time and are less likely to change their usual residence. Furthermore, to contain memory lapse, the sample is primarily

[3] Public rental housing units are units of quarters constructed by the Government for letting to low-income households, while subsidised housing units are units under some Government programmes to offer more affordable flats to families at discounted prices.

assigned up to October 2026, while November and December mainly serve as a buffer period for follow-up on outstanding cases and cases selected for quality assurance.

Risk-based identification

27. C&SD conducts risk-based identification to select households for targeted confirmation and updating. In particular, C&SD uses the following means to identify households reported in the first five months with a likelihood of change in socio-economic characteristics and the corresponding parts of the questionnaire that require updating:

a. adding risk assessment questions to the questionnaire for the first five months, e.g.:
   - whether household members will move in or out;
   - whether the person plans to change jobs (e.g. due to seeking better remuneration, upon completion of contract, or job turnover);
b. identifying selected characteristics based on existing census questions, e.g.:
   - final-year post-secondary students and unemployed persons, as they might take up employment later;
c. referencing administrative data, e.g.:
   - date of commencement of current tenancy for public rental housing and new registrations of water accounts.

Households are classified into five categories, as summarised in Table 1. Each category is subject to a specific selection proportion for confirmation checks. A higher proportion is applied to households expected to experience greater changes in socio-economic characteristics. As a prudent control measure, about 1% of households identified as “virtually no change” are selected for confirmation checks.

28. For selected cases in each category, field officers will verify with respondents to confirm or update the relevant parts of the questionnaire in the second half of 2026, without requiring respondents to go through the full set of questions again. This arrangement ensures cost-effectiveness and minimises respondent burden.

29. For the first two categories in Table 1, which are subject to 100% selection for confirmation, updated mid-year information will be collected for all records. For the other categories, where only a proportion of records is selected for confirmation, statistical modelling methods are to be applied to adjust records not selected for confirmation. The methods primarily involve the application of machine learning techniques to identify important features (e.g. age, industry, occupation, monthly earnings level, etc.) for constructing strata with more notable changes. Transition probabilities of stage changes (e.g. changes in economic activity status, and shifts across industries and occupations) and distribution drifts (e.g. income) will be estimated for these strata based on the results of confirmation checks and applied to records not selected for confirmation. In addition, administrative data from various sources will be used to ascertain the mid-year occupancy status and the usual residence of households and members, and to apply corresponding adjustments to the data.

Table 1 Confirmation checks: category, coverage of households, and proportion for confirmation

	Category		Coverage of households	Proportion for confirmation
1	Changes in usual residence and household formation	-	Households/members anticipating to move in or out	100%
2	Changes in economic activity status	-	Final-year post-secondary students;	100%
		-	unemployed persons;	
		-	employers/self-employed persons (e.g. small business owners and freelancers)	
3	Changes in employment income	-	Persons planning to change jobs (e.g. due to seeking better remuneration, upon completion of contract, or job turnover);	20%
		-	persons remunerated under selected payment modes (e.g. commission);	
		-	persons with secondary employment;	
		-	persons anticipating discretionary bonus in June 2026	
4	Changes in cost of housing	-	Households anticipating changes in rent paid or mortgage payment for their unit of quarters (e.g. due to tenancy renewals or mortgage adjustments)	10%
5	Others (virtually no change)	-	Households other than those listed above	1%

Data collected in June to December 2026
(at/after the census reference moment)

30. As regards households enumerated between June and December 2026, the information as at the census reference moment is collected. For households moving into the sampled unit of quarters after the census reference moment, the standing census practice is applied, under which the household currently living in the unit of quarters is enumerated and treated as the nearest neighbour for imputing the characteristics of the moved-out household. In addition, the mid-year occupancy status and the presence of households and members are to be assessed with the aid of administrative data.

V. Experience from the Small-scale Test Survey

31. C&SD conducted a Small-scale Test Survey (STS) of about 9 000 households from April to July 2025 to test various operational aspects of 26C, including questionnaire design, data collection systems, fieldwork management, and confirmation checks.

32. Confirmation checks were found effective in identifying households with changes, and the updates to socio-economic characteristics were successfully captured. Respondents were generally cooperative, and a high completion rate was achieved. In most cases, confirmation took only a few minutes. Feedback from respondents indicated that the risk assessment questions were generally clear and easy to understand, and that the confirmation process was easy to follow.

33. Assessments were conducted using the STS data to analyse how extending the data collection period affects the coherence of data in reflecting the mid-year position. The STS data revealed that socio-economic characteristics of households were generally stable over time, and only a few changes in usual residence were observed. In addition, analyses of administrative data from

different time-points showed that profiles of households and members exhibited a high degree of continuity over time. With modest mid-year adjustments, data collected over the extended data collection period shall represent the mid-year position accurately. The STS findings confirmed the feasibility and robustness of adopting the extended data collection period.

VI. Concluding remarks

34. As demonstrated in this paper, 26C represents a pioneering approach to modernising Hong Kong's population censuses. C&SD is advancing census innovations through three re-engineering initiatives, so as to enhance data quality, ease respondent burden, and reduce operational risks and costs. In particular, 26C makes more extensive and granular use of government administrative data and deepens cross-departmental data collaboration through closer partnerships with various government departments, thereby laying down a solid foundation for Hong Kong's transition from a traditional census approach to a combined census approach. In addition, an unprecedented full-year data collection strategy is adopted with a risk-based statistical mechanism for mid-year adjustments. This shall notably strengthen the operational resilience against uncontrollable events and achieve significant cost-effectiveness.

35. Census re-engineering is both methodologically and operationally complex. Implementing these initiatives requires careful planning and sustained capability building. Over the years, C&SD has been strengthening its organisational capability in statistical excellence, information technology and domain knowledge, thereby enabling the adoption of innovations to address emerging challenges in census re-engineering work. This includes strengthening AI and data science capabilities and developing cutting-edge AI applications to tackle complex business problems that were previously prohibitively difficult to solve with traditional methods and tools.

36. While C&SD is committed to implementing the re-engineering initiatives for 26C, public participation remains indispensable to its success. In closing, we call on all members of society to support 26C. Every response counts and contributes to Hong Kong's future. For the latest updates on 26C, please visit the 26C thematic webpage: www.census2026.gov.hk.

2024/25 STATISTICAL PROJECT COMPETITION FOR SECONDARY SCHOOL STUDENTS

Organising Committee* of the 2024/25 Statistical Project Competition

The 2024/25 Statistical Project Competition (SPC) for Secondary School Students, the 39th 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 responsibility.

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 submit their projects 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 *"Hang Seng Indexes Company Limited Prize for the Best Index Application"*, the *"Department of Decision Analytics and Operations, the City University of Hong Kong Prize for the Best Graphical Presentation of Statistics"* and the "*Division of Science, Engineering and Health Studies, PolyU CPCE Prize for the Most Creative Project"*.

To help interested participants prepare for the Competition, a Briefing Seminar for 2024/25 SPC was held on 19 October 2024. Professor CHEUNG Ka-chun of The University of Hong Kong, Chief Adjudicator, introduced the rules and regulations of the Competition. Representatives from the Census and Statistics Department (Mr. Wilson YUNG Wai-hei) and Hang Seng Indexes Company Limited (Mr. Danny CHU Wai-kit), sponsor of 2024/25 SPC, had introduced the use of official statistics and "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 19 October to 22 November 2024.

Group photo of speakers taken in the Briefing Seminar

Encouraging number of entries

In 2024/25 SPC, 150 entries and 71 entries were submitted for the Junior Section and the Senior Section respectively from 854 students of 81 secondary schools. While the total number of entries was similar to that in previous round, the number of participating schools had reached a record high, reflecting the overwhelming responses and promising support from the academia. Highlighting a variety of themes, the entries explored numerous socio-economic issues relevant to Hong Kong.

Adjudication panel led by Professor CHEUNG Ka-chun

An adjudication panel, led by the Chief Adjudicator, Professor CHEUNG Ka-chun, and comprised some 50 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

Dr. Alpha LING Man-ho of The Education University of Hong Kong and Professor CHEUNG Ka-chun for serving as the Chairpersons of the interview panel for Junior and Senior Sections respectively.

Prize Presentation Ceremony

The Prize Presentation Ceremony for the 2024/25 SPC took place on 21 June 2025 at the Auditorium of the Hong Kong Federation of Youth Groups Building. Professor CHAN Ngai-hang, President of the HKSS, Mr. CHAU Kam-tim, Acting Commissioner for Census and Statistics, Ms. Edith TSE Yuen-ching, Principal Assistant Secretary of the Education Bureau, Ms. Candy LAM Wing-ting, Chief Strategy Officer of Hang Seng Indexes Company Limited, Ms. Teresa NG Kit-yee, representative of the Department of Decision Analytics and Operations of City University of Hong Kong, Dr. Anthony LOH Wai-keung, representative of the Division of Science, Engineering and Health Studies of PolyU CPCE and Professor CHEUNG Ka-chun, Chief Adjudicator, were invited to the Ceremony to present prizes and trophies to the winning teams. Council members of HKSS, including Ms. Olivia OR Oi-yan and Dr. Wilson KWAN Chun-kit also attended the Ceremony to share the joy of the winning teams.

Group photo taken in the Prize Presentation Ceremony

Regarding the results of the Competition, students of Good Hope School, who used official statistics to study the medical negligence in Hong Kong, won the First Prize for the Junior Section. Students of Wa Ying College won the Second Prize and the Prize for the Best Index Application, while students of Diocesan Girls' School won the Third Prize. Another group of students of Diocesan Girls' School won the Best Graphical Presentation of Statistics, while students of Pui Ching Middle School won the Prize for the Most Creative Project.

Mr. CHAU Kam-tim presented the First Prize for the Junior Section to students of Good Hope 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, and won both the Best Graphical Presentation of Statistics and the Prize for the Most Creative Project. They applied official statistics to study tourism development and policies in Hong Kong. Students of St. Paul's Co-educational College won the Second Prize and the Prize for the Best Index Application, while another group of students of St. Paul's Co-educational College won the Third Prize.

Ms. Candy LAM presented the First Prize for the Senior Section to students of Stewards Pooi Kei College


2024/25 年度
中學生統計習作比賽頒獎典禮
Prize Presentation Ceremony of
2024/25 Statistical Project Competition
for Secondary School Students
香港統計學會
Hong Kong Statistical Society
教育局
Education Bureau
恒生指數
HANG SENG INDEXES


2024/25 年度
中學生統計習作比賽頒獎典禮
Prize Presentation Ceremony of
2024/25 Statistical Project Competition
for Secondary School Students
香港統計學會
Hong Kong Statistical Society
教育局
Education Bureau
恒生指數
HANG SENG INDEXES

Professor CHAN Ngai-hang, Ms. Edith TSE, Ms. Teresa NG, Dr. Anthony LOH and Professor CHEUNG Ka-chun presented prizes to winning teams

Social Media

Social media platforms (including a Facebook page and an Instagram for SPC) have been set up for promoting the SPC. Interested members can scan the QR codes below and follow our social media pages for updated information.

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 their adjudication process and ceremony.

**Organising Committee for the 2024/25 SPC :*

Dr. Thomas LO	Census and Statistics Department
Mr. CHAN Sau-tang	Education Bureau
Mr. Benjamin CHAN	Census and Statistics Department
Mr. Brian CHENG	Census and Statistics Department
Mr. Calvin CHU	Census and Statistics Department
Mr. James FONG	Census and Statistics Department
Mr. Tony HO	Census and Statistics Department
Mr. Marco KWAN	Census and Statistics Department
Mr. Ian NG	Census and Statistics Department
Dr. Zoe XIAO	Census and Statistics Department

NEWS SECTION

Personnel changes (New Appointments, Promotions and Retirements)

City University of Hong Kong (CityUHK) - Department of Decision Analytics and Operations

- Prof. Gavin Guanhao FENG was promoted to Associate Professor in 2025.
- Prof. Jingyu HE was promoted to Associate Professor in 2025.
- Prof. Biao CAI joined the department as Assistant Professor.
- Prof. Yiling XIE joined the department as Assistant Professor.

The Chinese University of Hong Kong (CUHK) - Department of Statistics and Data Science

- Prof. LIN Yingxin joined the department as Assistant Professor with effect from August 2025.
- Prof. Tony SIT was promoted to Professor with effect from September 2025.

The Education University of Hong Kong (EduHK) - Department of Mathematics and Information Technology

- Dr. SUN Daner was promoted to Associate Professor with effect from July 2025.
- Dr. LO Chung Kwan was promoted to Associate Professor with effect from January 2026.
- Dr. ZHANG Junyi joined the department as Assistant Professor with effect from August 2025.
- Dr. LI Xin Stephen joined the department as Assistant Professor with effect from September 2025.

- Dr. TING Fridolin Sze Thou joined the department as Senior Lecturer I with effect from September 2025.
- Dr. YE Huiyan joined the department as Research Assistant Professor with effect from December 2025.
- Dr. ZHANG Ying Bob joined the department as Research Assistant Professor with effect from December 2025.

The Hong Kong Polytechnic University (PolyU) - Department of Applied Mathematics

- Prof. LENG Chenlei joined the department as Chair Professor of Statistics and Machine Learning with effect from September 2025.

NEWS SECTION

Seminar on "Integration of AI in Statistical Work: Challenges and Opportunities"

A seminar on "Integration of AI in Statistical Work: Challenges and Opportunities" was co-organised by HKSS and the Department of Biostatistics of City University of Hong Kong (CityUHK) and was held on 7 January 2026 at CityUHK. Professor Xuming HE, the Chair of the Department of Statistics & Data Science, and Kotzubei-Beckmann Distinguished Professor of the Washington University in St. Louis, USA, was the speaker for the seminar. Professor HE was also a joint editor for the Journal of the Royal Statistical Society - Series B (Statistical Methodology), as well as the former President (2023-2025) of the International Statistical Institute.

Professor HE delivered his talk in the seminar.

Group photo taken in the seminar

In the seminar, Professor HE discussed the use of synthetic data and adaptive data collection, and highlighted the importance of statistical principles when AI is integrated into data-driven discovery and decision-making processes. The seminar was conducted in hybrid mode. More than 130 participants attended the seminar. A dinner with Professor HE was organised after the seminar.

Some snapshots of the speaker and audience taken during the seminar.

A dinner with Professor HE was organised after the seminar.

NEWS SECTION

Public Lecture on "Quality Statistics and Data For Everyone" on World Statistics Day

The United Nations General Assembly had designated 20 October 2010 as the first World Statistics Day (WSD) to acknowledge the importance of reliable and timely statistics for informed policy decisions, and had further decided to celebrate the World Statistics Day every five years on 20 October. The fourth WSD, commemorated on 20 October 2025 under the theme "Driving change with quality statistics and data for everyone", highlights the growing importance of high-quality statistics and data in a rapidly changing world, and that statistics facilitate informed research, planning and decision-making in various sectors.

To celebrate the fourth WSD, the Hong Kong Statistical Society (HKSS) jointly organised a public lecture on "Quality Statistics and Data For Everyone" with the Education University of Hong Kong (EdUHK), the Census and Statistics Department (C&SD), and the Education Bureau of the Hong Kong Government on 20 Oct 2025. The public lecture was held at the Tseung Kwan O Study Centre of EdUHK with 35 participants in total.

Professor John LEE, the President of EdUHK, gave a pre-recorded opening message at the event, while Mr. Leo YU, Commissioner for Census and Statistics, and Professor CHAN Ngai-hang, President of HKSS, officiated the event. Professor Philip YU, a professor at the Department of Mathematics and Information Technology of EdUHK, who is also the Associate Director of the University Research Facility of Data Science and Artificial Intelligence at EdUHK, introduced the speakers and moderated the Q&A Session.

Professor John LEE →

← Mr. Leo YU

Professor CHAN Ngai-hang →

Professor John LEE, Mr. Leo YU, and Professor CHAN Ngai-hang delivering their welcoming messages and opening speeches at the start of the public lecture.

Professor YU introduces the three speakers.

Three talks were presented: (1) "Leveraging AI and Statistics to Facilitate Data-Driven Business Decisions", delivered by Ms. Polly WONG, Former Marketing Science Partner of Meta Hong Kong; (2) "Innovation in the 2026 Population Census", delivered by Mr. Michael CHOI, a Statistician in C&SD; and (3) "Can Data Speak for Itself (Themselves)? – Basic Statistical Principles that Are Still Valid in the Era of Big Data", delivered by Professor LI Wai-keung, the Dean of the Faculty of Liberal Arts and Social Sciences, and the Research Chair Professor of Data Science at the Department of Mathematics and Information Technology of EdUHK.

Ms. Polly WONG (Left), Mr. Michael CHOI (Middle), and Professor LI (Right) giving their talks in the public lecture.

Group photo of the guests of honour and speakers (from left): Professor Philip YU, Mr. Michael CHOI, Ms. Polly WONG, Professor LI Wai-keung, Mr. Leo YU and Professor CHAN Ngai-hang.

The event received favourable responses from teachers in primary and secondary schools, students, government officials and the general public. To recapitulate the public lecture again, please visit the website: https://www.hkss.org.hk/index.php/events/list/academic/event/97-public-lecture-on-quality-statistics-and-data-for-everyone

Professor LI answered a question from the audience.

NEWS SECTION

HKSS' participation in the Fourth Guangdong-Hong Kong-Macao Greater Bay Area Statistical Forum

The Fourth Guangdong-Hong Kong-Macao Greater Bay Area Statistical Forum was held on 26th September, 2025. The forum, themed "Statistical Innovation and Technological Breakthroughs in the Guangdong-Hong Kong-Macao Greater Bay Area", aimed to foster statistical exchange and co-operation in the Greater Bay Area (GBA). The forum was attended by representatives of statistical organisations from Beijing, Guangdong Province, Hong Kong and Macao, etc. Relevant officials from the statistical systems of the cities in the GBA, as well as experts and scholars from national think tanks and renowned universities, also shared their valuable insights on topics such as applications of data science techniques, development of statistical talent, statistical reform and innovation, economic monitoring and forecasting, etc.

The emcee kicked off the opening ceremony of the forum.

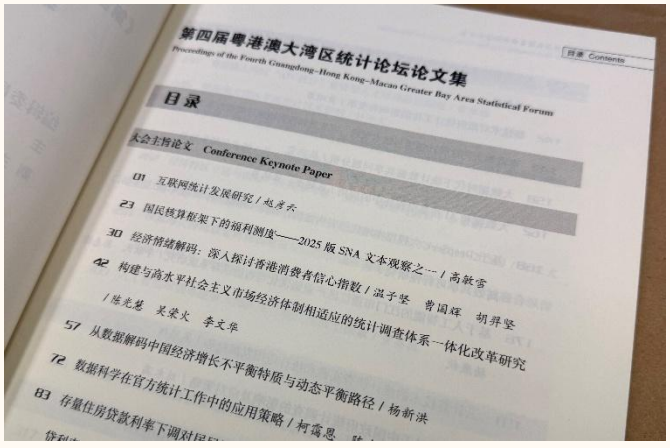
HKSS had been invited by Guangdong Provincial Bureau of Statistics to attend the forum and deliver a presentation. On behalf of HKSS, Dr. Geoffrey TSO Kwok-fai delivered a presentation on <<*Decoding Economic Sentiments: A Deep Dive into Hong Kong's Consumer Confidence Index*>> at the forum. Dr. TSO

introduced the compilation methodology of the Hong Kong Consumer Confidence Index (HKCCI), and the six sub-indices.

Dr. TSO delivered his presentation through a pre-recorded video

Dr. TSO also highlighted that the HKCCI serves as a crucial barometer for assessing economic conditions and consumer sentiment in Hong Kong; and the HKCCI, the Household Costs Indices, and the Consumer Price Index each serves a complementary role in providing a comprehensive view of economic conditions in Hong Kong. In addition, Dr. TSO contributed a paper of the same title, co-authored with Professor Alan WAN Tsz-kin and Mr. William HU Yijian, which was published in the proceedings of the forum.

Dr. TSO highlighted the merits of the HKCCI and its sub-indices.



The paper by Professor WAN, Dr. TSO and Mr. HU was published in the proceedings of the forum

NEWS SECTION

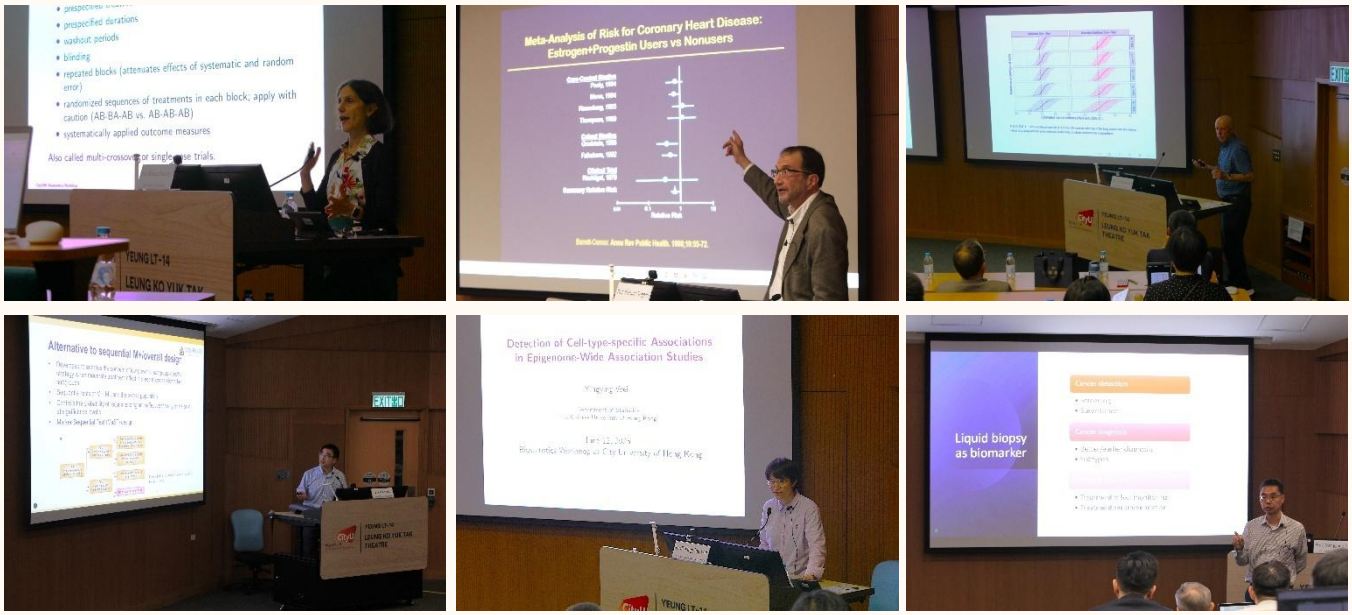
City University of Hong Kong Biostatistics Workshop 2025

With the support of HKSS, the Department of Biostatistics at City University of Hong Kong (CityUHK) hosted a 2-day Biostatistics Workshop during 12-13 June 2025 at CityUHK to explore the *Statistical Lessons of HK Nasopharyngeal Carcinoma Trials and Related Topics*. Professor CHAN Ngai-hang, the President of HKSS, who is also the Head and Chair Professor of the Department of Biostatistics at CityUHK, delivered the opening speech for the workshop.

Professor CHAN delivered the opening speech for the workshop

Group photo of Professor CHAN, speakers and some of the participants

A total of 8 speakers delivered talks during the workshop, including Professor Rebecca BETENSKY from the New York University, Professor Richard CHAPPELL from the University of Wisconsin, Professor Chen HU from the Johns Hopkins University, Professor Ian McKEAGUE from the Columbia University, Professor Menggang YU from the University of Michigan, Professor Yingying WEI of the Chinese University of Hong Kong, Professor Zudi LU and Professor Yi YANG from CityUHK. The workshop was conducted in hybrid mode, and more than 70 participants attended the workshop.



Professor BETENSKY (top left), Professor CHAPPELL (top middle), Professor McKEAGUE (top right), Professor HU (bottom left), Professor WEI (bottom middle) and Professor YU (bottom right) delivered their talks during the workshop.

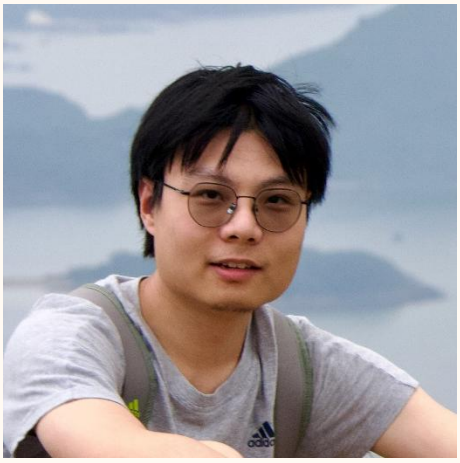
Professor BETENSKY (left), Professor HU (middle) and Professor CHAPPELL (right) in the Q&A session.

NEWS SECTION

Result of the HKSS - John Aitchison Prize in Statistics, 2025/26

By the deadline of December 19, 2025, high quality submissions were received for the prize. After careful deliberation, the adjudication panel decided to jointly award the HKSS - John Aitchison Prize in Statistics 2025/26 to both Dr. Youqian GAO (2025 PhD Graduate at the Chinese University of Hong Kong) and Dr. Zhiwei WANG (2025 PhD Graduate at the Hong Kong University of Science and Technology) for their papers, "*Word-Level Maximum Mean Discrepancy Regularization for Word Embedding*" and "*A Unified Framework for Identification of Cell-Type-Specific Spatially Variable Genes in Spatial Transcriptomic Studies*", respectively.

Dr. Youqian GAO



Dr. Zhiwei WANG

Dr. GAO's paper has been published in the Journal of the American Statistical Association, Vol. 121, Issue 553 (2026). His paper is a joint work with Ben DAI. His paper introduces a novel regularization framework, word-level MMD (wMMD), to address overfitting in word embedding models, a persistent challenge in natural language processing. By explicitly leveraging word-level distribution discrepancies across labels, the proposed method enhances robustness and generalization while effectively serving as a dimension-reduction mechanism. Supported by rigorous theoretical analysis and strong empirical performance on simulated and real-world datasets, wMMD consistently outperforms existing regularization techniques. Dr. GAO's research makes a

significant methodological contribution with clear practical relevance, advancing both the theory and application of embedding-based NLP models.

Dr. WANG's paper has been published in the Proceedings of the National Academy of Sciences of the United States of America, Nov 2025, Vol. 122, No. 46. His paper is a joint work with Yeqin ZENG, Ziyue TAN, Yuheng CHEN, Xinrui HUANG, Hongyu ZHAO, Zhixiang LIN and Can YANG. His paper presents the Mixture of Mixed Models (MMM), an elegant and powerful framework for identifying cell-type-specific spatially variable genes in spatial transcriptomics data. By directly modeling RNA count data, correcting for platform effects, and accounting for cell type composition, MMM achieves high statistical power while maintaining rigorous error control. Demonstrated across diverse datasets and technologies, including high-resolution and 3D tissue models, the method delivers robust and biologically meaningful discoveries. MMM represents a significant methodological advance, providing a versatile and principled tool for uncovering spatial gene expression patterns underlying complex traits and diseases.

Congratulations to the awardees. We would also like to express our gratitude to the organising committee of the Prize for their time and contribution including the adjudication work. The organising committee for the HKSS - John Aitchison Prize in Statistics 2025/26 is as follows:

Chairperson

Professor Alan WAN Tze-kin	City University of Hong Kong

Adjudication Panel

Professor Alan WAN Tze-kin	City University of Hong Kong *(Chair of Adjudication Panel)*
Professor Rob HYNDMAN	Monash University
Professor Jinchi LV	University of Southern California

Secretary

Mr. Evan DAI Lingqi	City University of Hong Kong