Appendix 3 Calculation of price indices of items by web scraping

Web scraping is a computer software technology that extracts information from websites, and the utilization of this technology allows efficient collection of information on products (goods and services) sold on websites. In general, information on websites includes data on products (goods and services) that have few sales records, but by obtaining detailed information on sales records from companies that operate the websites, the noise of those data can be removed, and the prices of products with the same quality, which is the premise for identifying price trends, can be tracked.

In light of the above circumstances and the increase in online purchases in recent years, the CPI is calculated by utilizing web scraping technology from the 2020-base for Airplane fares, Hotel charges, and Charges for package tours to overseas. By doing this, the price collection period (improvement from the collection of "price on a specific day of each month" to the collection of "daily price") and the number of plans adopted have increased dramatically, which contributes to the improvement of the statistical accuracy of the CPI.

For Airplane fares, Hotel charges, and Charges for package tours to overseas, price indices are calculated by following the steps below.

1 Airplane fares

(1) Reservation time and fare

The prices of various discount fares on all days of each month are adopted in accordance with the normal fares and the fare discount systems[55] of airlines and the purchase time of consumers. However, tickets for specific persons, such as first-class and business class tickets with enhanced services, as well as tickets for children and seniors, are excluded.

(2) Route

15 representative routes with a large number of passengers are selected, and all departure and arrival flights operated in each section are selected.

(3) Airline

Based on the passenger status of airlines, airlines with a high share of passengers are preferentially selected.

(4) Method of index calculation

Price indices are calculated by following the steps 1) and 2) below.

1) The minimum prices are calculated by route, airline, fare type, boarding date, and flight, and are simply averaged by the number of flights to calculate average prices by route, airline, fare type, and boarding date. In addition, the average price for each route, airline, and fare type is

[55] A discount fare system in accordance with the period of advanced purchase: 28 to 44 days before boarding date, 45 to 54 days before, 55 to 74 days before, and 75 days before or earlier.

calculated by averaging with weights using the share of the number of flights handled on weekdays (excluding days before holidays) and on days other than weekdays ($q_{0,d}$).

$$P_{t,a,b,c,d} = \frac{\sum_e P_{t,a,b,c,d,e}}{n_{t,a,b,c,d}}$$

$$P_{t,a,b,c} = \frac{\sum_d P_{t,a,b,c,d} q_{0,d}}{\sum_d q_{0,d}}$$

t: Comparison period, 0: Base period, *a*: Route, *b*: Airline
c: Fare type, *d*: Boarding date, *e*: Flight

2) The average prices are calculated by averaging with weights using the share of flights handled ($Q_{0,a,b,c}$) by route, airline, and fare type. Finally, the price index is calculated by dividing by the price in the base period.

$$P_t = \frac{\sum_{a,b,c} P_{t,a,b,c} Q_{0,a,b,c}}{\sum_{a,b,c} Q_{0,a,b,c}}$$

$$I_\mathrm{t} = \frac{P_t}{P_0} \times 100$$

t: Comparison period, 0: Base period, *a*: Route, *b*: Airline, *c*: Fare type

2 Hotel charges

(1) Reservation time and accommodation plan

In light of the release time of accommodation plans by travel agencies and the purchase time of consumers, a price of "Japanese-style room, one night with two meals" is adopted for inns, and a price of "Western-style room, one night with breakfast" is adopted for hotels for all days of each month*. However, by the following outlier exclusion process, plans with extremely high rates (or extremely low rates during a sale) as compared with general hotel charges are excluded.

(2) Accommodation facility

Based on the number of guests by travel destination (prefecture) and the facility scale (the number of people that can be accommodated) in the "Accommodation Travel Statistics Survey" (Japan Tourism Agency), about 400 representative accommodation facilities are selected by prefectures.

(3) Travel agency

Based on the transaction volume of travels handled by major travel agencies, travel agencies with a high share of transaction volume are preferentially selected.

* As a general rule, prices collected at the beginning of the month 2 months before the accommodation date are used.

(4) Method of index calculation

Price indices are calculated by following the steps 1) to 4) below. For the calculation, the data set for two months of current month (t) and the previous month ($t-1$) is used.

1) Outliers are excluded by following the steps below.

(a) The individual prices for each reservation website (s), accommodation date (a), accommodation facility (b), and plan (c) are set to $P_{s,a,b,c}$, and are then subject to logarithmic conversion.

$$Y_{s,a,b,c} = \log(P_{s,a,b,c})$$

(b) The average price and standard deviation are calculated for each reservation website, accommodation date, and accommodation facility. ($N_{s,a,b}$ is the number of plans)

$$Y_{s,a,b} = \frac{1}{N_{s,a,b}} \sum_{c=1}^{N_{s,a,b}} Y_{s,a,b,c}$$

$$\sigma_{s,a,b} = \sqrt{\frac{1}{N_{s,a,b}-1} \sum_{c=1}^{N_{s,a,b}} \left(Y_{s,a,b,c} - Y_{s,a,b}\right)^2}$$

(c) Individual prices whose difference from the average price are more than 3 times the absolute value of the standard deviation are set as outliers for each reservation website, accommodation date, and accommodation facility.

$$\left|Y_{s,a,b,c} - Y_{s,a,b}\right| > 3\sigma_{s,a,b}$$

2) For individual prices excluding outliers, the average price is calculated for each reservation website, accommodation date, and accommodation facility, and a data table having these as attributes is created ($N'_{s,a,b}$ is the number of individual prices excluding outliers).

$$Y'_{s,a,b} = \frac{1}{N'_{s,a,b}} \sum_{c=1}^{N'_{s,a,b}} Y_{s,a,b,c}$$

3) Missing values are complemented by following the steps below.

(a) By using the data table tabulated in 2), regression analysis is performed with the price $Y'_{s,a,b}$ as the explained variable, and the reservation website, accommodation date, and accommodation facility as the explanatory variables (dummy variables)[56].

[56] By performing regression analysis using the data set for two consecutive months, prices newly collected in the current month, as well as fluctuations in the average price attributed to the monthly entry and exit of accommodation facilities that no longer accept reservations from the current month, can also be adjusted collectively by the same regression coefficient.

$$Y'_{s,a,b} = \alpha + \boldsymbol{\beta}_s \cdot \boldsymbol{x}_s + \boldsymbol{\beta}_a \cdot \boldsymbol{x}_a + \boldsymbol{\beta}_b \cdot \boldsymbol{x}_b + \varepsilon$$

Explanatory variables

Reservation website: $\boldsymbol{x}_s = (x_{s,1}, \cdots, x_{s,S-1})$ S: Number of reservation websites

Accommodation date: $\boldsymbol{x}_a = (x_{a,1}, \cdots, x_{a,A-1})$ A: Total number of days in the previous month and the current month

Accommodation facility: $\boldsymbol{x}_b = (x_{b,1}, \cdots, x_{b,B-1})$ B: Number of accommodation facilities

(b) For the combination of the reservation website, accommodation date, and accommodation facility that has become a missing value, the attribute information thereof (reservation website: $\boldsymbol{x}_s{}'$, accommodation date: $\boldsymbol{x}_a{}'$, accommodation facility: $\boldsymbol{x}_b{}'$) is used to calculate the estimated price value $\widehat{y_{\text{mis}}}$ based on the estimated regression model, which is substituted as a complement value.

$$\widehat{y_{\text{mis}}} = \hat{\alpha} + \widehat{\boldsymbol{\beta}_s} \cdot \boldsymbol{x}_s{}' + \widehat{\boldsymbol{\beta}_a} \cdot \boldsymbol{x}_a{}' + \widehat{\boldsymbol{\beta}_b} \cdot \boldsymbol{x}_b{}'$$

4) From the complemented data set, the geometric mean price is calculated for the current month (t) and the previous month ($t-1$) respectively. By multiplying the ratio of these prices by the price index of the previous month, the price index of the current month is calculated.

$$P_t = \left(\prod\nolimits_{s,a,b} P_{t,s,a,b}\right)^{\frac{1}{N_t}} = \exp\left[\frac{1}{N_t}\sum\nolimits_{s,a,b} \log\left(P_{t,s,a,b}\right)\right]$$
$$= \exp\left[\frac{1}{N_t}\sum\nolimits_{s,a,b} Y'_{t,s,a,b}\right]$$
$$I_t = I_{t-1} \times \frac{P_t}{P_{t-1}}$$

3 Charges for package tours to overseas

(1) Reservation time and tour plan

The prices of tour plans (package tours with transportation and accommodations only) that do not include a sightseeing tour are adopted for all days of each month in line with the plan release time at travel agencies and the purchase time of consumers**.

(2) Travel destination

Destinations are selected from representative regions (Asia, North America, Europe, and Oceania) and cities with a large number of Japanese visitors. In addition, since charges for package tours are quite susceptible to regional situations, two or more cities are selected in principle from each region in order to grasp monthly price trends as stably as possible.

(3) Travel agency

Based on the transaction volume of travels handled by major travel agencies, travel agencies with a high share of transaction volume are preferentially selected.

(4) Airline and accommodation facility, etc.

As for the airlines that provide transportation services and for the grades of accommodation facilities, those that are stably supplied in each region and handled in large quantities are selected based on the statuses of travel handled by each travel agency.

(5) Method of index calculation

Price indices are calculated by following the steps 1) to 3) below.

1) By simply averaging by the number of tour plans, the average prices are calculated for each travel destination, travel agency, and departure date.

$$P_{t,a,b,c} = \frac{\sum_d P_{t,a,b,c,d}}{n_{t,a,b,c}}$$

t: Comparison period, 0: Base period, *a*: Travel destination, *b*: Travel agency, *c*: Departure date, *d*: Tour plan

2) By simply averaging by the number of days in the current month, the monthly average price is calculated for each travel destination and travel agency. In addition, by dividing by the price in the base period, the index for each travel destination and travel agency is calculated.

$$P_{t,a,b} = \frac{\sum_c P_{t,a,b,c}}{n_{t,a,b}}$$

$$I_{t,a,b} = \frac{P_{t,a,b}}{P_{0,a,b}} \times 100$$

t: Comparison period, 0: Base period, *a*: Travel destination, *b*: Travel agency, *c*: Departure date

3) The weighted average is calculated by using the transaction volume ratio by travel agency ($w_{0,a,b}$) at each travel destination, and the index for each travel destination is calculated. Finally, the price index is calculated by averaging with weights using the transaction volume ratio by travel destination ($w_{0,a}$).

$$I_{t,a} = \frac{\sum_b I_{t,a,b} w_{0,a,b}}{\sum_b w_{0,a,b}}$$

$$I_t = \frac{\sum_a I_{t,a} w_{0,a}}{\sum_a w_{0,a}}$$

t: Comparison period, 0: Base period, *a*: Travel destination, *b*: Travel agency