Expanding the use of Big Data for CPI in Japan

Seitaro Tanimichi, Takuya Shibata
Statistics Bureau of Japan
May 2023

Prepared for the Meeting of the Group of Experts on Consumer Price Indices
UNECE, June 2023, Geneva

Summary

The Statistics Bureau of Japan (SBJ) has been utilizing big data to calculate the consumer price index (CPI) and has greatly expanded the scope since the 2020-base year.

In the 2015-base year, the index was calculated using scanner data for four items: “personal computers (laptop)”, “personal computers (desktop)”, “tablet computers” and “cameras”. From the 2020-base, three items, “video recorders”, “PC printers” and “TV sets” were added to the index using scanner data.

The SBJ has been conducting experimental studies and pilot tests for the use of web scraping since 2015, and from the 2020-base, began actually producing indices for travel services (“airplane fares”, “hotel charges” and “charges for package tours to overseas”).

By expanding coverage, the use of big data has made it possible to produce more appropriate indices, with the number of prices increased significantly compared to previous field surveys, and to reduce the burden on local governments and price collectors.

This paper introduces a comparison of the 2020-base results using big data and the 2015-base results using field surveys for the same items, as well as the current status of studies aimed at expanding the use of big data.

1. Introduction

In the 2020-base revision of CPIs in Japan, the use of scanner data was expanded and internet sales prices by web scraping were newly adopted. In order to expand the use of big data, in light of the increase in online shopping in recent years and the development of information-gathering technology, around 2015 the SBJ started specific studies on the use of scanner data and the collection of online sales prices by web scraping. For the items to be adopted, we narrowed down the candidates by comparing the index created from the trial collection data with the current index and the percentage of online purchases. As a result, it was decided to expand the use of scanner data in recreational durable goods, and for travel services (airplane fares, hotel charges, and charges for package tours to overseas), and to shift from previous price surveys to collection of online sales prices using web scraping.

In addition to confirming that there were no legal problems such as copyright with web scraping, we requested the cooperation of site operators, improved the collection timing, and began operation in January 2020. Since August 2021, the SBJ has published indices calculated by expanding the use of such big data.

In this paper, we present the verification of the production of indices for items by using big data in the 2020-base and the status of studies toward the further use of big data in the 2025-base.

History of expanded use of big data in base revision of CPI

2000-base	Used scanner data for “personal computers (desktop)” and “personal computers (laptop)”
2005-base	Added scanner data for “cameras”
2010-base	Included the price by scanner data of “tablet computers” to “personal computers (laptop)”
2015-base	Separated “tablets computers” from “personal computers (laptop)”
2020-base	Used scanner data for “video recorders”, “PC printers” and “TV sets” Used web scraping data for “airplane fares”, “hotel charges” and “charges for package tours to overseas”

2. Details of studies and calculation methods of price indices using big data

(1) Use of web scraping data: example of “hotel charges”

In considering the use of web scraping for hotel charges, we conducted a questionnaire survey to examine trends in purchasing methods, time to make reservations, accommodation plans, selection of collection sites, etc. We also conducted price collection and index production by web scraping on a trial basis, and compared it with the index by conventional price surveys. As a result,

- The largest number of reservations were made via the Internet, and capturing the price trend of internet sales appropriately grasped the price trend of hotel charges.
- We confirmed that web scraping can stably collect internet reservation prices from each travel booking website.
- We had a prospect of a huge number of internet sales prices being accurately reflected in the indices, including quality adjustment, and it is expected that web scraping collecting daily prices contributes to the improvement of accuracy of indices.

Therefore, we decided to use the internet sales prices.

(Price collection sites)

According to the questionnaire results, the largest number of people used travel booking websites rather than websites of hotels. So, based on the status of the transaction volume handled by major travel agencies, travel agencies of booking websites with the highest share of the transaction volume are selected for web scraping collection of prices. In addition, as web scraping requires individual settings based on each website structure, it is practical and efficient to collect from a comprehensive booking site, in which we can collect many prices from the same site.

Table 1 : Reservation time and method (results of the questionnaire)

N = 2,448		RESERVATION TIME				
		Within a week	One to three weeks before	One month or more before	Unknown	Total
RESERVATION METHOD	Called hotels directly	3%	4%	5%	1%	13%
	Website of hotels	2%	7%	12%	1%	21%
	Travel booking site	7%	21%	29%	2%	59%
	Over the counter	0%	1%	2%	0%	3%
	Others	0%	0%	1%	0%	1%
	Unknown	0%	0%	1%	2%	3%
	Total	12%	33%	50%	6%	100%

(Accommodation plans and price collection time)

Depending on the release timing of accommodation plans at travel agencies and the timing of consumers' purchases, daily prices in each month of *ryokan* (Japanese-style inns), Japanese-style rooms, of one night with two meals plans and of hotels, Western-style rooms, of one night with breakfast are used. Plans with extremely high (or extremely low owing to a sale) prices relative to typical hotel charges are excluded during process of excluding outliers.

Table 2 : Cross table of room types and meal types (results of the questionnaire)

N = 2,448	WESTERN-STYLE ROOMS	JAPANESE-STYLE ROOMS	JAPANESE-WESTERN STYLE ROOMS	OTHERS	TOTAL
NO MEALS	24%	4%	1%	1%	29%
WITH BREAKFAST	24%	3%	1%	0%	29%
WITH BREAKFAST AND DINNER	11%	22%	7%	0%	40%
BREAKFAST, LUNCH AND DINNER INCLUDED	1%	1%	0%	0%	2%
OTHERS	0%	0%	0%	0%	0%
TOTAL	60%	30%	9%	1%	100%

As for price collection time, in principle, prices are collected at the beginning of the month, two months before the accommodation date. This is because, in the web scraping collection results obtained during the pilot study, the collection results one month before the accommodation date of some sites showed that the average price of some accommodations was abnormally high compared to that of the two-month prior collection due to the inability to collect low-priced plans because of full occupancy.

In addition, according to the results of long-term web scraping conducted between August 2017 and March 2018, limited to 30 accommodation facilities, the following trends were observed in the number of facilities where prices could be collected, and it was also found that there was a seasonal limit on advanced reservation. (Table 3)

- Prices for about 10% of accommodations four months ahead and about half of accommodations six months ahead were not listed on the booking site. Therefore, it was not possible to collect prices.
- Especially before November, prices from the following April (shaded cells) are posted considerably less than before, and there is a gap in the status of prices posted on the site at the time of change of the fiscal year.

Table 3: Number of accommodation facilities capable of price collection (N = 30)

	Reservation month										
Collection month	1 month ahead	2 month ahead	3 month ahead	4 month ahead	5 month ahead	6 month ahead	7 month ahead	8 month ahead	9 month ahead	10 month ahead	11 month ahead
2017 Aug	30	29	29	28	25	18	14	2	2	2	1
Sep	30	30	29	26	23	16	4	2	2	1	1
Oct	30	30	30	27	22	7	3	2	1	1	1
Nov	30	30	29	26	17	10	5	4	2	2	1
Dec	30	29	28	24	22	14	7	5	5	3	3
2018 Jan	29	29	27	26	26	14	9	6	5	5	5
Feb	29	28	28	27	26	18	12	5	5	5	3
Mar	29	29	28	27	26	17	10	6	6	3	2
Average	30	29	29	26	23	14	8	4	4	3	2
Collection percentage	100%	99%	96%	89%	79%	48%	27%	14%	12%	9%	7%

(Accommodation facilities)

Based on the number of guests and facility scale of capacity by travel destination (prefecture) in the Overnight Travel Statistics Survey (official statistics by Japan Tourism Agency), about 400 representative accommodations are selected.

Price collection by web scraping does not require consideration of the upper limit of the number of target facilities caused by resource constraints. However, unrestricted access to websites to obtain Internet sales prices is not possible in light of the load on the site. Therefore, it is necessary to set an appropriate number of target facilities.

In the pilot study, the standard error rate of the geometric average price was calculated using the experimentally collected data table, and the effect on the price index was taken into account. As a result, the number of facilities was set at 400, since the standard error rate for the increase in the number of facilities almost stopped decreasing and leveled off when the number of facilities exceeded 400.

(Calculation method of indices)

Using a two-month data set for the current month (t) and the previous month ($t-1$), the price indices are calculated according to the following procedures (1) to (4).

(1) Exclusions of outliers

In price collection, as all plans that match the conditions are collected, extremely high or low prices may be collected. Plans in such price range have large quality differences from other prices and may have temporarily lower prices, such as with a limited-time sale. Thus it is considered appropriate to exclude them as outliers when producing price indices. Therefore, the following procedure is adopted to exclude outliers.

(a) Define the individual prices as $P_{s,a,b,c}$ by booking website (s), by accommodation date (a), by accommodation facility (b) and by plan (c), and convert them to logarithms.

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

(b) Calculate average prices and standard deviations by booking 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) Any individual price that differs from the average price by more than three times the absolute value of the standard deviation for each reservation site, accommodation date and accommodation facility is considered as an outlier.

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

(2) Creation of a data table

For individual prices excluding outliers, average prices for each booking website, accommodation date, and accommodation facility are calculated, and a data table with these as attributions is created ($N'_{s,a,b}$ is the number of 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 value imputation

In the case of the average price after data cleaning, if the individual prices are not displayed on the site as a result of the site search by setting the reservation date and accommodation, the average value under this search condition cannot be calculated, which causes missing values in the data table. In the calculation of the average price in which missing values are ignored in the index calculation, the difference in missing by day of the week may make missing less random, resulting in a bias in the average price. In addition, attention should be paid to the imputation at the calculation stage of the average price because the result of the index calculation may change depending on the calculation order of the average. Therefore, a method of estimating and imputing missing values from regression analysis of data sets of actual measured values (regression imputation) is considered.

As the index calculation assumes a monthly chain-linking method, by performing regression analysis using a data set for two consecutive months, the same regression coefficient can be used to adjust the average price variation due to the entry and exit of accommodations on a monthly basis together, such as newly collected in the current month or those that no longer accept reservations from the current month.

(a) Using the data table aggregated in (2), regression analysis is performed with the price $Y'_{s,a,b}$ as an explained variable and reservation site, accommodation date, and accommodation facility as explanatory variables (dummy variables).

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

Explanatory variable

Reservation site: $\boldsymbol{x}_s = (x_{s,1}, \cdots, x_{s,S-1})$ S: The number of booking websites

Accommodation date: $\boldsymbol{x}_a = (x_{a,1}, \cdots, x_{a,A-1})$

A: Total number of days in the current month and the previous month

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

B: The number of accommodation facilities

(b) Based on the estimated regression model, in the combinations of booking website, accommodation date, and accommodation facility that lead to missing values of prices, estimate values of prices $\widehat{y_{\text{mis}}}$ are calculated using the attribution information (booking website: $\boldsymbol{x}_s'$, accommodation date: $\boldsymbol{x}_a'$, accommodation facility: $\boldsymbol{x}_b'$) and are substituted as imputed values.

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

(4) The data set after imputation is used to calculate the geometric average prices for the current month (t) and the previous month ($t-1$), respectively. These price relatives are multiplied by the price index for the previous month to calculate the price index for the current month.

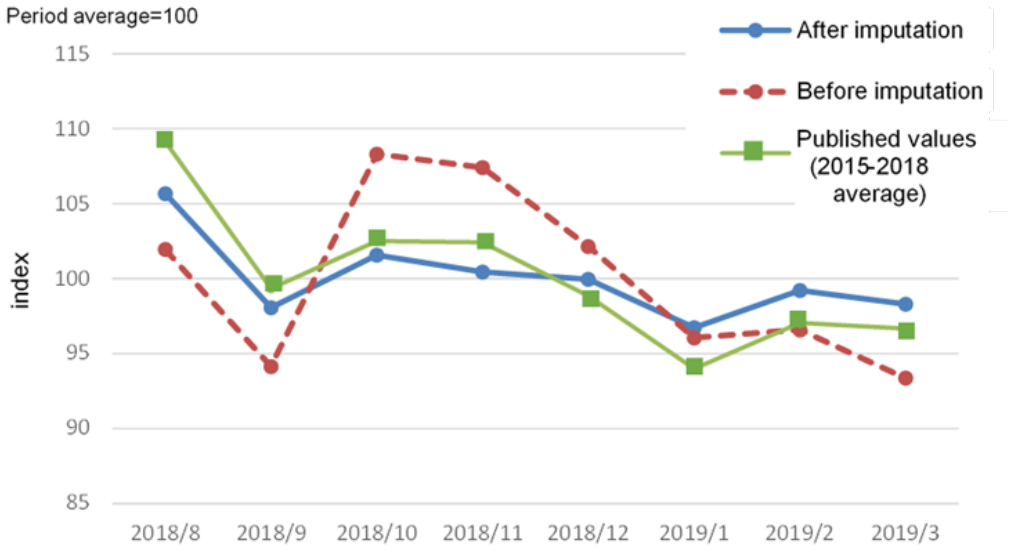
$$P_t = \left(\prod_{s,a,b} P_{t,s,a,b}\right)^{\frac{1}{N_t}} = \exp\left[\frac{1}{N_t}\sum_{s,a,b} \log\left(P_{t,s,a,b}\right)\right]$$

$$= \exp\left[\frac{1}{N_t}\sum_{s,a,b} Y'_{t,s,a,b}\right]$$

$$I_t = I_{t-1} \times \frac{P_t}{P_{t-1}}$$

Figure 1 shows the calculation results of the verification. By imputing missing values, it can be seen that the index has remained stable by the effect of adjusting the difference in month-by-month average prices due to differences in facilities. To examine seasonality, we compared the index after imputation with the average value for four years of published values from 2015 to 2018 and found that the index after imputation generally captured seasonal movements. In addition, the index in August was lower than the published value because the published value in 2018 largely increased owing to the effect of a calendar, but reflecting daily prices by web scraping removes the temporary effect of the relationship between survey date and a calendar. Conversely, the indices in December and January were higher than the published values, but this divergence was caused by the fact that the published values did not reflect prices during the busy period of year-end and New Year holidays, while the calculation values did. Thus calculation results are considered to reflect the actual condition.

Figure 1: Index calculation results



(2) Use of scanner data: examples of "TV sets"

Until the 2015-base, the price index of "TV sets" for the CPIs was calculated using prices collected through the specification designation method in the Retail Price Survey. However, while high-quality TVs with higher resolution and larger screens are becoming more prevalent, there is demand for conventional TVs due to the increasing number of single-person households and other factors, leading to greater diversification. To reflect these trends in the indices, we examined index creation using the hedonics method, which utilizes scanner data, as a method to create indices that do not rely on the specification designation method.

The following scanner data were used in the validation for the 2020-base revision.

- Period: Monthly data from October 2017 to March 2018
- Type: Liquid crystal display TV (not including organic EL TV)
- Region: Whole of country (about 2,500 outlets), including online shops
- Data size: Approximately 750 models, Unit sales: Approximately 220,000/month average
- Average unit price and sales quantities by model (total of outlet sales and online sales)
- Characteristics of each model, such as screen size and number of pixels displayed

Specifications	Examples
Release month	Year, Month
Tuner shape	Separate type, Integrated type, None
Screen size	3-inch type to 75-inch type
Number of pixels displayed	1366x768, 1920x1080, 3840x2160, etc.
D connector	D4x1, D5x1, None
PC input	D-Sub, None
Communication terminal	LAN, None
Card slot	SDXC, None
HDD capacity	0 GB to 2,000 GB

Internet	Capable, Incapable
Wireless function	IEEE802.11a/n, None
Audio output	10W+10W, 3W+3W, 5W+5W, etc.
HDMI connector	0 to 4
Link function	Available, Unavailable
Drive speed	Constant speed, Double speed
Recording media	HDD (external), HDD (internal/external)
High-definition capable	4K/2K, 8K, High-definition, Full high-definition, Incapable
Hybrid cast	Capable, Incapable

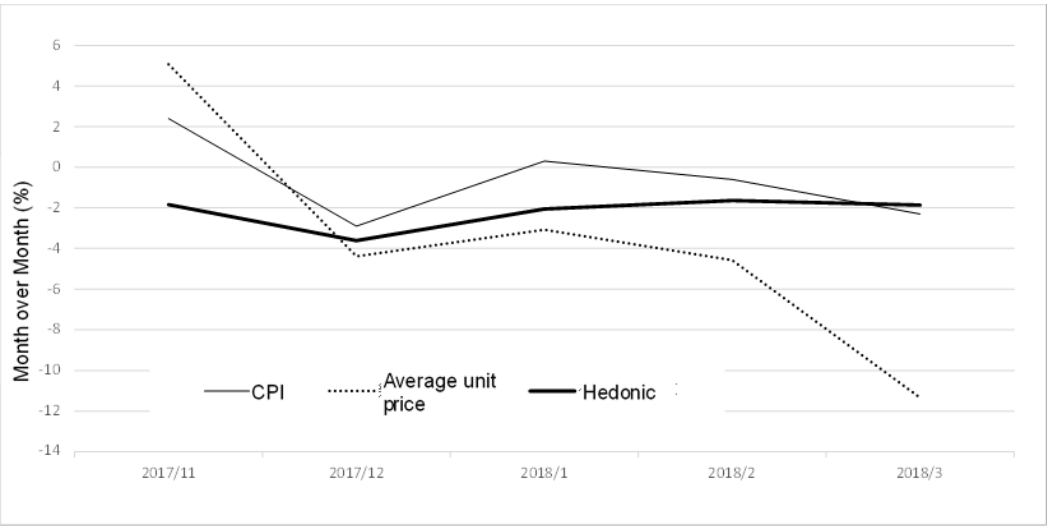
In terms of the product cycle, when observing the market share by release month from the scanner data as of March 2018, product models released in September 2017 still held about 30% of the market share in March 2018, more than half a year after launch, while models released within one year of launch held about 80%, those within one year and a half held 90%, and those within two years held almost 100%. In time series, the share of models released within a year and a half ranged from 80% to 90%, and the share of models released within two years transitioned at 95% or more, indicating that the product cycle is short compared to the frequency of base revisions of CPI (five years). It is conceivable that a long period of time after launch may result in a significant difference in quality from the new model, or a price drop greater than the difference in quality. For this reason, models after 24 months have passed since the launch are excluded from the analysis.

The regression model is set up as a semi-logarithmic regression model with the average unit price as an explained variable and with various characteristics such as specifications as explanatory variables. The explanatory variables were selected by the stepwise method from the characteristic values using scanner data of March 2018. For the month-over-month estimation, data from two consecutive months are pooled and analyzed using a regression model weighted by sales quantities to estimate the price relative between the two time points of which quality differences were adjusted.

As a result of the estimation, the result of the month-over-month estimation between November 2017 and March 2018 showed that the adjusted coefficient of determination adjusted for degrees of freedom remained stable over 0.95 in all the periods, indicating that its applicability to the hedonic regression model is good.

Figure 2 shows a comparison between the 2015-base CPI and the results of the month-over-month provisional calculation by the hedonic price index. Although there are differences in product models and price levels between the current CPI based on the specification designation method and the hedonic price index based on scanner data, the month-over-month provisional calculation values based on the hedonic price index show a difference of 0.4 to 4.7 points from the current CPI. As a result of the calculation, it was thought that the hedonic regression model using scanner data would enable stable quality adjustment and contribute to improving the accuracy of statistics, and therefore scanner data was used for TV sets in the 2020-base revision.

Figure 2: Comparison of the 2015-base CPI and calculation values



For PC printers and video recorders, a fixed-specification method is used, not a hedonic regression model. This is based on the following characteristics: these items have a long cycle of new products, the items have little difference in quality between the old and new products, the price of the items can be explained with small specifications, and the items have small weights.

3. Comparison of results using big data (the 2020-base) with results from field surveys (the 2015-base)

(1) Web scraping

For items using web scraping from the 2020-base, price collection conditions and the number of collected prices were compared with those of the 2015-base as shown in the table below, and the number of collected prices has increased significantly.

Item	Hotel charges	
Base	2015 Base	2020 Base
Collection conditions (main)	Prices on Friday and Saturday of the week including the 5th of every month	Prices of 1st to 31st of every month purchased two months in advance of accommodation
Number of collected prices	640	About 1 million

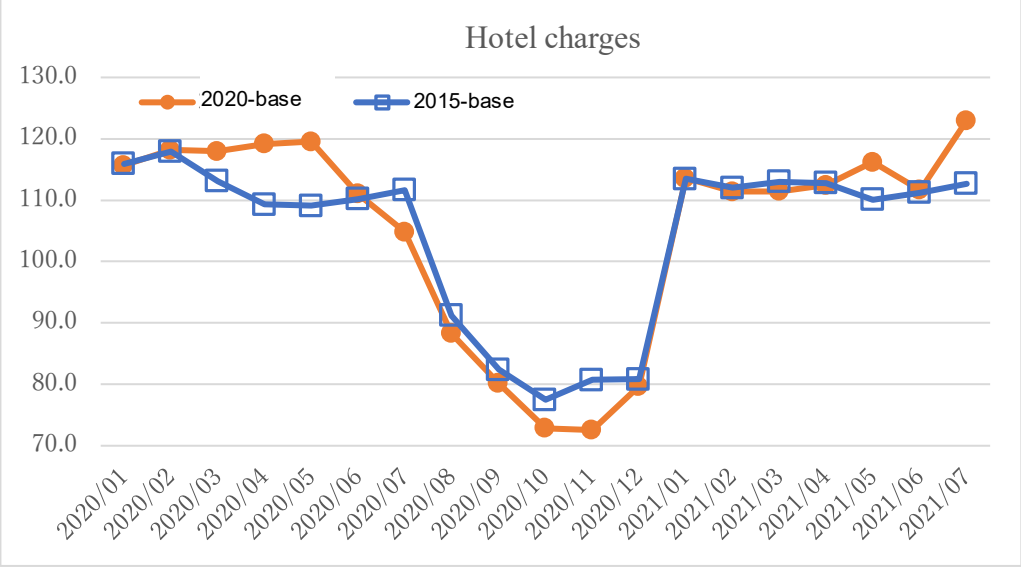
Item	Airplane fares	
Base	2015 Base	2020 Base
Collection conditions (main)	One flight each by adopted section and airline	All flights by adopted section and airline
Number of collected prices	2,604	About 2.5 million

Item	Charges for package tours to overseas	
Base	2015 Base	2020 Base
Collection conditions (main)	One flight by adopted city and travel company	All flights by adopted city and travel company
Number of collected prices	372	About 200,000

With regard to hotel charges, from January 2020 to July 2021, a comparison of the price index in the 2020-base for these items with the price index in the 2015-base (converted value as 2020 year = 100) yielded the following results.

The 2015-base index has fallen sharply in August 2020. On the other hand, the 2020-base index over the same period has been somewhat gradual compared to the 2015-base index. This is because the impact of the government's travel assistance program (reduction of hotel charges), which began in late July, was reflected from July in the 2020-base index, whereas the index of 2015, which only covered prices for a specific two days in early every month, did not show the impact of the program in July but reflected it from the following August. Web scraping has made it possible for policy effects to be reflected in the index in a timely manner.

In addition, the difference in the movements of the two indices from November to December 2020 may also be affected by the difference in the scope of accommodation dates covered and the timing of price collection. In the index of 2015, which only covers prices for a specific two days, the calendar around the survey date has affected the indices, but the introduction of web scraping has made it possible to cover all days of accommodation, which has made it possible to produce more stable indices.



With regard to travel services to which web scraping is introduced, it has become possible to produce more stable and appropriate indices by expanding coverage in general. "Hotel charges" were excluded from the price collection surveys conducted, which contributed to reducing the burden on collectors and local government officials.

(2) Scanner data

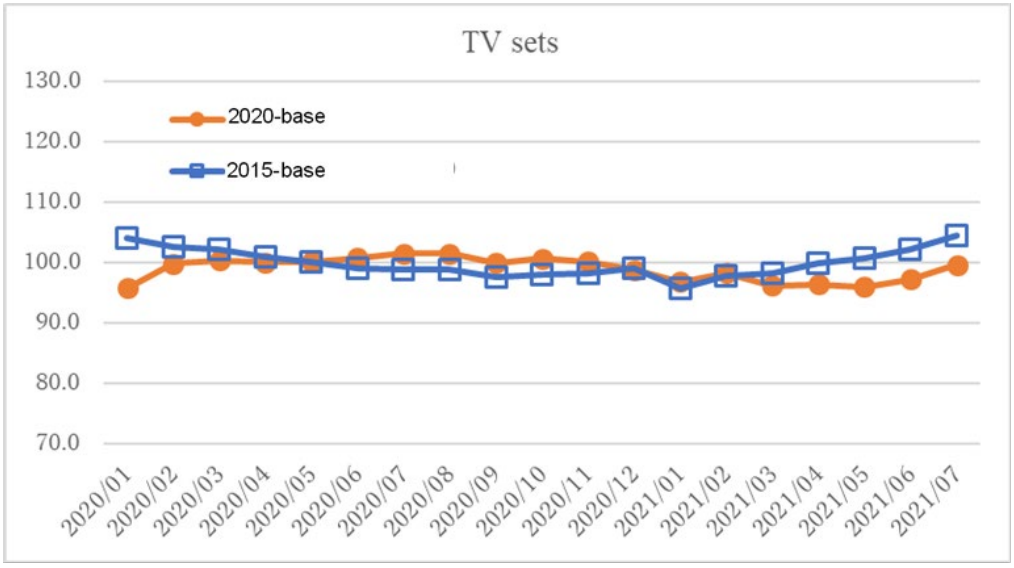
The table below shows the comparison of collection time of prices and the number of collected prices for items that use scanner data from the 2020-base with those in the 2015-base, and that the number of collected prices considerably increased.

	2015 Base			2020 Base		
Collection time and price	Price on any one of Wednesday, Thursday or Friday of the week including the 12th of each month			Prices from 1st to 31st of each month		
Item	Video recorders	PC printers	TV sets	Video recorders	PC printers	TV sets
Number of collected product models	6	1	8	23	46	600
Number of stores for collection	186	172	186	About 2,600	About 2,600	About 2,600
Number of collected prices	186	172	186	About 30,000	About 80,000	About 240,000

When comparing the price index in the 2020-base for these items with the price index in the 2015-base (converted value as 2020 year = 100) from January 2020 to July 2021, the following results were obtained.

- TV sets (hedonics method)

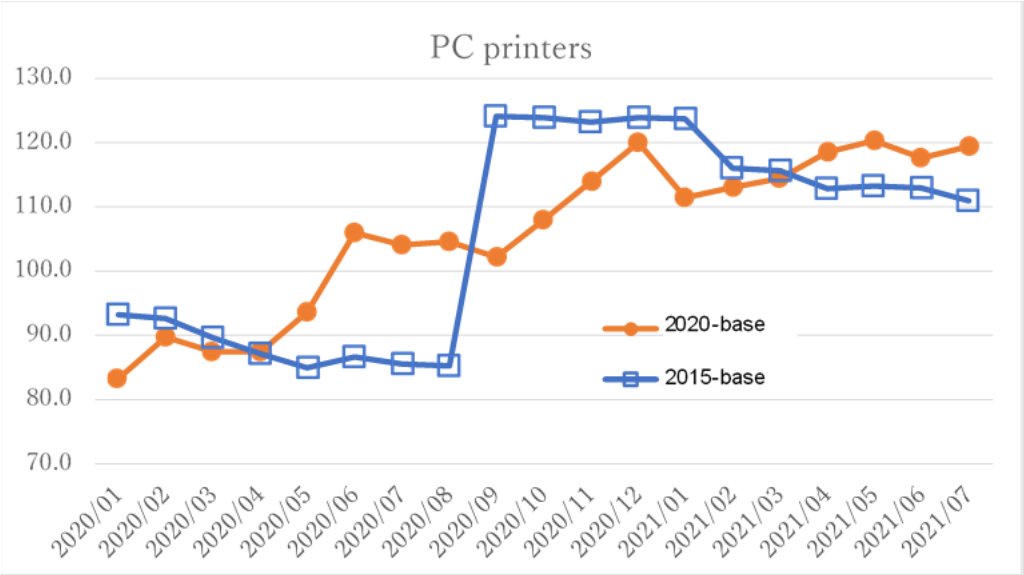
  While the prices of some specific product models are collected for the index of 2015, the 2020-base index covers all models (including online sales) included in the scanner data, so that the price trend after quality adjustment can be captured by the specification information. Specifically, the 2015-base index shows a downward trend from the spring of 2020 until the end of the year, while the 2020-base index shows an upward trend. The movement of the 2020-base index is also in line with the presumption that demand for televisions at home increased during this period, along with increased time at home.



- PC printers (fixed specification method)

  As the 2015-base index collects the price of one specific product model, the index changes depending solely on the model whose price increased in September 2020. On the other hand, the 2020-base index can capture models whose prices have increased since around May 2020 because multiple

models that fell under the selected specifications (including online sales) are included. Specifically, the movement of the 2020-base index is consistent with the presumption that since the spring of 2020, the demand for PC printers at home increased owing to the spread of remote working and classes to prevent the spread of COVID-19.



Based on the above, we believe that more appropriate index production has become possible for recreational durable goods for which scanner data is newly used by the expansion of coverage and quality adjustment using specification information. In addition, items for which the survey method was switched to price collection by scanner data are excluded from the scope of surveys by enumerators, and this contributes to reducing the burden on prefectures and enumerators.

4. Study to expand the use of big data

In light of the expansion of online sales, improvement of information-gathering technology, and further deterioration of the field survey environment, it is necessary to accelerate the use of big data for the CPI. Therefore, we will continue to study to make use of big data. In doing so, it is necessary to take into consideration newly occurring costs and issues, as well as the division of roles between field collection and prefectural surveys, and to prioritize areas that are expected to be cost-effective against budgetary constraints.

The items under consideration include white goods, foods, medical supplies, daily necessities and clothing. Of these, data for some items of white goods have already been shifted to scanner data, but it is expected that the extension to electric rice-cookers and microwave ovens will contribute to reducing the field survey burden on enumerators in the future. Scanner data is also expected to be used for food, medical supplies and daily necessities. On the other hand, in the case of foods, for example, there is no scanner data for prepared food. Therefore, the use of scanner data for some items may not substantially reduce the burden on enumerators.

For clothing, we are considering web scraping to collect prices for items such as one-piece dresses, slacks and children's trousers, in light of the growing size of the online sales market and the percentage of

purchases. As web scraping data for clothing contains a large number of related products in addition to the clothing being sought, it is necessary to extract equivalent products from these products, but since the necessary codes and names are often not present, it is difficult to filter them mechanically and it is not practical to extract them manually. Therefore, we are currently studying the construction of a machine learning model that automatically classifies products into equivalent products based on product descriptions (about 100 to 400 words) and image information.

To date, as for analyses using text information, we are verifying methods such as logistic regression, gradient boosting (Light GBM), and kernel SVM as models for classifying materials (cotton, chemical fiber, etc.), lengths (full length, short, etc.), seasons (spring/summer, fall/winter, etc.), and patterns (plain, floral, etc.). We are also verifying methods for analysis using image information such as ResNet and EfficientNet.

Although these methods can ensure a certain level of classification accuracy, practical applications require reducing the amount of images and shortening the computation time because of the large data capacity of images, and increasing the number of companies targeted for web scraping to secure a share of sales.

5. Conclusion

This paper introduced the expansion of the use of big data in the 2020-base revision. The use of big data has contributed to improving statistical accuracy by expanding coverage and reducing the burden on prefectures and enumerators. We will continue to conduct wide-ranging studies for accuracy improvement of the CPI and efficient price collection.