Statistics Japan

Expanding the use of Big Data for CPI in Japan

Seitaro Tanimichi, Takuya Shibata

Statistics Bureau, Japan

Meeting of the Group of Experts on Consumer Price Indices, 7-9 June 2023, Geneva, Switzerland

Outline

- Background
- Web Scraping data: hotel charge
- Scanner data
- Study for further expansion

Background

2000-base: Scanner data for “desktop computers” and “laptop computers”

2005-base: Added scanner data for “cameras”

2010-base: Included scanner data of “tablet computers” to “laptop computers”

2015-base: Separated “tablets computers” from “laptop computers”

2020-base: Web scraping data for “hotel charges”
“airplane fares”
“charges for package tours to overseas”

Scanner data for “video recorders”, “PC printers” and “TV sets”

Use of web Scraping data : hotel charges

- A questionnaire survey to examine
  - ✓ trends in purchasing methods,
  - ✓ time to make reservations,
  - ✓ accommodation plans,
  - ✓ selection of collection websites, etc.

Also

- Conducted price collection and index production by web scraping on a trial basis
- Compared with the index by conventional price surveys

- Capturing the price trend of internet sales grasped the price trend of hotel charges
- Web scraping can stably collect prices from each travel booking website
- A huge number of internet sales prices were accurately reflected in the indices

Web scraping contributes to the improvement of indices

Web Scraping (hotel charges) : Price collection sites

	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%

Web Scraping (hotel charges) : Accommodation plan

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%

Web Scraping (hotel charges) : Price collection time

- Prices are collected, in principle, at the beginning of the month, two months before the accommodation date

As for one month before the accommodation date,

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

Long-term web scraping conducted between August 2017 and March 2018

(for 30 accommodation facilities)

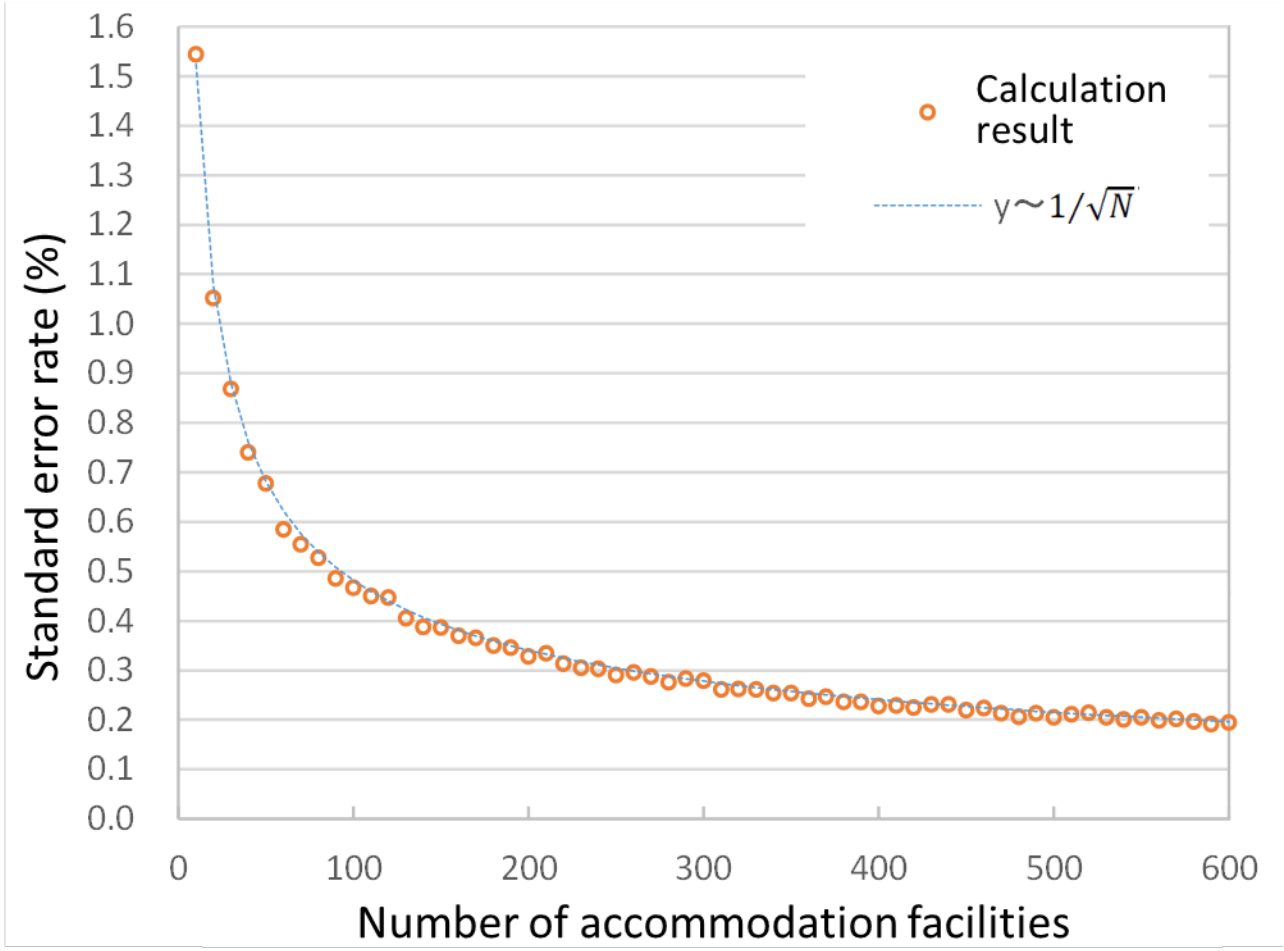
- Prices for about 10% of accommodations four months ahead and about half of accommodations six months ahead were not listed on the booking website
- Seasonal limit on the advanced reservation, a gap at the time of change of the fiscal year

Web Scraping (hotel charges) : Accommodation Facility

- About 400 representative accommodations facilities are selected

- While price collection by web scraping does not require consideration of the upper limit of the number of target facilities caused by resource constraints, unrestricted access to websites to obtain prices is not possible in light of the load on the website.
  ➔ It is necessary to set an appropriate number of target facilities.

- In the pilot study, the standard error rate of the average price for the increase in the number of facilities almost stopped decreasing and leveled off when the number of facilities exceeded 400



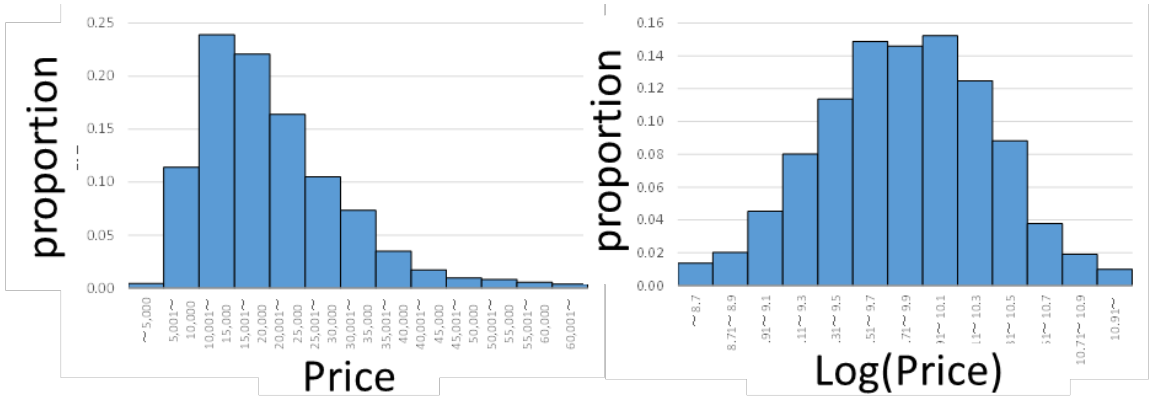
Web Scraping (hotel charges) : Calculation of indices

- 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

$P_{s,a,b,c}$ s: booking website,
a : accommodation date,
b : accommodation facility,
c : plan

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



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

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

➡ $Y_{s,a,b,c}$ is considered as an outlier if $\left|Y_{s,a,b,c} - Y_{s,a,b}\right| > 3\sigma_{s,a,b}$

Web Scraping (hotel charges) : Calculation of indices

(2) Creation of a data table

- Average prices for each booking website(s), accommodation date(a), and accommodation facility(b) are calculated,
- Data table with these as attributions is created

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

(3) Missing value imputation : Next Slide

(4) Calculation of index

- Data set after imputation is used to calculate 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.

$$P_t = \left(\prod_{s,a,b} P_{t,s,a,b}\right)^{\frac{1}{N_t}} = \exp\left[\frac{1}{N_t}\sum_{s,a,b}\log\left(P_{t,s,a,b}\right)\right] = \exp\left[\frac{1}{N_t}\sum_{s,a,b} Y'_{t,s,a,b}\right]$$

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

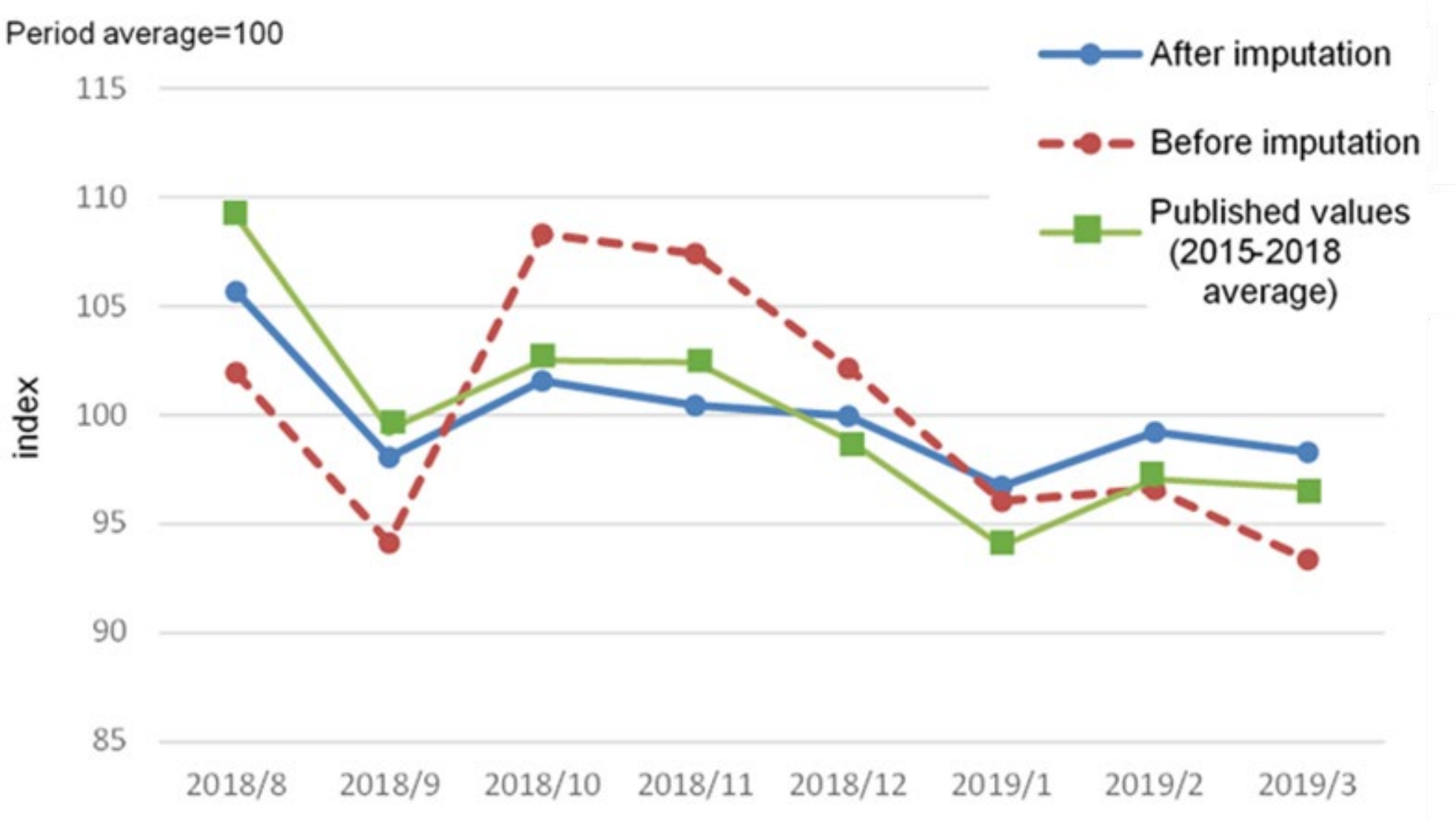
Web Scraping (hotel charges) : Missing value imputation

Accommodation date (X_a)	Booking site (X_s)	Facility (X_b)	Log Average Price (y)
2018/12/1	A	X	9.51
2018/12/1	A	Y	9.61
2018/12/1	A	Z	9.75
2018/12/1	B	X	
2018/12/1	B	Y	
2018/12/1	B	Z	
2018/12/1	C	X	9.58
2018/12/1	C	Y	9.69
2018/12/1	C	Z	9.85
2018/12/2	A	X	9.65
2018/12/2	A	Y	9.66
2018/12/2	A	Z	
2018/12/2	B	X	9.49
……	…	…	……

Accommodation date (X_a)	Booking site (X_s)	Facility (X_b)	Log Average Price(y)
2018/12/1	A	X	9.51
2018/12/1	A	Y	9.61
2018/12/1	A	Z	9.75
2018/12/1	C	X	9.58
2018/12/1	C	Y	9.69
2018/12/1	C	Z	9.85
2018/12/2	A	X	9.65
2018/12/2	A	Y	9.66
2018/12/2	B	X	9.49
……	…	…	……

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

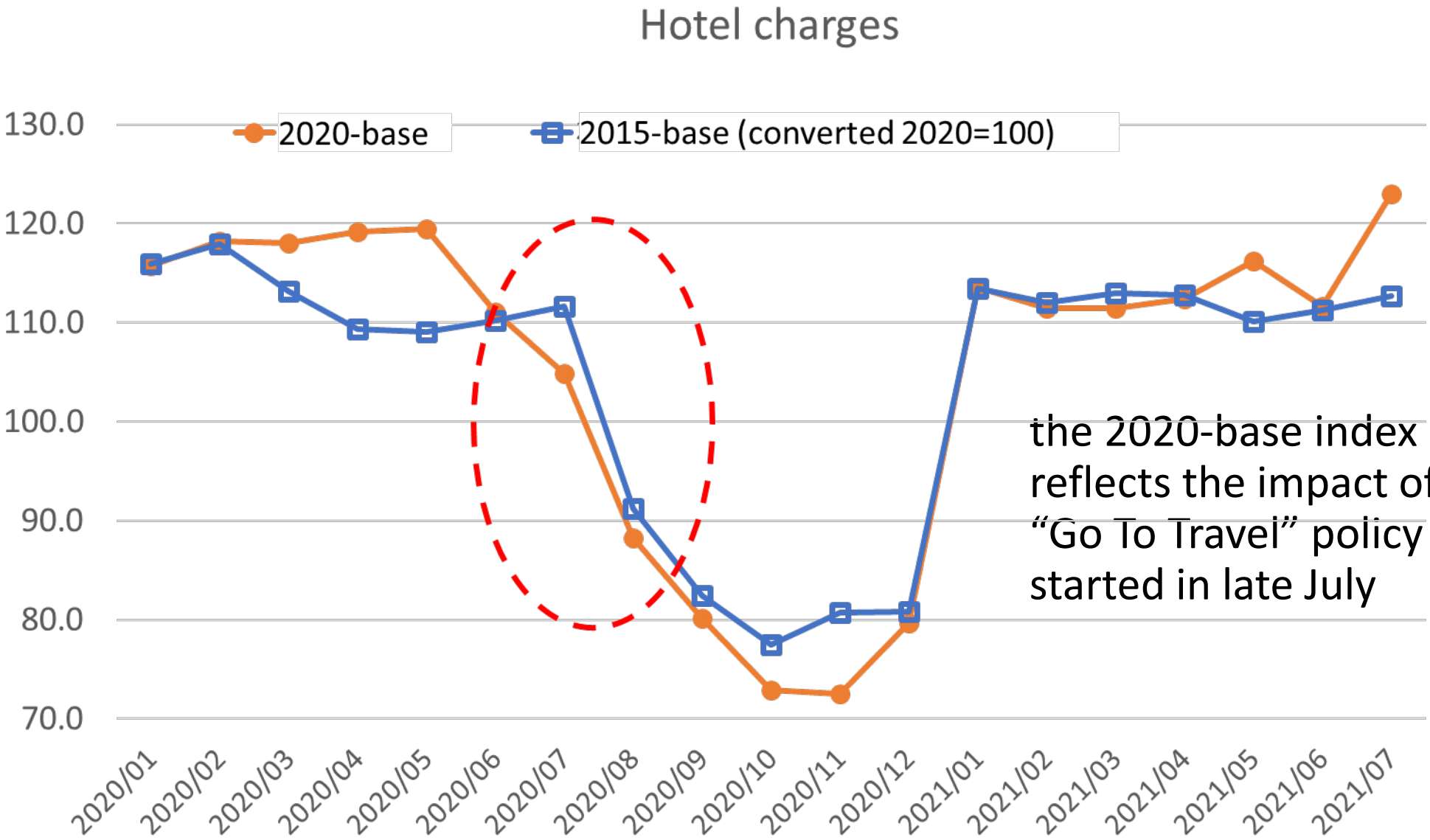
Use of web Scraping data : hotel charges



Use of web Scraping data : hotel charges

	2015-Base method (field collection)	2020-Base method (web scraping)
Collection conditions	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

Use of web Scraping data : hotel charges



the 2020-base index in July reflects the impact of the "Go To Travel" policy that started in late July

Use of Scanner data

- TV Sets
  ➔ hedonic model

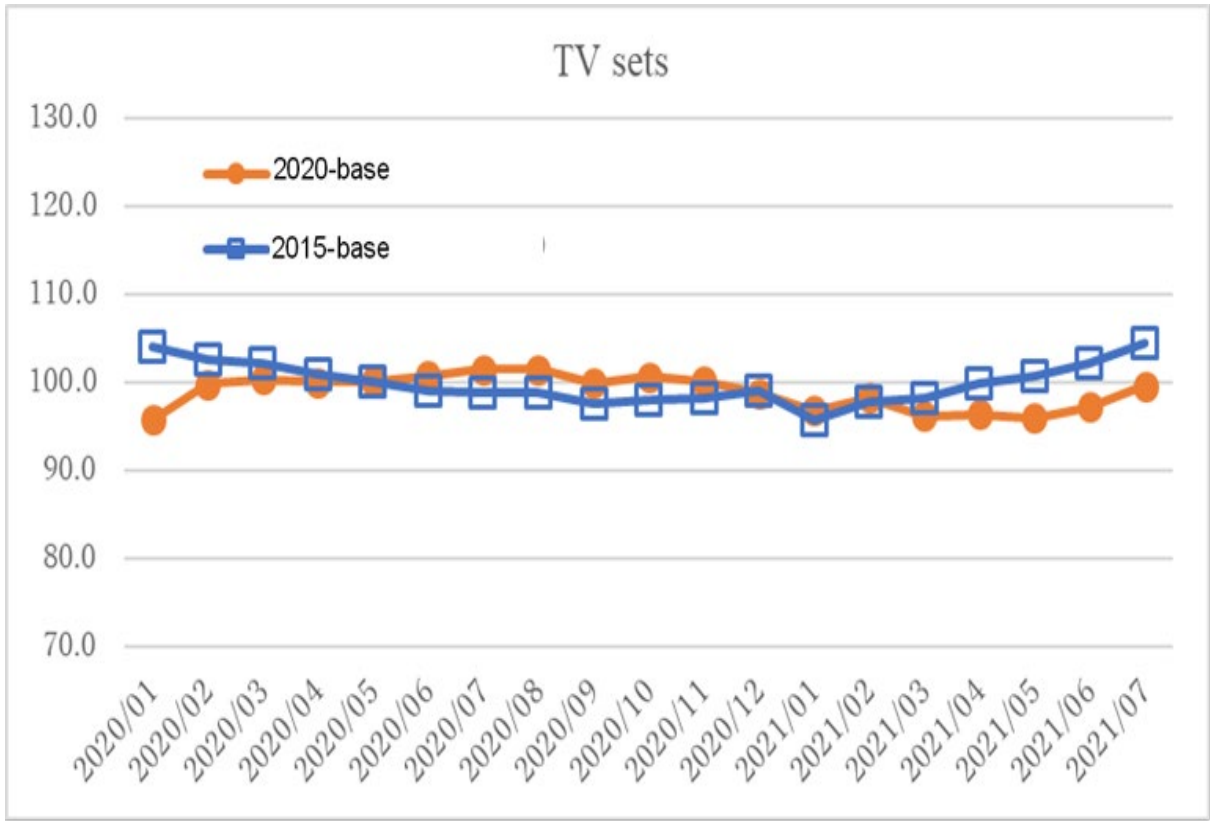
- PC printers, video recorders
  ➔ fixed-specification method

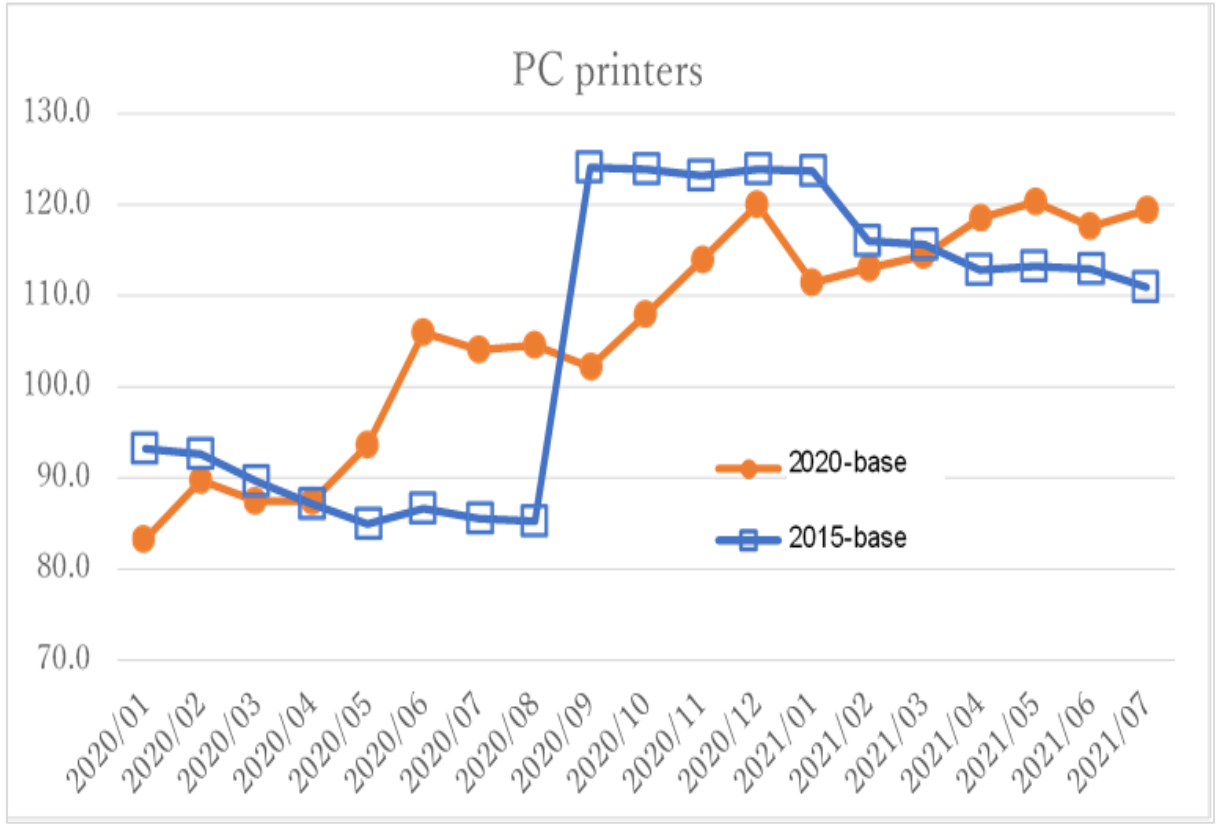
Specifications	Examples
Release month	Year, Month
Tuner shape	Separate type, Integrated type, None
Screen size	3-inch type to 75-inch type
Number of pixels displayed	1366x768, 1920x1080, 3840x2160, etc.
D connector	D4x1, D5x1, None
PC input	D-Sub, None
Communication terminal	LAN, None
Card slot	SDXC,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

Use of Scanner data

	2015 Base (field collection)			2020 Base (Scanner data)		
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

Use of Scanner data





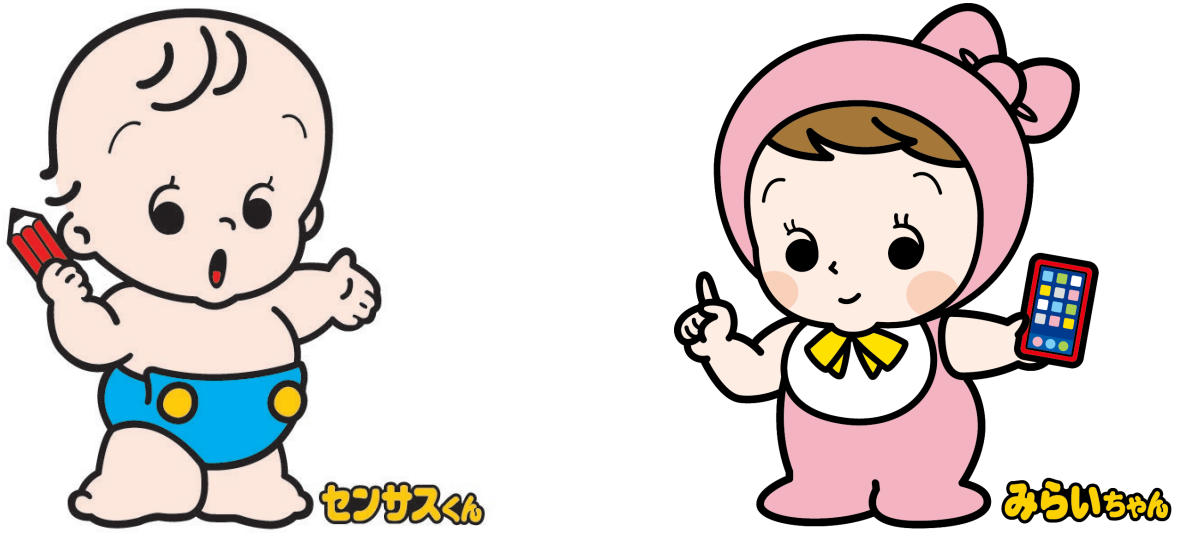
Study for further expansion of the use of big data

- It is necessary to accelerate the use of big data for the CPI
- The items under consideration include white goods, foods, medical supplies, daily necessities, and clothing

- For clothing, we are considering web scraping to collect prices for items such as one-piece dresses, slacks, and children's trousers
- As web scraping data for clothing contains a large number of related products, it is necessary to extract equivalent products from these products
- The necessary codes and names are often not present, it is difficult to filter them mechanically (and not practical to extract them manually)

➔Currently studying the construction of a machine learning model for automatically classifying products based on product descriptions (about 100 to 400 words) and image information

Thank you



mascot of Statistics Bureau of Japan

"*Census-Kun*" and "*Mirai-chan*"

(Master. Census) (Miss. Future)

Statistics Bureau of Japan

Seitaro Tanimichi
s.tanimichi@soumu.go.jp

Takuya Shibata
t5.shibata@soumu.go.jp