Improving the accuracy of the CPI by using big data in Japan

Hideaki Nakamura, Mone Imamura

Introduction

- Statistics Bureau of Japan (SBJ) has been calculating **the price indices of PCs since 2000 and those of digital camera since 2003, by hedonic method using scanner data**.
- To improve the accuracy of the CPI, SBJ is examining how to calculate the price indices by using big data in Japan. In this poster, we introduce some of the **tentative results of the trial calculation of the price indices of TV and mobile phone by using scanner data and those of airplane fares by using web-scraped data**.

Scanner data（TV・mobile phone）

- **Data used for the trial calculation** (from Jan. to Dec., 2018)

1. Scanner data

- Contains approximately **190,000** prices for TV, **200,000** prices for mobile phone per month
- Includes not only retail prices **but also online prices such as Amazon**
- is purchased from a marketing research company in Japan

2. Collected data used for the current index

- TV : Price collectors collect the retail prices of the models conform to the specifications that SBJ designates (screen size 32 inches, etc.).
- Mobile phone : SBJ selects survey models based on the data of shipments every month. And then, SBJ collects price data of them manually from main telecommunication carriers' websites.

- **Trial calculation**

1. Hedonic index $\ln p_T = \alpha_t + \beta_t \delta_{T,t} + \sum_k \gamma_{t,k} x_k$

where k is a model, x_k is a value of model k and $\delta_{T,t}$ is time dummy (1 when T=t, 0 when T=t-1)

2. Törnqvist index

3. Matched-model Jevons index

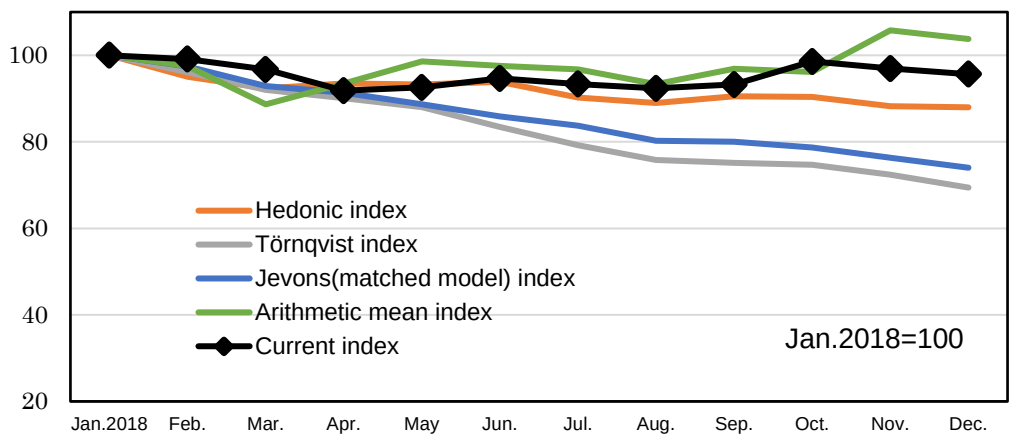
4. Matched-model Fisher index

5. Arithmetic mean index

6. Current index

- When survey models are changed, we adjust the qualities among models by using overlap method, regression equation, etc.
- Regarding the mobile phone, we adjusted the qualities by using overlap method in March, July and December.

TV



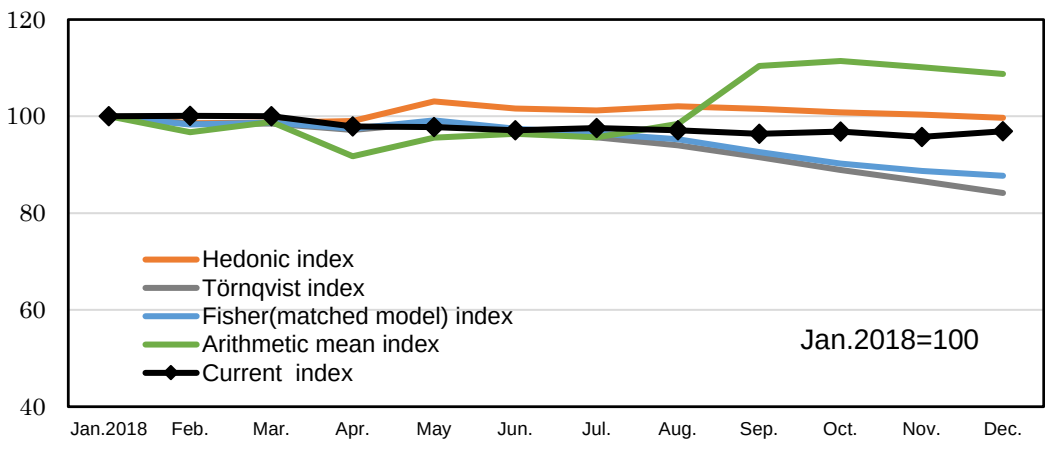
- **Explanatory variables used for calculating Hedonic index**

1 Time dummy 2 Display dummy (LCD=1, OEL=0) 3 Screen size 4 HDD memory capacity
5 TV tuner dummy (tuner separated from TV screen=1, other type of tuner=0)
6 Function of connecting the Internet dummy (with it=1, without it=0)
7 Resolution dummy (HDTV=1, the others=0) 8 Elapsed time from release date

- **Results**

- Among four trial indices, **fluctuation of the hedonic index, which covers much more TV models than the current index, is most similar to that of the current index**.
- The Törnqvist index, which is calculated with the sales amounts, has a marked downward bias because the fact that the less expensive TV models are, the more they are sold, is reflected.
- Chaining period-to-period matched-model Jevons index, not considering the sales amounts, has less downward bias than the Törnqvist index.
- Both of them have downward biases compared with the other indices.

Mobile phone



- **Explanatory variables used for calculating Hedonic index**

1 Time dummy 2 Resolution 3 The number of pixels of built-in camera 4 SIM-free dummy (sim-free model=1, otherwise=0) 5 Telecommunication carrier dummy 6 manufacturer dummy 7 Data capacity 8 weight by inch
9 Elapsed time from release date

- **Results**

- Among four trial indices, fluctuation of **the hedonic index is most similar to that of the current index in the long term.**
- Arithmetic mean index rose in September 2018, because new expensive mobile phone models with high quality were released. On the other hand, the hedonic index is stable because of the effect of the quality adjustment.
- The Törnqvist index and the matched model Fisher index have downward biases compared with the other indices as is the case of TV indices.

Web-scraped data (Airplane fares)

- **Current index**

- Select 10 main routes in Japan while referring to the number of passengers by route.
- Collect one or two price data by each route from the reservation sites of each airline company manually.
- Use three types of fares; normal fare, round-trip fare and the lowest discount fare.
- Calculate average prices and the price index using the following formula.

1. Calculate average prices of time t, route a, ticket type b and departure date c by using the number of passengers by route and airline companies.

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

2. Calculate average prices of time t, route a and ticket type b by using the number of dates in a month (n).

$$p_{t,a,b} = \frac{\sum_c p_{t,a,b,c}}{n}$$

3. Calculate average prices of time t and route a by using the sales amounts by ticket type (q).

$$p_{t,a} = \frac{\sum_b p_{t,a,b} q_{0,b}}{\sum_b q_{0,b}}$$

4. Calculate the price index by using the number of passengers by route.

$$I_t = \frac{\sum_a p_{t,a} q_{0,a}}{\sum_a p_{0,a} q_{0,a}} \times 100$$

t : time
0 : base period
a : route
b : ticket type
c : departure date
d : airline

- **Trial index by web-scraped data** (from Jan. to Sep., 2018)

- Select the same routes as the current index.
- **Collect the data (price, departure date, route etc.) of all the flights in each route** every day from reservation sites of major Japanese airline companies **by web scraping**. (This task is outsourced. The number of price data collected is between 500 thousand and 800 thousand in each month.)
- Use prices of **airline tickets for 28days, 45days, 55days, 75days in advance, of which discount rate are different from each other**
- **Interview main airline companies on the information of sales amounts** to verify the scraped data.
- Calculate average prices using sales amounts by ticket types and by holiday/weekday, obtained from the interview. The formula is shown below.

1. Calculate average prices of time t, route a, airline b, ticket type c and weekday/holiday d by using the number of flights in the route (m).

$$p_{t,a,b,c,d} = \sum_e p_{t,a,b,c,d,e} \,/\, m_{t,a,b,c,d}$$

2. Calculate average prices of time t, route a, airline b and weekday/holiday d by using sales amounts of airline companies and ticket types (q) as weights.

$$p_{t,a,b,d} = \sum_c p_{t,a,b,c,d} q_{0,b,c} \,/\, \sum_c q_{0,b,c}$$

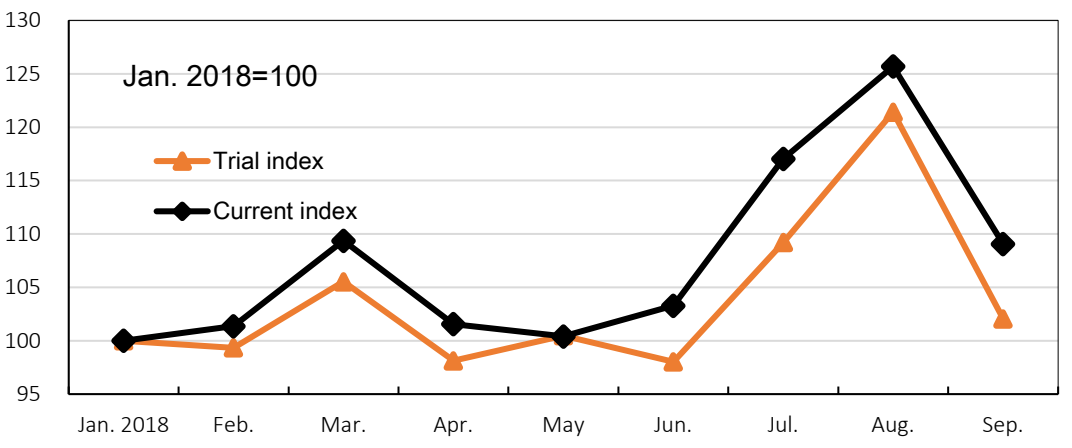
3. Calculate average prices of time t and weekday/holiday d by using sales amounts of airline companies and route (Q) as weights.

$$p_{t,d} = \sum_{a,b} p_{t,a,b,d} Q_{0,a,b} \,/\, \sum_{a,b} Q_{0,a,b}$$

4. Calculate average prices of time t by using sales amounts of "weekday (except the day before holidays)" and "holidays and the day before holiday" (r) as weights.

$$p_t = \sum_d p_{t,d} r_{0,d} \,/\, \sum_d r_{0,d}$$

t : time, 0 : base period, *a* : route, *b* : airline,
d : weekday(except the day before holiday) or "holidays and the day before holiday", *e* : flight
c : ticket type (normal fare, round-trip fare, discount fare (28days,45days,55days,75days in advance))



- **Results**

- **The fluctuation of the trial index is similar to that of the current index**. But, **the range of fluctuation of the trial index, calculated by using web-scraped data, is smaller than that of the current index**.
- In May 2018, the fluctuation of the trial index differs from that of the current index. This is because price data used for the current index and the trial index during the long vacation period (from the end of April to early May) differs.
- More specifically, the current index is calculated using only one or two price data of outward flights by route. On the other hand, the trial index is calculated using much more price data including inward flights. Thus, it reflects not only the demand increase at the end of April but also early May.

Next Steps

- From the tentative analysis of **scanner data**, we found that **the hedonic index is most similar to the current index**. We need to **prioritize items of which we should calculate the price indices by hedonic method with scanner data while considering the efficient use of research resources**.
- In this poster, we could utilize only the short-term data. **After the accumulation of long-term data, we plan to calculate price indices by multilateral methods such as GEKS-method**.
- In this trial, we **succeeded to collect price data stably and automatically by web scraping**.
- We perceive that we need to get the approval of the companies for collecting data from their websites regularly and **maintain good relations** with them in order not only to resolve the problems such as blocking but also **to get sales amounts data to update weights**.