2020 Asia–Pacific Statistics Week
A decade of action for the 2030 Agenda: Statistics that leaves no one and nowhere behind
15-19 JUNE 2020 | Bangkok, Thailand

Utilizing A Price Comparison Website to Produce Hedonic Price Indices for Mobile Phone

Listianingrum, Tri[1]; Nefriana, Radenroro[1]

[1] Statistics Indonesia

Abstract

A Consumer Price Index (CPI) is intended to measure pure price changes. The products whose prices are collected and compared in successive periods should ideally be perfectly matched; they should be identical concerning their physical and economic characteristics. This is problematic for commodities that change rapidly such as telecommunications and technology goods. The goods that are available on the market today are no more directly comparable with those which were available before. Measurement of price changes for those commodities is even more complicated by frequent disappearance of the goods that belong to the basket for CPI and the appearance of a brand-new type of that commodity that has a different quality than before. Thus, the price needs to be quality-adjusted.

Some approaches are available for dealing with the issue. One of them is by employing the hedonic method. The hedonic method can measure relative price changes while holding the quality and underlying characteristics constant by employing hedonic regression. This method is powerful but has a disadvantage: it needs a big set of data about the price and characteristics of the products which might be laborious to gather by enumerators.

Web scraping is a way to gather a big data set about price movement which also allows us to dig deeper under the price and collect the characteristics of products. This paper will elaborate on a study on utilizing price comparison website data which provides the data about the prices and the characteristics of products that are being sold on the internet. With those characteristics information available, the hedonic method was applied.

This research only focused on mobile phones, considering that mobile phone is a commodity that experiences a rapid change in CPI compilation. However, this research could be a benchmark to be applied to other commodities. Web scraping had been done weekly for the mobile phone’s price and characteristics from iprice.co.id from January to March 2020. Price indices were produced using 4 methods: fixed basket, matched model, time dummy model, and double imputation model.

.

Keywords: mobile phone; hedonic index; price comparison website

1. Introduction

For measuring inflations, deflating monetary magnitudes to reveal changes in real values in the National Accounts, and other uses, Statistics Indonesia (BPS) produces CPIs by using data that is collected through a consumer price survey, namely Survei Harga Konsumen (SHK). However, despite the importance of CPIs, there are some challenges in compiling them. First, a CPI is intended to measure pure price changes. The products whose prices are collected and compared in successive periods should ideally be perfectly matched; that is, they should be identical in respect of their physical and economic characteristics (ILO et al, 2004, p. 25). This is problematic for commodities that change rapidly such as telecommunications and technology goods. The goods that are available on the market today are no more directly comparable with those which were available before. Measurement of price changes for those commodities is even more complicated by frequent disappearance of the goods that belong to the basket for CPI and the appearance of a brand-new type of that commodity that previously had a different quality. Second, collecting a complete set of product characteristics through a field survey is resource-intensive. Moreover, given some circumstances (e.g. COVID-19), it is sometimes impossible to conduct field surveys.

When faced with measuring prices for products that undergo rapid quality change, the international best practice is to develop hedonic price indices, provided suitable source data are available (Trewin, 2005). The hedonic method is particularly well suited for comparing goods that can be thought of as comprising a bundle of underlying attributes, each of which is assumed to have its intrinsic values. Nevertheless, another problem exists. Product characteristics that are needed to perform this method are usually missing from survey results.

According to Cavallo and Rigobon (2016), big data is prospective to improve statistics and empirical research in economics. Hypothetically, it is also promising to be used for the construction of CPIs. This is reasonable because a vast number of online prices are displayed on e-commerces (even though it has to be kept in mind that big data surely cannot fulfill the exact concept and definition used in SHK). It also allows us to elicit a more complete set of product characteristics that can be laborious to gather via a field survey. Furthermore, this can also be done amid some circumstances that do not allow for field surveys to be conducted.

Considering the aforementioned opportunities, a study on utilizing a price comparison website to produce hedonic price indices for a product with a rapid quality change has been done. The scope of the research was limited for mobile phones with the hope that this research could be a benchmark to be applied to other commodities. The reason for choosing mobile phones was because the electronic category ranked third most frequently purchased products by households online in Indonesia (Ministry of Communication and Information Technology, 2016). The change in mobile phone products is frequent because the market for mobile phones is moving fast and new products are introduced every month. Also, in the case of mobile phones, specifications can be considered independent, measurable attributes. By comparing prices and specifications of various mobile phones, the hedonic regression model can assign values to each of the particular features that are identified as "price-determining".

Related works on price index methodology have also been done in some countries. Statistics New Zealand has used the hedonic model for used cars since 2001. In 2011 the model was updated by fitting the log of price, adding extra characteristics, and adding squared terms for the age of the car and size of the engine (Bentley & Krsinich, 2017). Statistics Belgium used scraped data to construct the hedonic model for consumer electronics and second-hand cars in 2018 (Loon & Roels, 2018). In Indonesia, a study has been conducted by utilizing big data to develop an alternative Residential Property Price Index (RPPI) for the secondary market (existing house), employing the hedonic method as a quality-mix adjustment to calculate robust asking price indices given the availability of property characteristics data (Rachman, 2019).

2. Methodology

Figure 1 illustrates the methodology used in this research.

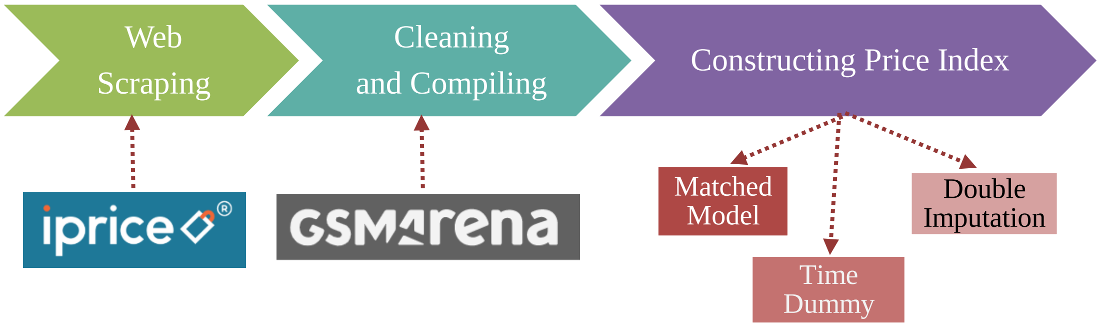

Figure 1. The research method

Web Scraping

Collecting online prices is not trivial because they are posted on hundreds of different e-commerce that lack of homogeneous structure and format. This makes the data cannot be easily collected and used for CPI calculation. Differently, price comparison websites work by collecting product information, including pricing, from participating retailers and then displaying that collective information on a single results page in response to a search query. This makes the data structure from various e-commerce more homogeneous and hence easier to compare. For the research, There are 3 kinds of data needed to be scraped: types of products sold the most at each period; price for each type of product from retailers at each period; and characteristics for each product. Those data can be obtained from a price comparison website. Producing price indices by utilizing big data from a price comparison website has also been done before in Japan (Abe & Shinozaki, 2018).

A price comparison website, price.co.id, was chosen as the main data source for scraping after considering several criteria: the availability of the data needed, its popularity in Indonesia, and its scraping feasibility. Subsequently, web scraping began by understanding the structure of the website. The scraper was developed by using Python with the BeautifulSoup package. To capture the price movement, scrapings were conducted every Wednesday from January 1st, 2020, till the end of March 2020. Besides the price information, 35 specification variables of each mobile phone displayed were also collected to construct a hedonic regression.

Further, to limit the scope while retaining the purpose of this study, some selections were done. Top mobile phone vendors in the 3rd quarter of 2019 in Indonesia were Oppo, Xiaomi, Samsung, Vivo, and Realme with a total market share of 94% (Canalys in Khoirunnisa, 2018). Given that information, they were selected as the target scope. For each product on sale, iprice.co.id displays many prices from various online retailers to compare. Hence, for each mobile phone, only the first 4 prices displayed on each e-commerce were recorded, alongside the screen size, screen resolution, density, RAM, internal memory, battery capacity, and many other features information. Whereas there are 11 e-commerce covered: Shopee, Tokopedia, Blibli, Lazada, Lazmall, Tokopedia, Arjuna Electronics, Personal Digital, Bukalapak, Amazon, and Blanja. They are some of the most visited e-commerce sites in Indonesia. Thus, it was expected that they are representative of online mobile sales in Indonesia.

Cleaning and Compiling

Mobile phone types usually come in various combinations of internal memory and RAM sizes. For example, the Xiaomi Note 8 Pro type comes in a choice of a combination of internal memory and RAM: 64GB 6GB RAM, 128GB 6GB RAM, 128GB 8GB RAM, and 256GB 8GB RAM. The price depends greatly on the size of the internal memory and RAM in it. So, if that information is not obtained, the data cannot be used.

Unfortunately, the data obtained from scraping suffered some problems. Not only some of the memory and RAM information was missing, but the specifications information was also inconsistent. Therefore, data from iprice.co.id was equipped with a specification database that was built based on gsmarena.com to maintain data consistency and reliability. Gsmarena.com is a well-known website as a reference for mobile phone information which currently occupies 535th rank on Alexa globally.

Constructing The Price Indices

There were 4 models to compare: the fixed basket, matched model, time dummy model, and double imputation model. The fixed basket and matched model use direct comparison for prices in a successive period without regression function. Meanwhile, time dummy and double imputation utilize hedonic regression for calculating the price index.

Mobile phones' information was scraped weekly (t) from iprice.co.id (Sample/S) from week 1 to week 13. Some products were available in the earlier period, but not the latter (Death/D). Conversely, other products were available in the latter period, but not earlier (Birth/B). There was also a set of products that are available in both periods (Matched/M). These phenomena are illustrated in Figure 2.

S^{t-1}	S^t
D^{t-1}	▨
M^{t-1}	M^t
▨	B^t

Figure 2. Samples from period t-1 and t

The fixed basket and matched model are almost similar, except the fixed basket selects only products that are available through all periods (t). The price index is directly calculated by comparing price in period t to the base period (week 1). Meanwhile, the matched model selects matched products in 2 successive periods (Mt-1 and Mt) to calculate a relative price for 2 consecutive periods and then use it to produce a price index. In the fixed basket, the number of products is fixed for each period, whereas in the matched model it could vary depending on the matched products in 2 consecutive periods.

The time dummy hedonic price index uses the hedonic regression to directly determine a price index by making time as one of the predictors. The coefficient for each period is an index number. It is done by pooling entire samples for all periods 0 to t and then uses this large pooled dataset to directly determine an index number which measures price change from the base period to the current period.

The double imputation is an adaptation from research conducted by ABS on constructing hedonic price indices for personal computers. It is a combination of the matched model and the time dummy model. It uses a matched model to determine a price change over two consecutive periods for Mt-1 and Mt, and use hedonic regression to measure price movements for Bt and Dt-1. Both then combined to produce a price index for two consecutive periods. This price index is then chained to provide a measure of price change from a base period to the current period.

The approach begins with the construction of a pooled data set for two consecutive periods: t-1 and t. Thus, the data set used for modeling is:

$$S_{POOLED}^{t-1,t} = D^{t-1} \cup M^{t-1} \cup M^{t} \cup B^{t}$$

This differs from the time dummy in two key ways. First, pooling only occurs for two consecutive periods. Second, the resulting price index is not the final measure of change from period t-1 to period t but instead is used in later stages of the calculation.

The double imputation index is constructed from the two indices: $I_{MM}^{t-1,t}$ for observations matched between the two periods; and $I_{2TD}^{t-1,t}$ for new items or discontinued lines. The double imputation price index measuring price change from period t-1 to period t is then a weighted geometric mean of the two-component indices.

$$I_{DI}^{t-1,t} = \left[I_{MM}^{t-1,t}\right]^{f_m}\left[I_{2TD}^{t-1,t}\right]^{1-f_m}$$

where $\boldsymbol{f_m}$ is the weighted fraction of matched observations from period t to period t-1.

To construct a price index from some earlier price reference (base) period, a chained series id formed as

$$I_{DI,Chain}^{t-1,t} = I_{DI,Chain}^{t-1,t} \times I_{DI}^{t-1,t}$$

3. Result

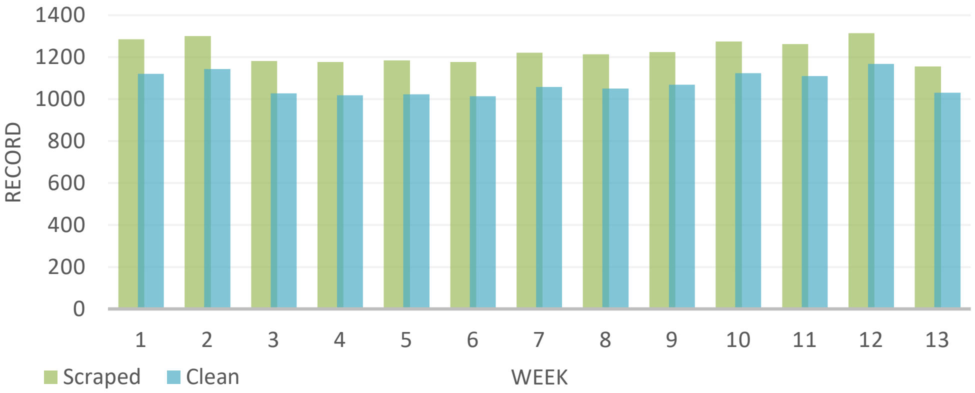

Figure 3. Amount of weekly scraped and cleaned data

The scraping process took about 30 minutes on average. Problems in scraping occurred when the HTML structure of the website changed so the scraping program needed to be modified. For each scraping process, a CSV file was generated containing the product name, 44 price fields from all e-commerces, and 35 product specifications to be used as explanatory variables in hedonic regression. However, the product specifications scraped from iprice.co.id was often suffering from inconsistency and incompleteness. To deal with that, a phone specification database was developed based on gsmarena.com as a reference to clean the data. The database contained full specification information of 430 types of mobile phone for the 5 mobile phone vendors included in this research. Figure 3 shows the number of records being scrapped every week and the amount of data after cleaning.

The regression model uses ln(price) as the response variable and the stepwise method for the predictor selection. In this study, from 38 specifications information available, the regression finally resulted in 10 variables that significantly affected mobile phone price: RAM, internal storage, screen size, screen density, battery, weight, vendor, material, simcard type, screen type, type of front glass, NFC, LED notification light, iris scanner, fingerprint scanner, radio FM, and proximity.

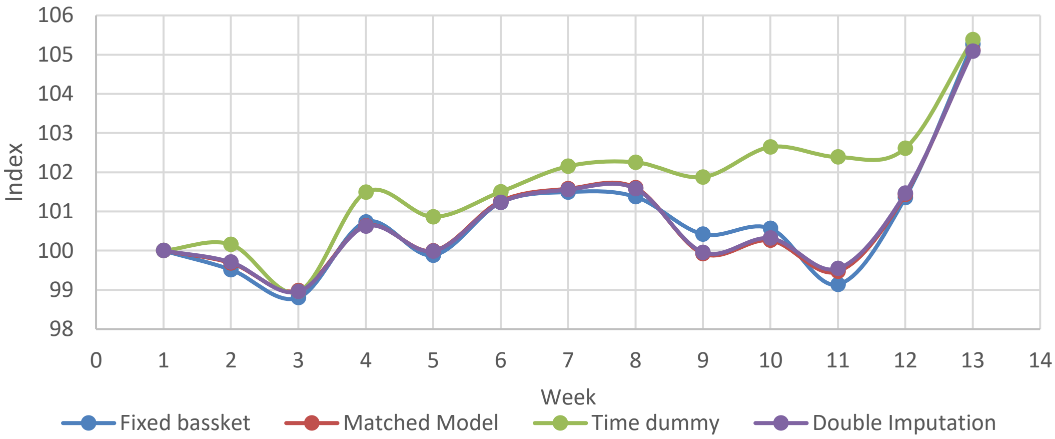

Figure 4. The weekly price index for mobile phone January to March 2020

Figure 4 illustrates a comparison of indices that resulted in the study. The double imputation and matched models appeared to coincide. This was because the proportion of unmatched data was very small compared to those matched, so the resulting indices were almost the same as the ones resulted from the matched model. Generally, the indices movements for the fixed basket matched and double imputation models were almost the same, while for the time dummy indices they were different because they purely utilized the hedonic model. Because the model only had an R squared of 80 percent, there was 20 percent diversity that cannot be explained from the model.

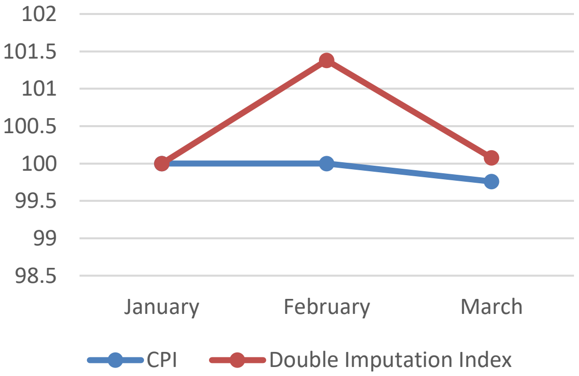

Figure 5. Price index comparison with CPI (January 2020=100)

Figure 5 shows the double imputation indices resulted in this study that tend to be more volatile than the CPIs for mobile phone compiled by BPS. However, drawing further conclusions between the two kinds of indices would still be too early because the series was very short.

4. Discussion, Conclusion, and Recommendations

a. A price comparison website was beneficial for collecting online price data for a product from various e-commerce sites without the need to visit them one by one. Nonetheless, a comparison website alone was not enough to be used as a basis for the compilation of hedonic indices, because sometimes there were problems in the completeness and consistency of the product's specifications. A database for products' specifications that was more accurate was essential as a reference for validating the data.

b. Data scraping was reliable for collecting a large amount of data in a short time with minimal resources. It would even allow for producing daily price indices. However, the cleaning process was still important since scraped data did not appear as perfect and ready. In the future, a detailed methodology needs to be formulated to capture price changes and eliminate noise from big data.

c. The time dummy model gave the most distorted results than the other methods. The hedonic model performed best when used as an indirect approach and combined with other methodologies to produce indices.

d. Some products displayed on the price comparison website were old products. It was possible that most of them still appeared because some sellers still had old stocks. It might be necessary for future researches to filter the data based on the year of production to avoid old stocks that are rarely found on the market presently.

References:

1. Abe N, Shinozaki K. 2018. *Compilation of Experimental Price Indices Using Big Data and Machine Learning: A Comparative Analysis and Validity Verification of Quality Adjustment*. Bank of Japan Working Paper Series. 18-E-13.
2. Bentley A, Krsinich F. 2017. *Toward Big Data CPI for New Zealand*. Paper presented at the Ottawa Group 2017. Eltville, Germany.
3. Badan Pusat Statistik. 2013. *Pedoman Survei Harga Konsumen*. Jakarta: BPS.
4. Cavallo A, Rigobon R. 2016. *The Billion Prices Project: Using Online Prices for Measurement and Research*. Journal of Economic Perspectives. 30 (2): 151-78. doi: 10.1257/jep.30.2.151.
5. International Labour Office *et al*. 2004. *Consumer Price Index Manual: Theory and practice*. Retrieved on November 1, 2019, at http://www.ilo.org/wcmsp5/ groups/public/---dgreports/---stat/documents/presentation/wcms_331153.pdf.
6. Khoirunnisa. 2019. *Top 5 Vendor Smartphone di Indonesia Q3-210*. Retrieved on May 1, 2020, at https://selular.id/2019/11/top-5-vendor-smartphone-di-indonesia-q3-2019/.
7. Loon KV, Roels D. 2018. *Integrating big data in the Belgian CPI. Meeting of the Group of Experts on Consumer Price Indices.* Geneva, Switzerland.
8. Ministry of Communication and Information Technology. 2016. ICT Indicators Infographic.
9. Rachman, AN. 2019. *An Alternative Hedonic Residential Property Price Index for Indonesia Using Big Data: The Case of Jakarta*. International Conference on Real Estate Statistics. Luxembourg: Eurostat.
10. Trewin, D. 2005. *The Introduction of Hedonic Price Indices for Personal Computers.* Canberra: ABS.