HEDONIC PRICE INDEX FOR MOBILE PHONE BASED ON E-COMMERCE DATA

TRI LISTIANINGRUM

**APPLIED STATISTICS
GRADUATE SCHOOL
IPB UNIVERSITY
BOGOR
2021**

STATEMENT ABOUT THESIS, INFORMATION SOURCES, AND COPYRIGHT ASSIGNMENT

I hereby declare that the thesis entitled “Hedonic Price Index for Mobile Phone Based on E-Commerce Data” is my work with the direction of the supervising committee and has not been submitted in any form to any college. Resources originating or quoted from works published and unpublished from other writers have been mentioned in the text and listed in the Bibliography at the end of this thesis.

I hereby assign the copyright of my writing to the IPB University.

Bogor, January 2021

Tri Listianingrum
NIM. G152184524

RINGKASAN

TRI LISTIANINGRUM. Indeks Harga Hedonis untuk Telepon Seluler Berdasarkan Data *E-Commerce*. Dibimbing oleh FARIT MOCHAMAD AFENDI dan BAGUS SARTONO.

Big data memberikan peluang untuk membangun kumpulan data yang sesuai dengan kebutuhan penelitian. Salah satu perhatian global mengenai pemanfaatan *big data* adalah dalam penggunaannya untuk mengukur indeks harga. Seiring dengan perkembangan situs *e-commerce* yang pesat, banyak informasi mengenai harga dan spesifikasi produk yang tersedia di internet. Informasi ini memungkinkan peneliti untuk menghitung indeks harga hedonis untuk produk yang mengalami perubahan kualitas yang cepat seperti telepon seluler.

Salah satu metode untuk mengumpulkan *big data* adalah dengan menerapkan pengumpulan data skala besar di web dengan proses *scraping* yang diautomatisasi. Pada penelitian ini dibuat *web scraper* dengan menggunakan bahasa *python* untuk mengekstrak informasi tentang produk, harga dan karakteristik dari berbagai *e-commerce* yang ditampilkan di website IPrice.co.id. Ekstraksi data dilakukan setiap minggu dari Januari hingga Juni 2020 dan mengumpulkan lebih dari 34.000 amatan secara total. Sebuah database yang berisi spesifikasi lengkap dari 458 jenis telepon seluler dikembangkan berdasarkan informasi dari gsmarena.com untuk memvalidasi data hasil *scraping*.

Metode yang digunakan untuk menghitung indeks harga hedonis adalah *double imputation method*. Salah satu langkah dalam metode ini adalah membuat model regresi hedonis untuk dataset yang terdiri dari dua periode yang berurutan. Karena adanya pencilan dan nilai *leverage* yang tinggi, regresi kekar dengan estimasi MM (*Multi-Stage Method*) digunakan untuk mendapatkan estimasi parameter yang andal dan efisien. Tidak adanya informasi mengenai kuantitas produk sebagai dasar pembobotan membuat peneliti melakukan beberapa alternatif skenario pembobotan. Yang pertama adalah pembobotan berdasarkan tahun rilis telepon seluler, yang kedua adalah dengan menggunakan pangsa pasar sebagai *proxy* untuk pembobotan, dan yang ketiga adalah kombinasi keduanya.

Hasil penelitian menunjukkan bahwa pergerakan harga berdasarkan transaksi daring yang ditangkap dalam penelitian ini tidak mencerminkan transaksi luring yang tercermin dalam Indeks Harga Konsumen (IHK) yang dihasilkan oleh Badan Pusat Statistik (BPS). Indeks mingguan yang dihasilkan memiliki fluktuasi yang lebih besar daripada IHK. Pembobotan yang didasarkan pada tahun rilis dan pangsa pasar menghasilkan indeks yang paling stabil.

Kata kunci: Indeks Harga Daring, Regresi Hedonis, *Web Harvesting*

SUMMARY

TRI LISTIANINGRUM. Hedonic Price Index for Mobile Phone Based on E-Commerce Data. Under supervision from FARIT MOCHAMAD AFENDI and BAGUS SARTONO.

Big data provides an opportunity to build datasets that fit specific requirements for research. One of the global concerns about big data is how to utilize it for measuring price index. Due to the proliferation of e-commerce sites, a vast amount of prices information followed by the product specifications are available on the internet. This information would allow researchers to calculate hedonic indices for the products undergo rapid quality change like a mobile phone.

A method to gather big data is by implementing a large-scale data collection on the web with an automated scraping process. In this study, a web scraper was built using python language to extract information about product, price and characteristics from various e-commerce displayed on the IPrice.co.id website. The data extraction was conducted weekly from January to June 2020 and gathered more than 34,000 records in total. A database containing comprehensive specifications of 458 types of mobile phones was developed based on information from gsmarena.com to validate the scraped data.

The method used for calculating the hedonic price index is the double imputation method. One of the steps in this method is to construct a hedonic regression model for datasets consists of two consecutive periods. Due to the presence of outliers and leverage points, robust regression with MM (Multi-Stage Method) estimation was used to get a reliable and efficient estimate of the parameters. Due to the lack of information on the product quantity for weighting, several alternative weighting scenarios were taken. The first is weighting based on the release year mobile phone, the second is by using market share as a proxy for weighting, and the third is the combination of both.

The result shows that price movement based on online transactions captured in this study do not reflects the one from offline transactions captured in the Consumer Price Index (CPI) produced by Badan Pusat Statistik (BPS). The weekly indices have a wider fluctuation than the CPI. Weighting based on both release year of mobile phone and market share produces the most stabilized index.

Keywords: Hedonic Regression, Online Price Index, Web Harvesting

HEDONIC PRICE INDEX FOR MOBILE PHONE BASED ON E-COMMERCE DATA

TRI LISTIANINGRUM

Thesis
as one of the requirements to obtain
Master of Science
in
Applied Statistics Program

APPLIED STATISTICS
GRADUATE SCHOOL
IPB UNIVERSITY
BOGOR
2021

digitally signed

dsign.ipb.ac.id

Examiner on Thesis Examination: Dr. Ir. Erfiani, M.Si.

Thesis Title : Hedonic Price Index for Mobile Phone Based on E-Commerce Data
Name : Tri Listianingrum
Student ID : G152184524

Approved by

Supervisor:
Dr. Farit Mochamad Afendi, S.Si., M.Si.

digitally signed
disign.ipb.ac.id

Co-Supervisor:
Dr. Bagus Sartono, S.Si., M.Si.

digitally signed
disign.ipb.ac.id

Acknowledged by

Head of Applied Statistics Program:
Dr. Kusman Sadik, S.Si., M.Si.
NIP 19690912 199702 1 001

digitally signed
disign.ipb.ac.id

Dean of Graduate School
Prof. Dr. Ir. Anas Miftah Fauzi, M.Eng.
NIP 19600419 198503 1 002

Digitally signed by:
Anas Miftah Fauzi
[3FE49AA995DD0C4F]
Date: 13 Feb 2021 14.43.15 WIB
Verify at disign.ipb.ac.id

Examination Date:
January 20th 2021

Graduation Date:
January 29th 2021

PREFACE

Alhamdulillahirabbil‘alamin, praise is mere to the Almighty Allah SWT for the gracious mercy and tremendous blessing that enabled me to finish this study in the middle of the Covid-19 pandemic. This study conducted from July 2019 to January 2021 is a thesis titled “Hedonic Price Index for Mobile Phone Based on E-Commerce Data”.

I would never have been accomplished this thesis without the support of many people. Therefore, I wish to show my appreciation to my research supervisors: Dr. Farit Mochamad Afendi, S.Si., M.Si. and Dr. Bagus Sartono, S.Si., M.Si. for their valuable assistance in developing this thesis. I want to thank my husband and son for their patience and understanding, in times when this woman is switching to college student mode, and for always blesses me with statistically significant joy in life. I wish to extend my special thanks to college mates from BPS STT 2018 for the companion and collaboration throughout the journey.

I hope that the result of this thesis could be an added value for science, especially in the development of price index methodology in Indonesia. Above all, hopefully, the completion of this study could make me a wiser and humbler person.

Bogor, January 2021

Tri Listianingrum

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

I INTRODUCTION

1.1 Background

Today, people produce a massive amount of new data on the internet real-time as a digital footprint of our day to day life. De Mauro *et al.* (2016) discussed the formal definition of big data is the information asset characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value. According to Cavallo and Rigobon (2016), big data is prospective to improve statistics and empirical research in economics. Big data provides an opportunity to stop treating data as given because it enables anyone to build a dataset to fit specific research requirements. One of the global concerns about the use of big data in official statistics is on how to utilize it for measuring price index.

One of the official price indices that are produced regularly in Indonesia is the Consumer Price Index (CPI) which is organized by Badan Pusat Statistik (BPS) as the Indonesian National Statistics Office. The compilation of Indonesian CPI uses the data collected through a monthly survey called Survey Harga Konsumen (SHK). The enumeration is performed by visiting outlets, merchants or markets selected and collect the price of the commodities. BPS also utilizes online sources for compiling CPI, like in gathering the price of gasoline and diesel from Pertamina website and price of mobile phone credit from official provider websites. However, in this case, online prices captured are official prices whose value is consistent with offline prices. It doesn’t necessarily reflect the price movement for online transactions.

The proliferation of e-commerce sites made a vast number of prices information are available on the internet. While the data dispersed across countless webpages, advances in automated scraping method now allow anyone to design and implement a massive scale of data collection on the web. While detailed information for each good is available, the new and disappearing product can be quickly detected and accounted for (Cavallo and Rigobon 2016). At a glance, online data collection offering efficiency and effectiveness because it is relatively cheap and enables to produce a large dataset in a small amount of time. However, online prices are “found data”, in the sense that measuring price index is not the actual purpose of its appearance. It tends to have very comprehensive coverage for a haphazard subset of transactions (Bentley and Krsinich 2017). Thus, the data gathering demands a methodical step to obtain a representative sample.

In Indonesia, e-commerce enthusiasts are increasing. Google’s and Temasek’s study found that in 2018, 18 million people which is 7% of the population fell into the category of online buyers. Indonesia’s Internet economy has more than quadrupled in size since 2015 at an average growth rate of 49% a year (Google and Temasek 2019). However, current SHK methodology mainly captures offline transactions on the store. Now, due to the shifting behaviour of consumer towards buying things online, it is necessary to formulate an online price index.

According to Deloitte Consumers Insights Survey (2018), ‘Digital gadgets, computers and accessories’ ranks fourth in the category of goods most often purchased online in Indonesia. With the accelerating speed of technological advancement, mobile phones have become a mandatory component of people' daily performance. Based on research by a data analytics company - Statista, mobile phone users were 63.3% in 2019, and it is estimated to reach at least 89.2% of the population in Indonesia in 2025. It is amongst commodities that taken into account for computing the CPI.

Telecommunications and technology goods are often experiencing rapid improvement over time, and the same applied to mobile phones. The release of a new type of mobile phone having enhancement in technology occurred in a short time. Measurement of price changes for those commodities would have to consider the appearance and disappearance of new and old goods from the market. The direct comparison method assumes that the quality difference between old and new products is ignorable and thus processes price change due to quality changes as zero. Continually ignoring small changes in the quality of replacements can lead to an upward bias in the index (ILO *et al.* 2004, p. 112).

When faced with measuring prices for products that undergo rapid quality change, international best practice is to develop hedonic price indices, provided appropriate source data are available (Trewin 2005). The hedonic method is particularly well suited for comparing goods that considered of as comprising a bundle of underlying attributes, each of which is assumed to have its intrinsic value. A mobile phone comes with particular technical and performance characteristics which are known as specifications. These make them differentiated products with many alternative designs and selling prices in the market. According to Lancaster (1966), consumers derive utility from the attributes of the product rather than the product itself. It means, consumers do not purchase a mobile phone as it is; but rather they buy bundles of its specifications and features that come with it. Mobile phone price depends on the set of specifications embodied such as brand, battery duration, display size, weight, and camera resolution that makes it suitable to be analyzed with the hedonic method.

Formulation of an online price index is an issue that is widely discussed by many countries. It is a way to embrace the digital era where retail is not only comprised of the traditional market. It gained popularity when online retail made a subtle appearance with steady escalation showing promising growth in the future. Price data from e-commerce are practical to formulate a price index for the online retail transaction. Web scraping allows gathering a data set about price movement and collect the characteristics of the products at once. With characteristics information available, we could employ a quality adjustment with a hedonic method for estimating the price index.

There are several related works on the development of price index methodology from many countries. Stats New Zealand has been using a hedonic model for used cars since 2001 and by 2011 updated the model by fitting log of price, adding more characteristics, and adding squared terms for age of car and size of the engine as discussed by Bentley and Krsinich (2017). Statistics Belgium in 2018 used scraped data to construct a hedonic model for consumer electronics and second-hand cars (Loon and Roels 2018).

In Indonesia, Rachman (2019) conducted a study by utilizing big data to develop an alternative residential property prices index (RPPI) for the secondary market (existing house) to challenge Bank Indonesia's existing survey-based RPPI.

The study employed hedonic methods as a quality-mix adjustment to calculate robust asking prices indexes given the availability of property characteristics data.

1.2 Objectives

This study aims to:

1. formulate a hedonic regression model for mobile phone's online retail price,
2. calculate online price indices for mobile phone by implementing the double imputation method.

1.3 Benefit

This study could provide a preliminary investigation for CPI compilation in BPS to shift towards big data approach. It also gives an overview of the possibility to implement a hedonic method for quality adjustment in the long run. Although this study only covers one commodity, it can be a benchmark to apply the hedonic method for others in the future.

1.4 Limitations

This research aspires to produce the best possible results with the available resources. Some aspects of this research are objects of limitations to achieve the objective of the study while respecting time boundaries. The following list is limitations to this study.

1. Web scraping is conducted weekly; every Wednesday from January to June 2020.
2. The source of data scraping is IPrice.co.id web page.
3. The ideal dataset to measure CPI should be a real-time data covering price and volume (quality and quantity) information for all products. The target population of transactions should have three dimensions: product, place and time (Bentley and Krsinich 2017). However, the current scraping method could not generate such comprehensive dataset. In this study, the researcher measured the price index without information about quantity. Furthermore, there is no geographical information on the target population of transactions: it consists only of product and time dimensions.

II METHODS

2.1 Research Framework

Figure 1 shows the details of the research framework. Transactions between seller and buyer occur online and offline. Reflecting on the methodology, SHK only captures price movements in the offline one. This research aims to figure price movement in online transactions. The main steps of this research are: scraping the data from the internet, cleaning data, hedonic regression analysis, and calculating index using double imputation method.

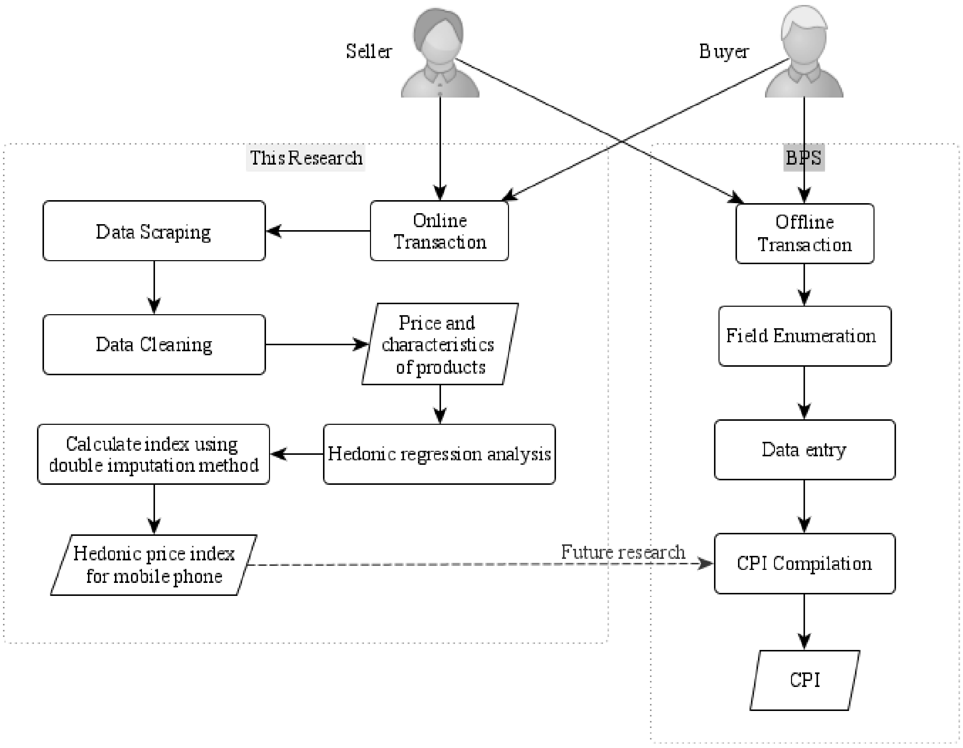


Figure 1 Research framework

The primary data-gathering technique in SHK is the field enumeration. To capture online transactions, we could collect the data through web-scraping. Online retail prices on e-commerce represent the price for a product in that specific time. Collecting this data continuously and consistently would provide an alternate data source for compiling price indices. Web scraping also allows collecting more detailed information like the characteristics of a product.

This study use regression analysis with price as the response variable and characteristics as explanatory variables. It is also known as hedonic regression. In this case, the model decomposes the price of a mobile phone into its characteristic. This hedonic model is required to calculate the price index using the double imputation method.

Figure 1 reveals that this study is closely related to the process of producing CPI in BPS. CPI incorporated the price changes of hundreds of commodities that compose it, one of them is the mobile phone. This study generates a price index for mobile phones measured based on online transactions. The hedonic price index for mobile phone as a result of this study could complement the CPI calculation in BPS to reflect the share of online price movement.

Figure 2 illustrates the step-by-step CPI calculation process in SHK Online, an information system for processing CPI in BPS. The index resulted from this study could provide the relative price of a commodity, for this case is the mobile phone. However, there are some aspects of the CPI that are not available in this study: the preliminary survey, weighting, and geographic dimension. Thus, to investigate the best way to incorporate it should be conducted in further research to ensure CPI’s comparability over time.

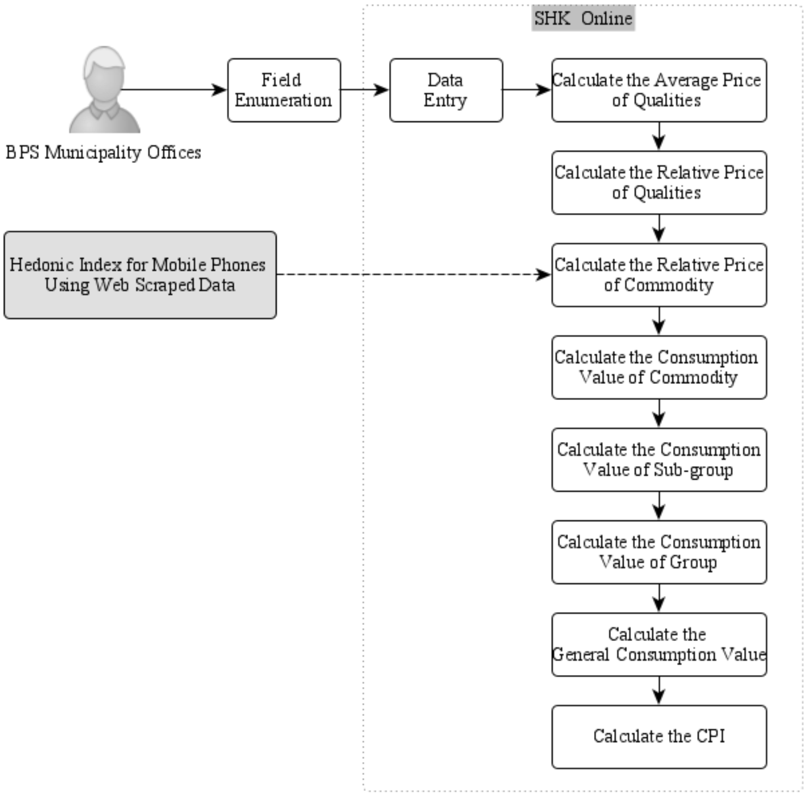

SHK Online
BPS Municipality Offices
Field Enumeration
Data Entry
Calculate the Average Price of Qualities
Calculate the Relative Price of Qualities
Hedonic Index for Mobile Phones Using Web Scraped Data
Calculate the Relative Price of Commodity
Calculate the Consumption Value of Commodity
Calculate the Consumption Value of Sub-group
Calculate the Consumption Value of Group
Calculate the General Consumption Value
Calculate the CPI

Figure 2 Hedonic price index for CPI compilation

2.2 Data Scraping

Figure 3 illustrates the process of data scraping. It begins with website selection, an examination of the website's structure, development of web scraper, and the last is data extraction.

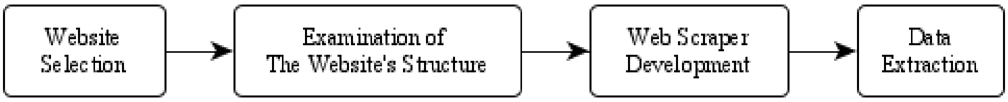


Figure 3 Data scraping process

The first and critical part of the process is selecting the website as the data source. Quality of the data depends on it. Unfortunately, some websites have unstructured HTML code which makes the information intricate to extract. Websites often protect their data from scrapers by applying various methods. Brody (2014) denoted among the best practices to protect website content are: rate the limit of individual access, require a login for access and change the website's HTML regularly. Therefore, feasibility is a primary aspect to consider in website selection.

There are three kinds of information required for the research: types of mobile phone available in the market, its price and specifications. That information is available on a price comparison website. A price comparison website works by collecting product information, including pricing, from participating retailers and then displaying that collective information on a single results page in response to a shopper's search query. Compiling the price index by utilizing big data from price comparison website had also been done before in Japan by Abe and Shinozaki (2018). Yet, relying on third parties such as price comparison website has its drawback because the third party is filtering and altering the samples from the beginning. Nonetheless, collecting data directly from e-commerce would require enormous effort, particularly for matching the products across e-commerce. Therefore, the researcher decide to choose a price comparison website as the data source in this study.

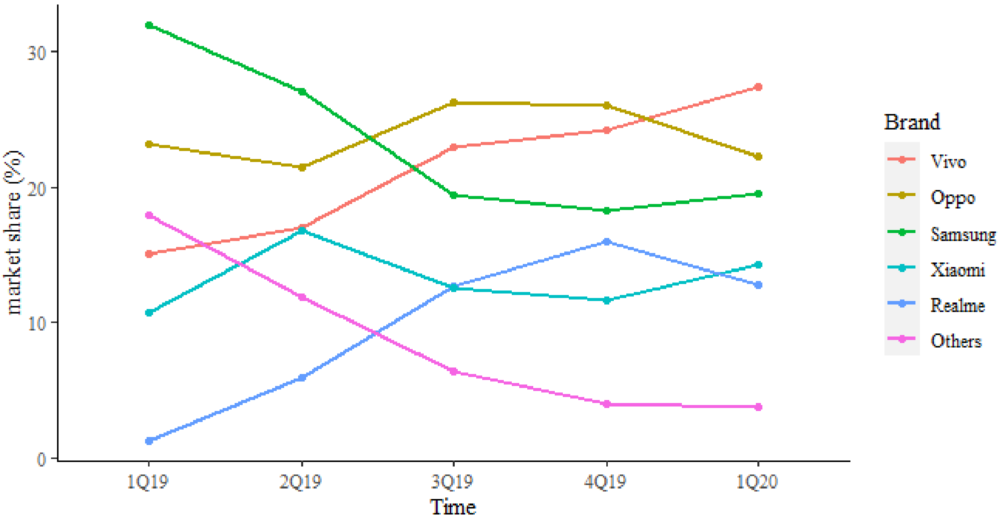
The second step is examining the chosen website structure by identifying the tags containing the desired information in the HTML code. Every web scrapers are usually uniquely built to explore one website. Preparation for data extraction includes investigating the structure of the website and examining how to extract the information needed. Websites are build using HTML (Hypertext Markup Language), along with CSS (Cascading Style Sheets) and JavaScript. HTML elements are separated by tags and directly introduce content to the web page (Oheix 2018). The tags in the website's HTML code which contain the information were identified to build a web scraper.

The third is developing a web scraper. The web scraper was developed using python language with BeautifulSoup package. It was used continuously throughout the research period. However, sometimes a minor modification was needed if there are changes on the HTML structure of the website which disturb the data extraction.

The last step of data scraping is data extraction. A schedule was made to get the data series. The data extraction was conducted weekly and the schedule was every Wednesday from January to June 2020. The total number of datasets obtained in that period is 26.

In this study, the website chosen is IPrice.co.id (further referred to as IPrice). Listianingrum and Nefriana (2020) found that IPrice can be relied on for collecting online price data for a product from various e-commerce sites without the need to visit them one by one. Nonetheless, IPrice alone was proven to be lacking in providing sufficient data for the compilation of hedonic indices. Sometimes there were problems in the completeness and consistency of the product's specifications.

Acquiring data is an important part of this research. In IPrice, the information displayed is very dynamic and updated in real-time. Working with big data such as this means dealing with high volume, velocity and variety data. The art is to find a pattern in the complexity instead of the being trap in the noises. Three aspects related to data scope are considered to obtain representative samples. It includes which kinds of mobile phones would be taken, which e-commerce would be selected for its price to be obtained and how broad is specification details would be recorded.



Source: IDC (2000)

Figure 4 Indonesia’s market share for mobile phone from 1^{st} quarter of 2019 to 1^{st} quarter of 2020

The first aspect is on how to limit the mobile phone products to be taken for the sample. The amount of mobile phone data displayed the IPrice fluctuates. As an overview, there were 9937 mobile phone products from 25 brands displayed on September 27th 2020. It reflects the amount of data at other times. Working with all brands marketed in Indonesia would be ineffective since the majority of the market share only contributed by several vendors. According to research by Canalys, top smartphone vendors in the 3rd quarter of 2019 occupied by Oppo, Xiaomi, Samsung, Vivo and Realme with total market share 94% (Khoirunnisa 2019). International Data Corporation (IDC) also published Indonesia’s market share for mobile phone from 1st quarter in 2019 to 1st quarter in 2020 in Figure 6. The information displayed in Figure 6 supports the decision to choose the five brands as the target samples.

The second aspect to consider is the selection of e-commerce to get the product’s price information. The idea is to get as many samples as possible, but also to maintain comparability. The types of e-commerce that appear on IPrice pages might vary between times. Therefore, the selected e-commerces are those which appears at the beginning of the scrapping. If in the midway through new e-commerce appears, it will not be recorded to ensure comparability over periods. There are 11 e-commerces in the target sample: Shopee, Tokopedia, Blibli, Lazada, Lazmall, Tokopedia, Arjuna Electronics, Personal Digital, Bukalapak, Amazon, and Blanja. For each e-commerce displayed in a product, a maximum of 4 prices would be gathered. IPrice insight reported in the 2nd quarter 2020 Shopee, Tokopedia, Bukalapak and Blibli are in the five most visited e-commerce in Indonesia respectively. Based on that, the 11 e-commerces chosen are expected to reflect the online market in Indonesia.

The third aspect is the specification detail. Initially, the specifications were taken by scraping from IPrice. The target is the complete field information displayed on IPrice at the beginning of the scraping period. Amongst the fields that are recorder are: screen size, screen resolution, density, RAM, internal memory, battery capacity, and many other features information. Throughout the journey, if a new field of specification appeared would be neglected. However, due to some issues, specification data cannot be used directly. Instead, it is used as the basis for creating a specification database. Table 1 depicts example of entries of the scraped data. The complete structure of the scraped data is available in Appendix 1.

Table 1 Example of entries in the scraped data

Product	P_shopee1	…	P_blanja4	RAM	Internal Storage	…	3DTouch
Oppo A1K	1.350.000	...		2GB	32GB	...	
OPPO F5	2.250.000	...				...	No
Samsung Galaxy A20s 64GB Hitam	2.405.000	...		4GB	64GB	...	
Samsung Galaxy M10 Charcoal Black	1.608.000	...		2GB	16GB	...	
Xiaomi Mi Note 2	2.825.000	...	5.900.000	6GB		...	

2.3 Data Cleaning

Appendix 1 shows that the scraped data contains the name of the product (column 1), its prices from various e-commerce (column 2 to 45) and the specifications detail (column 46 to 78). However, Table 1 shows that the entries are not ready to be used in the analysis due to several issues. Names of the products remains unstandardized which would be troublesome for matching over periods. Missing values on price attributes is not a problem, but considerable amounts of missing values on the specifications detail would be problematic. Some attribute entries also have an undesirable format for the analysis; numeric fields are presented in character. For example, the attributes RAM and internal storage which should have numeric entries have a 'GB' affix attached. Data cleaning aims to overcome these issues amongst others.

The first strategy to overcome the missing value on specifications is to fill it with valid information whenever possible. A reliable source is selected to complete the missing value. Each type of mobile phone has a unique set of specifications. We could obtain the information from various websites to fill the missing values in the specifications. This process can be done except for the attributes of RAM and Internal Storage because the size of RAM and Internal Storage in a mobile phone type is not unique; it usually has several combinations of RAM and internal memory.

Table 2 Example of entries in the specification database

No	Type	Brand	Size (inch)	Resolution (pixel)	...	Gyro
1	oppo5x	oppo	5,5	2.073.600	...	Yes
2	oppoa1	oppo	5,7	1.036.800	...	No
3	oppoa12	oppo	6,12	1.094.400	...	Yes
4	oppoa1k	oppo	6,1	1.123.200	...	No
5	oppoa3	oppo	6,2	2.462.400	...	Yes
...	...	...	...	...	...	...
458	samsungz3	samsung	5	921.600	...	No

The cleaning process starts with building a database for product specifications in Table 2. The database is initially established by the information available in scraped data contains specifications of each type of mobile phones recorded. The initial database is then supplemented and validated by the information taken from gsmarena.com. Gsmarena.com is a popular website as a reference for mobile phone information. There are 458 types of mobile phones in the specification database based on the types that are available in the market in the scraping period. The complete structure of specification database is available in Appendix 2.

The following is step by step data cleaning process.

1. Merge data from the all scraping period by adding the scraping date as the “Date” attribute.
2. Delete all of the attribute for specifications, except “RAM” and “Internal Storage”.
3. Create the “Type” attribute by extracting string information indicated the mobile phone type from the “Product” attribute. This variable represents types of mobile phone regardless the color, size of RAM and internal storage variation. This step has to eliminate the additional information from the product name to produce a unique entry for every type of mobile phone. It is done by tokenization of the product name and identify unnecessary word to delete and then regrouping the remaining words without space. This attribute is particularly important to be a link to the specification database. Table 3 depicts the difference of the “Type” attribute from the displayed product name.

Table 3 Example of entries in Product and Type attributes

No	Product	Type
1	Oppo Realme 2 Pro 64GB Blue Ocean	opporealme2pro
2	Samsung Galaxy M30s Putih	samsunggalaxym30s
3	Samsung Galaxy S9 64GB Midnight Black	samsunggalaxys9
4	vivo V5s Baru Emas	vivov5s
5	Xiaomi Mi 4c 16GB Putih	xiaomimi4c

4. Identify the rows in which information about RAM or Internal Memory is missing. Investigate the possibility of obtaining this information and put it on the table. There are some ways to impute the values, sometimes the information

could be available in name of the product, as illustrate in Table 3 row 1,3 and 5. Another possibility, if the product happens to have only unique combination of RAM and Internal Storage than it could be obtained from the information in gsmarena.com. This step is done by manual checking for every row.

5. Delete rows on which information about RAM or internal storage is still missing and cannot be obtained. This step would eliminate quite number of rows from the dataset. However, since RAM and internal storage is a mandatory information, this step has to be taken.
6. Create attribute “ID” by combining “Type”, “RAM” and “Internal Storage” columns. ID is the basis for comparing prices over time as well as determining observations for the regression analysis. Table 4 shows how the “ID” attribute is created.

Table 4 The formation of the ID attribute

No	Product	Type	RAM (GB)	Internal Storage (GB)	ID
1	OPPO A1	oppoa1	4	64	oppoa1_4_64
2	Oppo A1K	oppoa1k	2	32	oppoa1k_2_32
3	Oppo A1K Hitam	oppoa1k	2	32	oppoa1k_2_32
4	OPPO A3s 16GB Merah	oppoa3s	2	16	oppoa3s_2_16
5	OPPO A3s 32GB Purple	oppoa3s	3	32	oppoa3s_3_32

7. Grouping data by “ID” and “Date” attributes and summarize the price median from all e commerce. The median is taken as the price benchmark for each ID in a scraping period.
8. Create final dataset by fully join dataset from step 7 with specification database. Example of the final dataset is displayed in Table 5. The final dataset contains information from scraped data (column 1 to 5) and specifications information from specifications database (column 6 to 43). The complete variables on final dataset is displayed in Appendix 3.

Table 5 Example of entries in the final dataset

Date	ID	Median (Rupiah)	RAM (GB)	Internal Storage (GB)	…	Gyro
20200101	oppo5x_6_64	10.079.000	6	64	…	Yes
20200122	oppo5x_6_64	9.999.000	6	64	…	Yes
20200325	oppo5x_6_64	9.199.000	6	64	…	Yes
20200304	oppo5x_6_64	9.999.000	6	64	…	Yes
20200401	oppo5x_6_64	9.999.000	6	64	…	Yes

2.4 Hedonic regression

Hedonic price model decomposes the price of a product into respective components that determine the product price (Martinez and Garmendia 2010). The basic approach in constructing hedonic models is using linear regression with price as the dependent variable and attributes as independent variables. However, Gu and Xu (2017) found that the functional form of hedonic models are varied, and include the classic linear feature model, the logarithmic price model, the semi-logarithmic price model, and the semi-parametric model. The general format of the hedonic price regression is a semilogarithmic model as follow:

$$Ln(P) = \beta_0 + \beta_i Z_i + e_i$$

Ln(P) is the natural logarithms of price which in this model is represented by its median, β_0 is the constant coefficient, β_i is the regression coefficient of characteristic variables, Z_i refers to the characteristic variable and e_i refers to the error term. Estimation for parameters initially used the Ordinary Least Squares (OLS) method. In the case of linear regression, and when the residuals are normally distributed, the two methods OLS and Maximum Likelihood Estimation (MLE) lead to the same optimal coefficients.

OLS method assumes the error term has zero mean $[E(e_i) = 0]$, constant variance $[Var(e_i) = \sigma^2]$, and zero covariance $[Cov(e_{i,}e_j) = 0]$ (Gujarati and Sangeetha 2007). A linear regression model that involves multiple explanatory variables has an additional condition to be met, i.e. the absence of multicollinearity between explanatory variables. The presence of high collinearity among the independent variables can increase the standard error of the estimated regression coefficients. The model is tested for multicollinearity using variance inflating factor (VIF).

In regression analysis the use of least squares method often fails to provide good fits to the data containing outlier and leverage (Maronna et al. 2019). Outlier is an observation whose dependent variable value is unusual given its value on the predictor variables. On the other hand, an observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviate from its mean (Rousseeuw 1984).

Amongst the methods to diagnosis outliers is studentized residuals. Studentized residuals can be obtained as,

$$Studentized\ Residual = \frac{e_i}{S_{(i)}\sqrt{1 - h_{ii}}}$$

$S_{(i)}$ is the standard deviation of the residuals where ith observation is deleted and h_{ii} is the ith diagonal entry in the hat matrix, $H = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. If a studentized residual exceed +2 or -2 the observation is an outlier. h_{ii} is also a measurement for leverage. When a leverage > 2p/n then it is a matter of concern.

Cook’s distance (or Cook’s D) is another measure for diagnosing outlier. It can be defined as:

$$Cook's\ D = \frac{{e_i}^2}{pMSE}\left(\frac{h_{ii}}{(1 - h_{ii})^2}\right)$$

with p is the number of parameters to be estimated and MSE is the mean square error of the regression model. If Cook's D > 4/n, the observation is an outlier.

Modern statistical theory provides an alternative to outlier rejection, in which outlying observations are retained but given less weight. This approach is known as robust statistics. The well-known methods of robust estimation are M (Maximum Likelihood) estimation, S (Scale) estimation and MM (Multistage Method) estimation. Das *et al.* (2015) found that when outliers present in data OLS estimation gives very misleading result while M and MM estimation do a better job. Second, for data with leverage points, OLS and M estimation yield misleading outputs, whereas, MM estimation gives expected results. Finally, OLS and M estimation gives deceptive results for data with both leverage points and outliers, but MM estimation gives the best results.

MM estimation was introduced by Yohai (1987). MM estimation procedure is to estimate the regression parameter using S estimation that minimizes the scale of the residual from M estimation and then proceeds with M estimation. MM estimation aims to obtain estimates that have a high breakdown value and more efficiency. MM estimator is the solution of

$$\sum_{i=1}^{n} \rho_1(u_i) Z_{ij} = 0 \;\; or \;\; \sum_{i=1}^{n} \rho_1 \left(\frac{Y_i - \sum_{j=0}^{k} Z_{ij} \hat{\beta}_j}{s_{MM}} \right) Z_{ij}$$

where s_{MM} is the standard deviation obtained from the residual of S estimation and ρ is a Tukey's biweight function:

$$\rho(u_i) = \begin{cases} \frac{{u_i}^2}{2} - \frac{{u_i}^4}{2c^2} + \frac{{u_i}^6}{6c^2}, & -c \le u_i \le c; \\ \frac{c^2}{6}, & u_i < -c \; or \; u_i > -c. \end{cases}$$

Susanti et al (2014) stated the algorithm to compute MM estimator are:

1. Estimate regression coefficients on the data using the OLS.
2. Test assumptions of the classical regression model.
3. Detect the presence of outliers in the data.
4. Calculate residual value $e_i = y_i - \hat{y}_i$ of S estimate.
5. Calculate value of $\hat{\sigma}_i = \hat{\sigma}_s$.
6. Calculate value $u_i = \frac{e_i}{\hat{\sigma}_i}$
7. Calculate weighting value of the MM method

$$w_i = \begin{cases} \left[1 - \left(\frac{u_i}{4.685} \right)^2 \right], & |u_i| \le 4.685; \\ 0, & |u_i| > 4.685. \end{cases}$$

8. Calculate $\hat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$ using weighted least square (WLS) method with weight w_i

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{MM}} = (\boldsymbol{Z}'\boldsymbol{W}^{\boldsymbol{MM}}\boldsymbol{Z})^{-1}(\boldsymbol{Z}'\boldsymbol{W}^{\boldsymbol{MM}}\boldsymbol{y})$$

9. Repeat steps 5-8 to obtain convergent value of $\hat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$
10. Test to determine whether independent variables have significant effect on the dependent variable.

In this study, the MM estimator algorithm process was carried out using rlm() function of the MASS R software package.

2.5 Double imputation method

The Double Imputation method is adapted from hedonic price index practice for personal computers conducted by Trewin (2005) for the Australian Bureau of Statistics (ABS). The researcher choosed this method because it had been assessed against several concerns regarding both the indirect and direct methods for hedonic price indices. The double imputation price index, distinguish the extrapolation concern, improving predictive power, remove the issue of revision of earlier data, increase sensitivity to changes in characteristics, and providing transparency to users.

There are 3 kinds of hedonic price index proposed in the paper hedonic price index practice for personal computers for ABS by Trewin: indirect hedonic price index, direct hedonic price index and double imputation index. An indirect hedonic price index uses the hedonic function to predict prices, which in turn are included in more traditional index number formulae. Hedonic imputation is one particular application of an indirect hedonic price index. A direct hedonic price index uses the hedonic function to directly determine a price index. The result of the regression modelling process is itself an index number, which measures the price change from the current period over some base period. The double imputation index is a combination of the direct and indirect approaches

The double imputation index uses a direct approach to determine a price change over two consecutive periods, and in a manner very similar to indirect approaches this imputed price change is then used to impute price movements for both new and superseded models. These price movements are combined with matched model price movements 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 to two consecutive periods: t-1 and t. A sample of products from each period is taken (Sample / S). Some products will be available in the earlier period, but not the latter (Death / D). Other products will be available in the latter period, but not the earlier (Birth / B). There will also be a set of products which are available in both periods (Matched / M). This situation is illustrated in Figure 5.

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

Figure 5 Illustration of the samples from period 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 multi-period pooled approach used in direct approach 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 is instead used in later stages of the calculation.

If we consider quantitative explanators $(z_{1i}, \dots, z_{mi})$ and indicator variables $(z_{m+1i}, \dots, z_{ki})$ the models took the form of:

$$\ln(\mathrm{p}_i^{\tau}) = \hat{\beta}_0 + \hat{\beta}_1 \ln z_{1i} + \cdots + \hat{\beta}_m \ln z_{mi} + \hat{\beta}_{m+1} z_{m+1,i} + \cdots + \hat{\beta}_k z_{ki} + d_{ti} + \varepsilon_i \qquad (1)$$

with: $\tau = t-1, t$ and $d = \left\{ \begin{array}{l} 1, i \in\ S^{t} \\ 0, i \in\ S^{t-1} \end{array} \right.$

For this process, a regression is run each period to determine a parsimonious form by previously conducting a variable selection process, and to estimate the parameters $\beta_0, \dots, \beta_k, \delta$.

From equation (1), the prediction for the prices from perion t-1 is given by:

$$\ln(\hat{\mathrm{p}}_i^{t-1}) = \hat{\beta}_0 + \hat{\beta}_1 \ln z_{1i} + \cdots + \hat{\beta}_m \ln z_{mi} + \hat{\beta}_{m+1} z_{m+1,i} + \cdots + \hat{\beta}_k z_{ki} \qquad (2)$$

And prices from period t are estimated by:

$$\ln(\hat{\mathrm{p}}_\mathrm{i}^{\mathrm{t}}) = \hat{\beta}_0 + \hat{\beta}_1 \ln \mathrm{z}_{1\mathrm{i}} + \cdots + \hat{\beta}_\mathrm{m} \ln \mathrm{z}_{\mathrm{mi}} + \hat{\beta}_{\mathrm{m}+1} \mathrm{z}_{\mathrm{m}+1,\mathrm{i}} + \cdots + \hat{\beta}_\mathrm{k} \mathrm{z}_{\mathrm{ki}} + \hat{\delta}_1 \qquad (3)$$

Taking the difference from equation (2) and (3),

$$\ln\left(\hat{\mathrm{p}}_\mathrm{i}^{\mathrm{t}}\right) - \ln\left(\hat{\mathrm{p}}_i^{t-1}\right) = \hat{\delta}_1$$

$$\ln\left(\frac{\hat{\mathrm{p}}_\mathrm{i}^{\mathrm{t}}}{\hat{\mathrm{p}}_i^{t-1}}\right) = \hat{\delta}_1$$

$$\frac{\hat{\mathrm{p}}_\mathrm{i}^{\mathrm{t}}}{\hat{\mathrm{p}}_i^{t-1}} = \exp\,(\hat{\delta}_1)$$

A hedonic two period pooled time dummy price index at period t with the price at period t-1 is given by:

$$I_{2TD}^{t-1,t} = \exp(\hat{\delta}_1) \times 100$$

The matched model price index from the matched sample is:

$$I_{MM}^{t-1,t} = \prod_{\mathrm{i} \in S^{t-1} \cap S^{t}} {\frac{\hat{\mathrm{p}}_\mathrm{i}^{\mathrm{t}}}{\hat{\mathrm{p}}_i^{t-1}}}^{1/N_{S^{t-1} \cap S^{t}}}$$

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 of price from period t-1 to period t is then a weighted geometric mean of the two component indices is described as equation (4).

$$I_{DI}^{t-1,t} = \left[I_{MM}^{t-1,t}\right]^{f_m} \left[I_{2TD}^{t-1,t}\right]^{1-f_m} \qquad (4)$$

$\boldsymbol{f_m}$ 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.

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

The processes described above are represented in Figure 6.

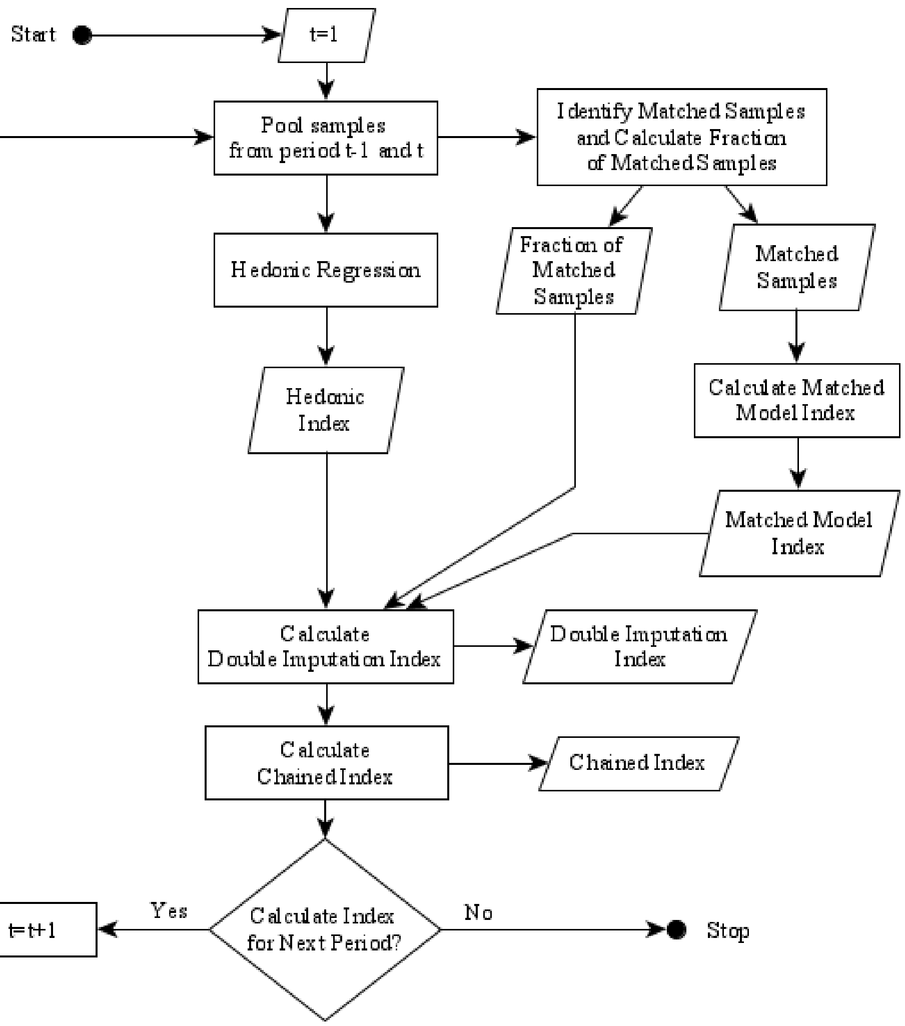

Start
t=1
Pool samples from period t-1 and t
Identify Matched Samples and Calculate Fraction of Matched Samples
Hedonic Regression
Fraction of Matched Samples
Matched Samples
Hedonic Index
Calculate Matched Model Index
Matched Model Index
Calculate Double Imputation Index
Double Imputation Index
Calculate Chained Index
Chained Index
t=t+1
Yes
Calculate Index for Next Period?
No
Stop

Figure 6 Flowchart of the double imputation method

III RESULT AND DISCUSSION

3.1 Data Exploration

Figure 7 shows the result of the scraping process and part of the cleaning process. The number of observations being retrieved on each week's scraping process is depicted with the red line. The blue line indicates the number of observations with complete information as a result of the cleaning process. The amount of data fluctuates greatly over time between 900 to 1300 observations. In general, the amount of scraping results from January to March is higher than from April to June. The weekly average number of observations gathered in the first quarter was more than 1228, while for the second quarter was only 1102. Around 11% of the total retrieved were eliminated each week for having issues such as incompleteness and inconsistency.

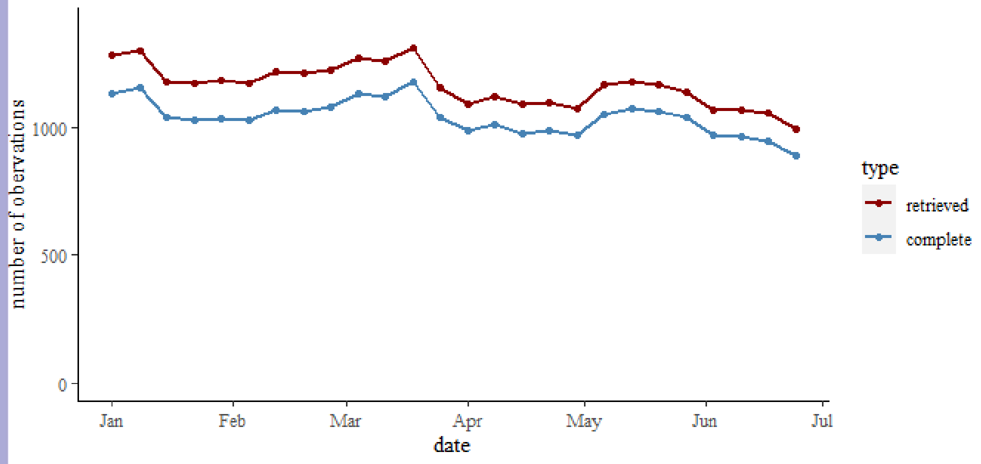


Figure 7 Amounts of weekly scraped data

There is a decreasing trend on the amount of scraped data over the periods as illustrated in Figure 7. In this study, the rise and fall found in scraped data is something that cannot be controlled. However, there was a change in the appearance of the user interface on IPrice at the end of March. This might have caused a sharp decrease in the amount of scraped data at the end of March. During the scraping period, modifications were made to the web scraper three times due to changes in IPrice's HTML structure. Change of HTML structure might be a sign of changes in the algorithm that affect the data displayed.

The observations presented in Figure 7 is not the final dataset for the analysis. That data still contains redundancy on the products recorded in each period. The next process is grouping data by date and ID to eliminate the redundant observation and get a benchmark for the price, which is the median. Figure 8 shows the number of observations of the final dataset.

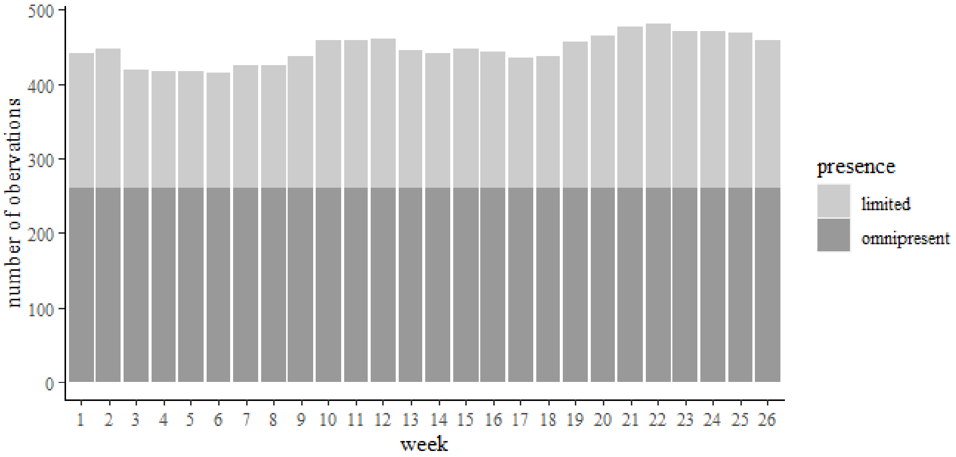


Figure 8 Number of observations per week in the final dataset

The average number of observations per week in the final dataset is 447 observations. The observations are divided into the omnipresent and limited present. Omnipresent refers to the observations in the final dataset which appear in every period while limited means the observations that not appeared in some periods; it may appear in 1 to 25 time periods but not entirely. The number of omnipresent observations is fixed at 260 in all periods. It means that there are 260 mobile phone types with a particular size of RAM and internal storage that always gathered in the scraping process and passed through the data cleaning. The proportion of limited presence observations is relatively high, which is almost 50 percent of the total observations. It suggests that forming an index with price comparisons with a fixed basket of mobile phone for all periods is not sufficient for this analysis.

The size of the final dataset as in Figure 8 is 11617 observations. However, several observations have peculiar value on the price attribute. For example, there is a mobile phone with price reached 144 million rupiahs. The bizarre price information is not valid for the analysis. In this study, it is assumed that the reasonable selling price of mobile phones is below 40 million rupiahs. There are five observations eliminated for it does not meet the condition. The final dataset shrank into 11612 rows in size.

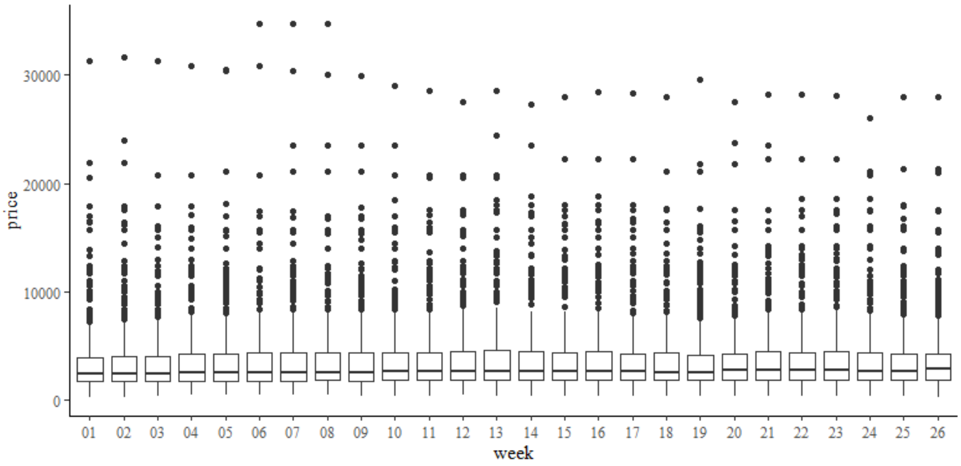


Figure 9 Boxplot of mobile phone price per week (thousand rupiah)

Figure 9 shows the distribution of price (in thousand rupiahs) per week based on the filtered data. It appears that the majority of mobile phone price lies under 5 million rupiah. Even so, many mobile phones are far more expensive than others which are indicated by the appearance of an upper outlier. This appearance of outlier designates that the price data is right-skewed.

Figure 10 exposes comparisons of the weekly mean of price movement to the mean of several specifications which comprise the quality of a mobile phone: the size of RAM, size of internal storage and resolution of the main camera. The four charts show an upward trend. There was considerable fluctuation in the mean of prices at the end of March and the end of July. Despite the oscillations, the average price has a general upward trend with the price mean ranging from 3.4 to 3.7 million rupiahs. An upward trend means that the price of mobile phones is getting more expensive over time. However, the price increase is associated with increases in quality. The second graph in Figure 10 shows that in the first week the average of RAM was only around 3.6, while in the last week the average RAM reached 3.8. It means that the quality of mobile phones in terms of RAM size had improved over time. The same phenomenon happened to the internal storage and resolution of the main camera. From the beginning to the end of the scraping period, the average internal storage increased from 59 to 65.9 GB while the main camera increased from 16 to 18.9 MP.

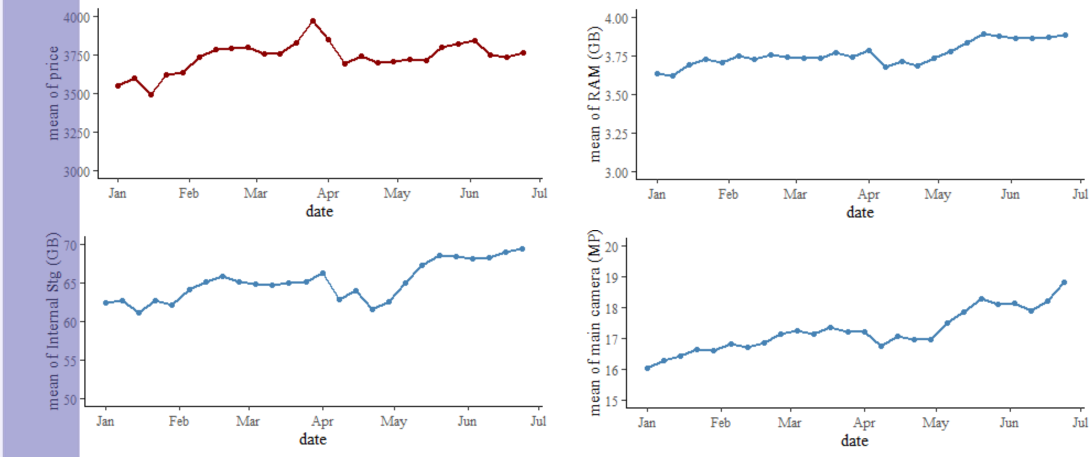


Figure 10 The series of means for price, internal storage, RAM and main camera from January to June 2020

The role of the hedonic method is to remove the impact of the change in characteristics on the price changes. The first graph in Figure 10 suggests that the price of mobile phones is always increasing. However, it might not be the case if we only compare prices for products having equal quality. If the effect of the quality characteristics is incorporated into the hedonic model, a pure price change of mobile phone could be obtained.

Logarithmic transformation is a convenient means of transforming a highly skewed variable into one that is more approximately normal (Feng *et al*. 2012). Figure 9 shows that the price distribution is right-skewed hence using its logarithmic transformation is preferable. Figure 11 illustrates the comparison between the use of actual price and ln(price) as the response variable to one of the explanatory variables: RAM. When using Ln price, boxplot sizes become more balanced for each RAM size. This evidence supports the use of the logarithmic form of price data for modelling the data in this study.

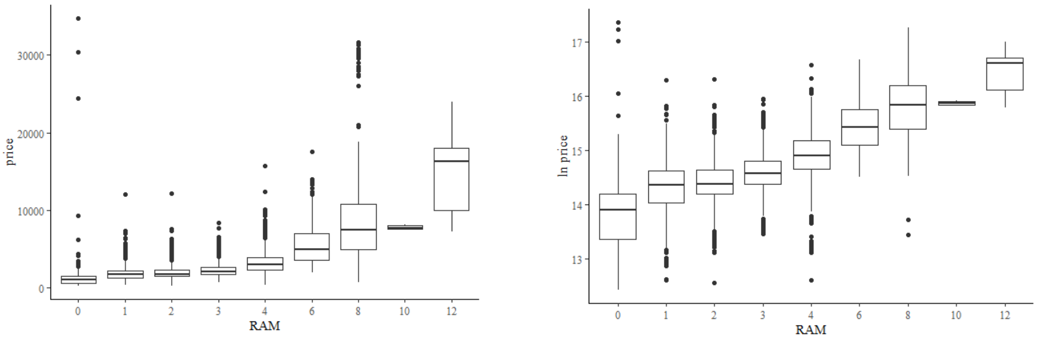


Figure 11 The comparison between price and Ln price as response variable

3.2 Hedonic Regression Analysis

There are two main reasons to estimate a model: prediction and inference. A model that aims for prediction is often treated as a black box, in the sense that the exact form of the model is not the primary concern provided that it yields accurate predictions. However, we are often interested in understanding the way that the response variable is affected as by the change of its predictors. In this situation, we wish to estimate the model, unless, the main goal is not necessarily to make predictions. Instead, we want to understand the relationship between X and Y. Hence it cannot be treated as a black box because we need to know its exact form (James *et al.* 2013). This ultimate goal of hedonic regression analysis in this study is for inference by modelling the online retail price of a mobile phone, based on its specifications such as brand, size of RAM, and size of internal storage.

Though seem less sophisticated than modern statistical learning approaches, linear regression is still useful and widely used. The linear model has distinct advantages in terms of inference and, on real-world problems, is often surprisingly competitive to non-linear methods. Linear models also allow for relatively simple and interpretable inferences although may not yield as accurate predictions as some other approaches.

In the case that too many variables used in a multiple regression model, there could be issues of interpretability. Irrelevant variables lead to unnecessary complexity in the resulting model. By removing these variables, we can obtain a model that is easier to interpret. One of the approaches to eliminate irrelevant variables is subset selection. This approach involves identifying a subset of the predictors that we believe to be related to the response. We then fit a model using least-squares on the reduced set of variables.

There were 40 candidates for explanatory variables in the dataset. Variable selection is then conducted to reduce model complexity. The elimination is carried out using the subset selection approach. This approach is implemented by experimenting literature study and data exploration to get a subset of variables which could explain the variation of the price.

In 2019 Waseem Ahmad, Tanvir Ahmed and Bashir Ahmad studied the effect of smartphone attributes on their retail price using a log-linear hedonic price model in Pakistan. In that study, nine variables were selected as candidates for explanatory variables: RAM, InternalStorage, Brand, Size, Camera, Battery, Weight, NetworkGeneration, and RadioFM. However, because the scraping process had produced much other information which might not exist in the previous study, some candidate variables are added from the scraped data such as Density, Core, NFC, Protection, and Material. The exploration for the best model is done by experimenting with models and applying the stepwise selection method. Some of the criteria considered in choosing a model are avoiding multicollinearity, maximizing the R square adjusted value and keeping the model simple. Stepwise selection leads to the result to eliminate the camera variable out of the model. The final variables used in the model are listed below.

1. Battery
This feature represents battery capacity measured in milliampere-hour (mAh). The functionality of the smartphone is severely affected by battery life. With the introduction of smartphones which equipped with various applications, energy consumption has increased. Figure 12 shows the average battery capacity for mobile phones released in 2011 to 2020. Figure 12 shows that the battery capacity of a mobile phone is increasing over time. However, there is a tradeoff between large battery capacity and weight as the battery capacity increases the weight of the mobile phone.

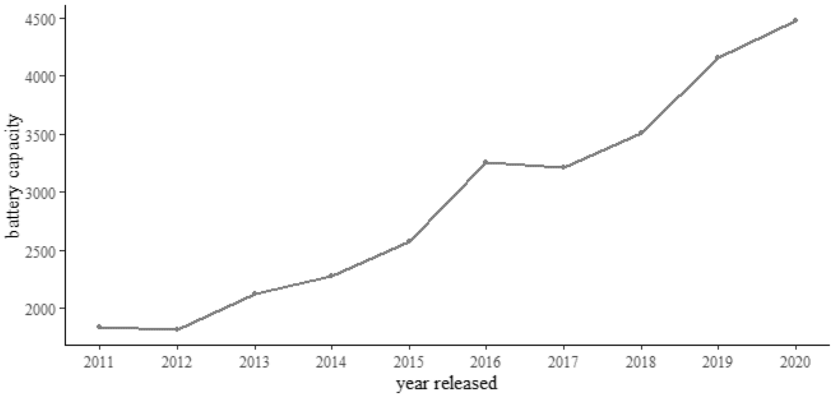


Figure 12 The trend of battery capacity from 2011 to 2020 (mAh)

2. Core
Core refers to the number of cores in the processor. A processor, also known as Central Processing Unit (CPU) may consist of single or multiple cores: Dual, Quad, Hexa, and Octa-core. Processor cores distribute the work that comes in when a phone is in use. One core has a maximum number of instructions it can process within a certain amount of time. If a lot of actions are performed, a queue of sorts will form. If this queue gets too long, part of it will go to the next core. The number of cores could scale up the phone’s performance (Marcel, 2020). Based on the specification database, most mobile phones in 2011 only have a single-core processor, but nowadays most mobile phones’ processor contains eight cores, or what is called hexacore.
3. Density
Density refers to the screen’s pixel density per inch / PPI (pixels per inch). It is the unit of measure used to quantify the number of pixels found on a square

inch surface measured. A higher PPI, or pixel density, means more detailed display on the screen. It means better images, better fonts, smoother lines, or in other words, higher quality (Neagu, 2017). This variable can also represent variable resolution which is not included in the model because the size times the density is the resolution.

4. InternalStorage
   Internal storage is used to store files such as document, pictures and videos. Large memory means more data storage capacity on the phone. According to the data, the maximum size of the mobile phone's internal storage has reached 1000 GB (GigaByte) or to be exact is 1024 GB. The values of this attribute are: 0, 4, 8, 16, 32, 64, 128, 256, 512, 1000.
5. RAM
   RAM is the memory that a phone uses in current working. A phone with more RAM is more responsive to inputs and is good at working in lots of applications simultaneously. The correlation between RAM and the response variable in this study is quite large: 0.745. It indicates that the size of RAM has a positive relationship with the mobile phone's price.
6. Size
   Size refers to the screen size in inch. Screen size relates to display performance. Larger screen enables the user to watch videos, play games, browse the internet conveniently. However, phones having large screen size might be less practical.
7. Weight
   The weight of the phone is measured in gram. Design characteristics like weight usually are reckoned by the users and can influence the demand for a phone. Some people argue that lighter is better. However, there's a sort of stigma that weighty handsets acclaimed to be more durable.
8. Brand
   According to Durianto (2014) people who have brand awareness of a product will tend to choose a brand name that is already known first and the price latter thus it will become a consideration for consumers in a purchasing decision . A study by Paramansyah *et al.* (2020) revealed that brand awareness variable has a significant influence on purchase intention. It means that a customer might be willing to pay more for a trusted phone brand. The brands analyzed in this study are Oppo, Vivo, Realme, Samsung and Xiaomi.
9. Material
   Material refers to the substance constructing the mobile phone body. It is divided into: Metal, Glass, Ceramic, and Plastic.
10. NFC
    By definition from gsmarena, NFC (Near Field Communication) is a short-range high-frequency wireless communication technology that enables the exchange of data between devices over about a 10 cm distance. This feature is gaining popularity that increasingly there are more and more mobile phones launched with the NFC feature. Samsung has the most types of mobile phone with NFC. This variable is classified into 'Yes' for those mobile phones equipped with NFC, and 'No' for the contrary.

11. RadioFM
Mobile phone with RadioFM enables the user to access the FM network. However, this feature seldom appeared in a recently released mobile phone because nowadays FM radio is often accessed via the internet.

The following hedonic model is developed to identify factors determining the mobile phone price.

$$Ln(PRICE_i) = \beta_0 + \sum \beta_1 Battery_i + \beta_2 Core_i + \beta_3 Density_i + \beta_{10} InternalStorage_i \\ + \beta_5 RAM_i + \beta_6 Size_i + \beta_7 Weight_i + \beta_{8i} Brand_i + \beta_{9i} Material_i \\ + \beta_{10i} NFC_i + \beta_{11i} RadioFM_i$$

where:
$Ln(PRICE_i)$ is natural logarithm of price of mobile phone I, and
β_j is regression coefficient of characteristic variable, j=1,2,…,11.

All of the explanatory variables in the hedonic model together significantly affect the price of the smartphone. Adjusted R squared is at 73.38%, means 73.38% of price variation that can be explained by 11 explanatory variables in the model. No VIF value of each variable exceeded 10, i.e. the rule of thumb maximum values (Gujarati and Sangeetha 2007), which means there is no evidence of multicollinearity in the model.

Table 6 The estimates of hedonic model coefficients based on OLS method

Variable	Coef	SE Coef	T-Value	P-Value	VIF
Constant	13,2101	0,0424	311,54	0,000	
Battery	-0,0001	0,0000	-10,85	0,000	4,66
Core	0,0297	0,0027	11,16	0,000	2,75
Density	0,0009	0,0001	16,07	0,000	2,50
InternalStorage	0,0013	0,0001	21,70	0,000	2,29
RAM	0,0931	0,0029	31,73	0,000	4,10
Size	0,0668	0,0119	5,63	0,000	6,75
Weight	0,0046	0,0003	17,77	0,000	3,52
Brand					
Realme	-0,1351	0,0193	-7,01	0,000	1,35
Samsung	0,0909	0,0118	7,72	0,000	2,85
Vivo	0,0561	0,0133	4,22	0,000	1,56
Xiaomi	-0,2404	0,0113	-21,28	0,000	2,44
Material					
Glass	0,0671	0,0115	5,81	0,000	1,56
Ceramic	0,1656	0,0269	6,16	0,000	1,27
Plastic	-0,2099	0,0086	-24,48	0,000	1,57
NFC					
Yes	0,2172	0,0098	22,19	0,000	1,91
RadioFM					
Yes	-0,1353	0,0090	-15,11	0,000	1,45

Table 6 reveals the estimation of the hedonic model’s coefficients. All variables have a significant effect on the price. The rise of every numeric variable except the size of battery capacity are significantly increasing the price of a mobile phone, namely core, density, internal storage, RAM, size and weight. Meanwhile, the coefficient estimation results on dummy categorical variables give quite distinctive results.

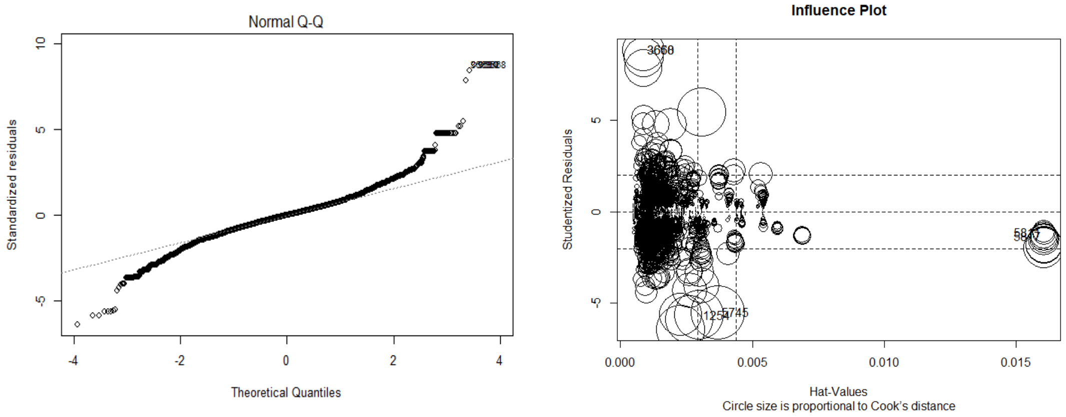


Figure 13 The normal qq plot of residuals and the influence plot of the OLS hedonic model

However, Normal Q-Q Plot of standardized residual in Figure 13 reveals that the residual from the OLS Hedonic Model has a longer or heavier tail than normal. The violation of the normal distribution means that the OLS estimators are no longer MLE estimators. The heavy-tail indicates the presence of outliers in the model which is confirmed by the influence plot in Figure 13. The influence plot reveals the presence of outliers based on the studentized residual, leverage based on hat values and influence observations based on Cook’s distance. These outliers may have a strong influence on the method of least squares in the sense that they “pull” the regression equation too much in their direction (Montgomery, Peck, & Vining, 2012). The presence of outliers in the data could also cause heteroscedasticity. Carroll and Ruppert (1988) discussed the presence of outliers and leverages affects the distribution of data thus the assumption of homoscedastic error variance breaks down and heteroscedasticity sets in. Heteroscedasticity also found in this OLS Hedonic Model which is identified by the Breusch-Pagan test with BP score = 733.63 and p-value < 2.2e-16.

As discussed earlier by Das *et al.* (2015) robust regression with MM estimation gives proper result for data with both leverage points and outliers compared to OLS and M estimation. Therefore, a process to re-estimate the parameters of OLS using robust regression models with MM estimates was conducted. Table 7 shows that the estimated parameters of the OLS and MM estimator results are generally very close, except for the dummy variables of Samsung and Vivo brand which shows opposite direction. MM estimation gives more efficient estimation for the parameters in the model compared to OLS.

The results of the study show that mobile phones with higher battery capacity do not necessarily have higher prices. The battery capacity represents the amount of power the battery can hold. A mobile phone is an essential item for daily lives, but its’ usefulness lasts as long as there is power in it. Higher performance requires more power. However, the battery life depends on the types of hardware inside the smartphone from the processor to display and any other features packed into it.

Table 7 The estimates of hedonic model coefficients based on OLS vs MM estimation methods

Variable	OLS		MM estimation	
	Coef	SE Coef	Coef	SE Coef
Constant	13,2101	0,0424	13,3073	0,0365
Battery	-0,0001	0,0000	-0,0001	0,0000
Core	0,0297	0,0027	0,0189	0,0023
Density	0,0009	0,0001	0,0012	0,0001
InternalStorage	0,0013	0,0001	0,0014	0,0001
RAM	0,0931	0,0029	0,0901	0,0025
Size	0,0668	0,0119	0,0291	0,0102
Weight	0,0046	0,0003	0,0053	0,0002
Brand				
Realme	-0,1351	0,0193	-0,1525	0,0166
Samsung	0,0909	0,0118	-0,0107	0,0101
Vivo	0,0561	0,0133	-0,0144	0,0115
Xiaomi	-0,2404	0,0113	-0,2961	0,0097
Material				
Glass	0,0671	0,0115	0,0841	0,0099
Ceramic	0,1656	0,0269	0,1559	0,0232
Plastic	-0,2099	0,0086	-0,1962	0,0074
NFC				
Yes	0,2172	0,0098	0,2691	0,0084
RadioFM				
Yes	-0,1353	0,0090	-0,0793	0,0077

There are several primary items to mobile phone performance, amongst them are the number of cores in the processor and the RAM size. Together, they determine how responsive a smartphone behaves. Each item determines the performance in a different dimension. The number of cores relates to the number of tasks which could be executed at a time, while RAM provides memory from where the processor derives commands. Therefore, the bigger RAM and cores mean more commands could be accessed by the processor and more process could be done at one time respectively.

Phones are available in a wide variety of display sizes. However, the quality of the display depends on its density. Table 7 concludes that bigger and clearer display is still one of the main factors that increase the price. There are several reasons for the preference of large-sized screen size. Firstly, these phones are great for reading e-books, watching videos and running more applications side by side. Secondly, they are helpful in better typing. Thirdly, mobile users who want to use their device for seeking information, i.e. internet browsing, are more efficient if they have a device with a large screen size (Ahmad, *et al.* 2018).

People are fond of the phone’s ability to store more data such as text, games, photos, videos, and music on the handset. Internal storage has a significant positive effect on phone prices. The results obtained in this study are consistent with the findings of Montenegro and Torres (2016). They confirm that other things being equal, storage size had a significant influence on mobile phone prices.

Some mobile phone attributes like the battery, screen size, etc. add to the weight. In this study, weight is highly correlated with other variables such as screen size and battery capacity. However, the VIF value indicated no multicollinearity between those variables. It is surprising to find in Table 7 that weight has a significant positive effect on the price. However, the weight could be a sign of quality phone to represents the specifications not covered in the study.

Mobile phone buyers often use the brand as a proxy for quality (Ahmad et. Al. 2018). OLS and MM estimation gives different result on the brand. OLS shows that given the same quality, mobile phone from Samsung and Vivo brands have higher prices than those from Oppo brands while Realme and Xiaomi are cheaper than Oppo while MM estimation shows that Oppo is the highest price of all. It indicates that Oppo, Samsung and Vivo are the premium brands for smartphones in Indonesia. As for customers, buying a phone from the brands Realme and Xiaomi will save money than to buy it from other brands. Although it seems strange why people want to buy a mobile phone of the same quality at a higher price from a particular brand, there is an explanation for this. It is due to some other aspects that are not counted in the model like the buyer’s familiarity and the use of brand as a proxy for status symbol.

Mobile phones with plastic materials are the cheapest compared to those with glass, ceramic and metal materials. New technology like NFC significantly increases the price of a mobile phone. What is strange is the FM radio feature which has a negative effect on the price. It is because radio features are not considered crucial for mobile phones to have in the current era.

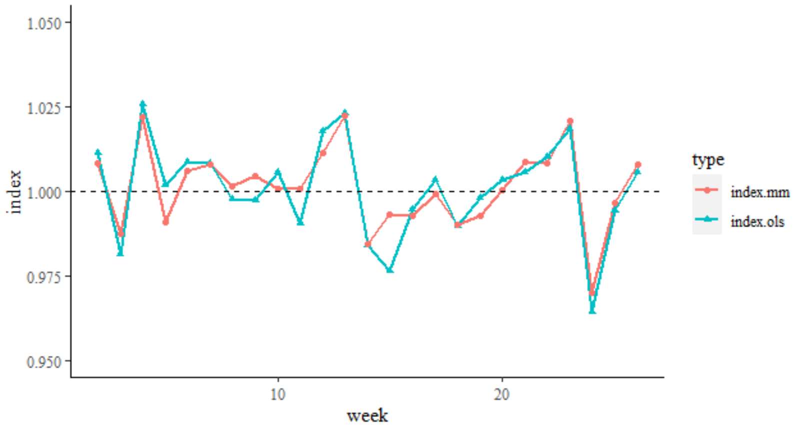


Figure 14 The comparison of hedonic price index based on OLS and MM estimation method

Figure 14 reveals that hedonic price index based on the OLS model is more volatile than the one from robust regression with MM estimation. The standard deviation for the OLS index is 0.014 while for MM index is 0.012. To get a robust estimate of the coefficient, the MM index is chosen to be calculated in the double imputation method.

3.3 Double Imputation Index

The double imputation index calculation uses the method described in Figure 6. It illustrates that the regression model is only part of the calculation process. Equation (4) reveals that the double imputation price index measuring price change of price from period t-1 to period t is then a weighted geometric mean of the two index components: the matched model index and hedonic index. The influence magnitude of the regression model on the double imputation index value is determined by the proportion of unmatched observations in each pair of periods.

Table 8 represents the percentages of matched and unmatched observations from two consecutive periods, which needed for calculating the double imputation index. The table shows that the percentage of unmatched observations in each pair of periods is relatively small with an average of 7.41%. It means that the hedonic index value obtained from the regression model will not have much effect on the resulting final index value.

Table 8 Percentages of matched and unmatched observations in the dataset from two consecutive periods

No	Period	Pair	Matched	Unmatched	% Matched	% Unmatched
1	1-2	453	435	18	96,03	3,97
2	2-3	454	413	41	90,97	9,03
3	3-4	423	414	9	97,87	2,13
4	4-5	420	415	5	98,81	1,19
5	5-6	424	410	14	96,70	3,30
6	6-7	428	414	14	96,73	3,27
7	7-8	429	421	8	98,14	1,86
8	8-9	441	419	22	95,01	4,99
9	9-10	464	430	34	92,67	7,33
10	10-11	467	450	17	96,36	3,64
11	11-12	467	452	15	96,79	3,21
12	12-13	466	439	27	94,21	5,79
13	13-14	463	424	39	91,58	8,42
14	14-15	471	418	53	88,75	11,25
15	15-16	465	424	41	91,18	8,82
16	16-17	461	416	45	90,24	9,76
17	17-18	462	409	53	88,53	11,47
18	18-19	475	417	58	87,79	12,21
19	19-20	490	430	60	87,76	12,24
20	20-21	509	431	78	84,68	15,32
21	21-22	500	456	44	91,20	8,80
22	22-23	498	452	46	90,76	9,24
23	23-24	491	449	42	91,45	8,55
24	24-25	494	444	50	89,88	10,12
25	25-26	486	441	45	90,74	9,26

Figure 15 compares the matched model index to the double imputation index. It shows that the lines between the matched model (blue line) and the double imputation (red line) indices tend to coincide. Although, in some parts, there is a small gap between the two lines. This gap implies that there is a correction factor for the index resulting from the matched model. The matched model only considers products that appear in 2 periods, while the double imputation considers the death and birth of the products on the market. In this context, death means the Mobile phone IDs that are available in the earlier period, but not the latter while birth means Mobile phone IDs that are available in the latter period, but not the earlier.

Figure 15 shows that the prices of mobile phones from January to June 2020 has a flat trend even though there are large fluctuations at several points. It differs from the series of the average price shown in Figure 10 that specifies that there is an uptrend in the average mobile phone price around those periods. It means that the price increase in figure 10 happened simultaneously with an increase in the quality of the mobile phone. If a comparison of price carried out by paying attention to the quality, then the price of mobile phones tends to be stable.

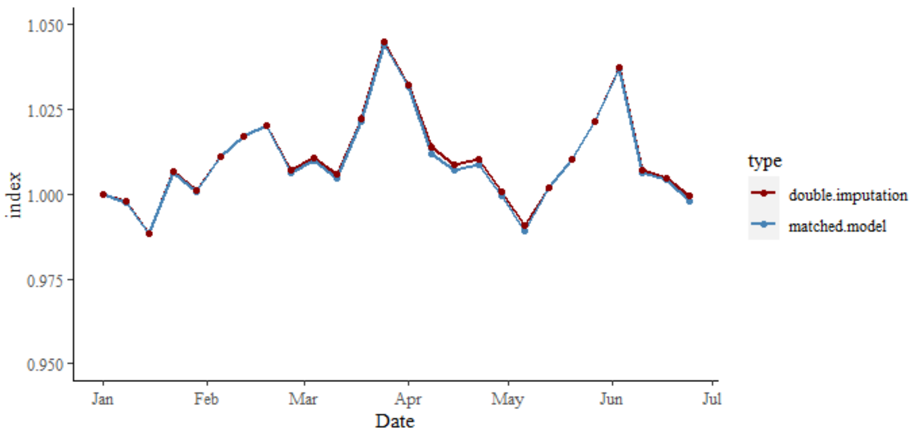


Figure 15 The comparison between the double imputation and matched model indices

There are two peaks identified in Figure 15 that appears at the end of April and early June. This result indicates that the price of mobile phones increases substantially in both periods. However, a dramatical increase may also be a product of noises in the scraping results. As previously explained, choosing big data as a data source comes with a risk, because big data tends to contain a lot of noises. Besides, choosing IPrice as the source of sample selection also has the peril to get a set sample that is unrepresentative because IPrice has already altered it in the first place. Figure 7 also shows that there was a sharp decline in scraping results during these peak periods that are suspected could affect the index fluctuation.

Identification is carried out to the of scraping results. It turns out that many old types of mobile phones appear in scraping results. Due to the rapid development and innovation in mobile phone products, these old mobile phones would have a large technology gap with the newest one. People tend to choose the latest released mobile phone over the very old one.

Figure 16 delineates the number of mobile phone products each week with colors indicates the year release of the product. The results seem to show a reassuring pattern that the number of phones released in 2020 is increasing over time. The majority of observations are the mobile phone released from 2017 to

2019. However, there is a sizable proportion of mobile phones released in 2016 and below. There are even mobile phones released in 2011 recorded in the scraping results. IPrice displays the data based on the price displayed on the e-commerce page. However, there is no guarantee that it is an up-to-date page. The bottom line is that the sales of old types of mobile phones might not reflect the actual situation in the market.

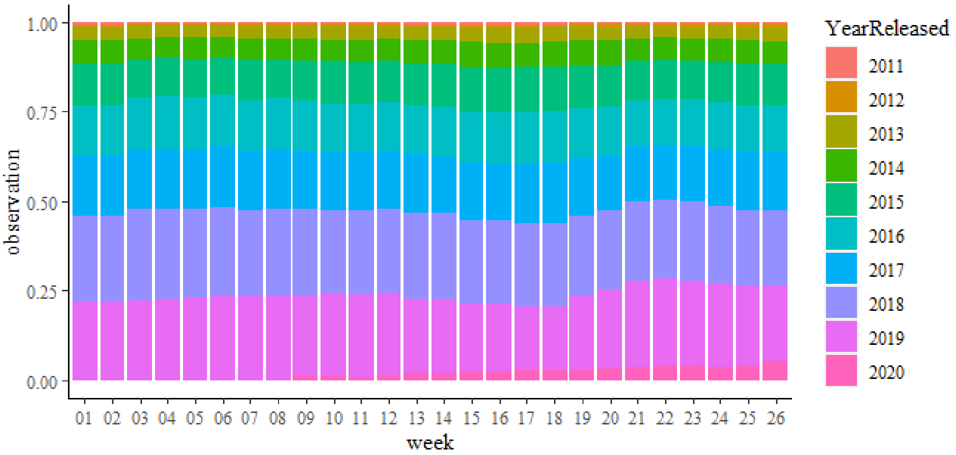


Figure 16 Number of weekly observations by year release of the mobile phone

The problematic drawback of scraping data is the absence of quantity information. Quantity is crucial to acknowledge which products are sold more in the market to serve as a weighting basis. With no weighting, each type of mobile phone that appears will be treated as equal, even though their contribution to the price increase is different. Because of that, several weighting scenarios were carried out. The first one takes into account is the release year mobile phone previously described. The second is by using market share as a proxy for weighting. And the third is the combination of both.

Scenario 1: Weighting by YearReleased

The idea of this weighting scenario is to give more weight to mobile phone from later release, assuming there are more transactions for these types of mobile phones rather than the old one. Unfortunately, any reference for determining the weighting value based on the release year have not been found. The researcher therefore only holds to the idea that older phones should have less weight than the newer. The observations were divided based on the variable YearReleased into four groups and then assigned the weight values in the following Table 9.

Table 9 Weighting scenario 1

Group	YearReleased	Weight
1	>=2019	4/10
2	2018	3/10
3	2017	2/10
4	<=2016	1/10

Figure 17 shows the double imputation index using weighting scenario one. The weighting scenario one based on mobile phones' release year was not very influential at the beginning. Nevertheless, in the second quarter, the index produced with weighting was lower than the one without weighting. It means that if the release year of mobile phone is weighted, it turns out that the increase in mobile phone prices in the second quarter is not as high as that measured in double imputation method without weighting.

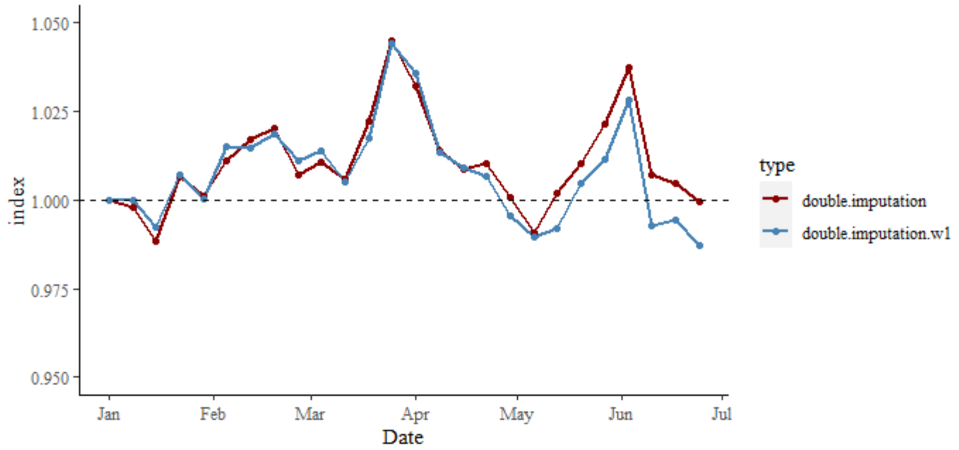


Figure 17 The comparison of hedonic price indices of the mobile phone between the use of weighting scenario 1 and without weighting in double imputation method

Scenario 2: Weighting by Brand

In Figure 18, we can see that the majority of the mobile phones in the dataset are from the Xiaomi and Samsung brands, in contrast, the are very few mobile phones from Realme brand compared to the other brands. Whereas, Figure 4 shows that Realme sum 12.8% of the total market share for mobile phone in Indonesia. If weighting by the brand is not performed, the major influence to the index value would be from the price movements of the Xiaomi and Samsung mobile phones. With weighting scenario 2, the calculation can be evaluated by assigning the right proportion for each brand’s contribution.

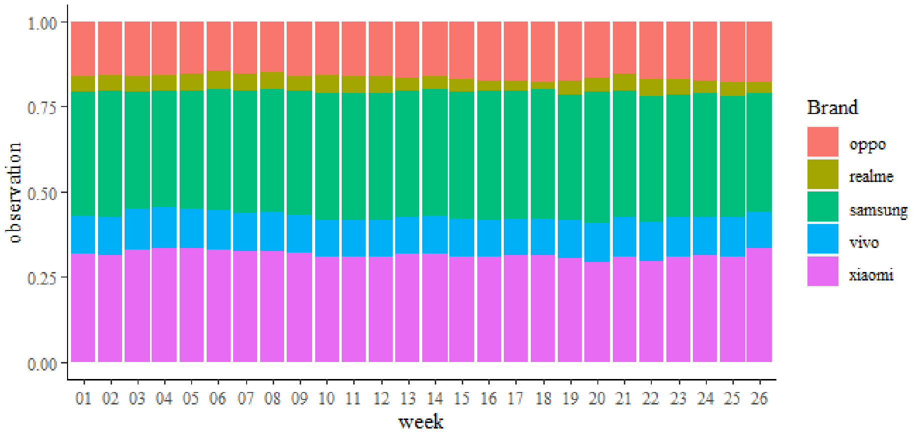


Figure 18 Number of weekly observations by brand

Weighting scenario two is the weighting based on the brand. The weighting value is taken from Indonesia's mobile phone market share in 1st quarter 2020 by IDC (Figure 4). Table 10 illustrates the weighting values for each brand.

Table 10 Weighting scenario 2

Group	Brand	Weight
1	Oppo	22,2/96,2
2	Realme	12,8/96,2
3	Samsung	19,5/96,2
4	Vivo	27,4/96,2
5	Xiaomi	14,3/96,2

Figure 19 shows the comparison of the double imputation index with and without the weighting scenario 2. The results show that in general, the index that has been corrected by the brand have values that are not as high as its original index, especially in the first quarter. It indicates that the observations from the dataset that are dominated by certain brands have impacts on the double imputation index produced.

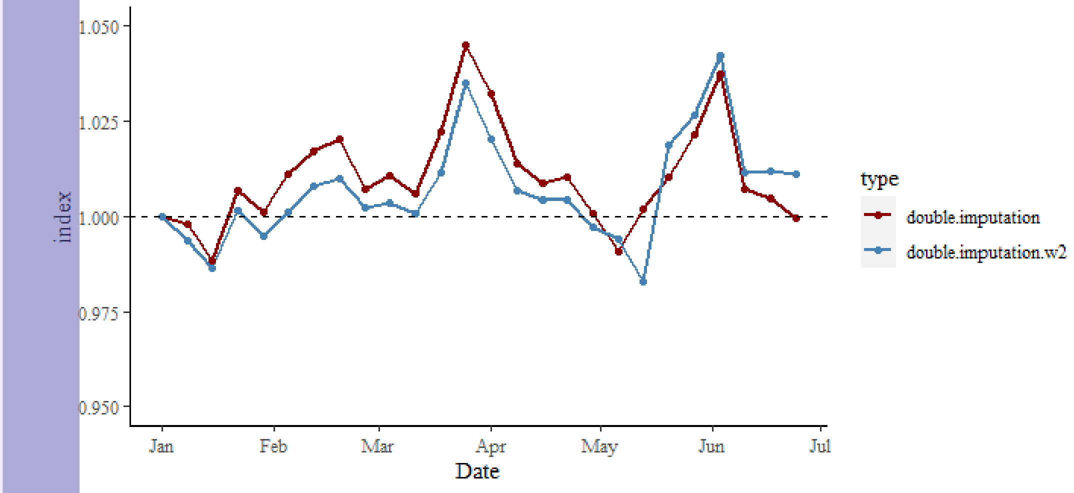


Figure 19 The comparison of hedonic price indices of the mobile phone between the use of weighting scenario 2 and without weighting in double imputation method

Scenario 3: Weighting by YearReleased and Brand

Weighting scenario 3 is the combination of weighting based on the year of release in scenario 1 and the brand of the mobile phone in scenario 2. The results in Figure 19 shows the comparison between all of the indices: the unweighted double imputation index, double imputation index with weighting scenario 1 (w1), double imputation index with weighting scenario 2 (w2), and double imputation index with weighting scenario 3 (w3). Even though calculated with different weighting scenario, it turns out that the movement direction of the four indices is still relatively consistent.

There are no reference parameters to assess the reliability of this index produced in this study. Therefore, the reference for choosing the best model is to look at the stability of price movements. Because it is assumed that the use of big data as the main source will cause noise that can disrupt price movements to become more volatile. In this study, the index that is considered the best is the one using weighting scenario 3 (w3). The w3 index has the smallest standard deviation between the indices displayed in Figure 20 with 0.0125 compared to 0.0133 for the unweighted index, 0.141 for w1, and 0.136 for w2. It means, compared to other weighting scenarios, weighting scenario 3 produces a relatively more stable index. The results of the overall weekly index value are presented in appendix 4.

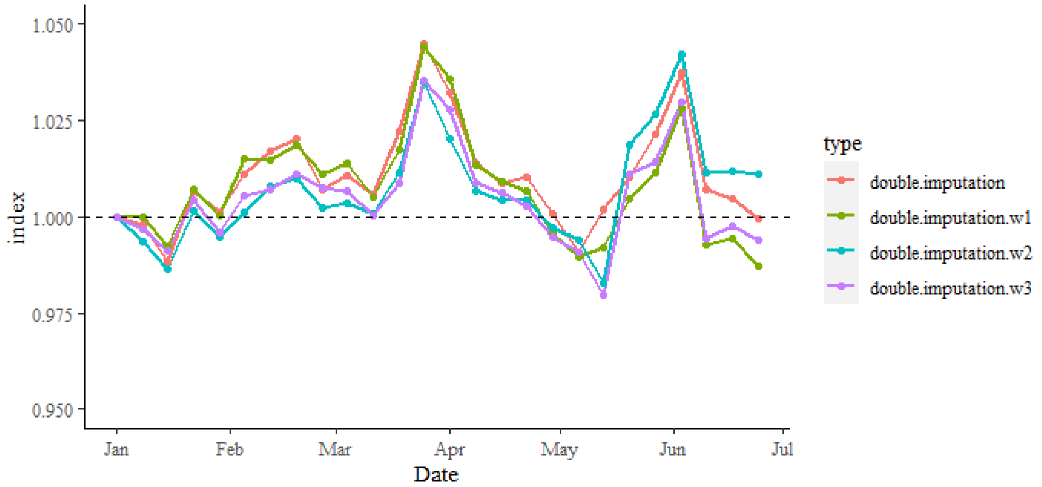


Figure 20 The comparison of hedonic price indices of the mobile phone between the use of all weighting scenario and without weighting in double imputation method

Even though the double imputation index with weighting scenario 3 has the smallest standard deviation, test for equality of variance shows that the weighting scenario 3 has not been proven to provide an unequal variance from the unweighted index. It is indicated by the results of the Lavene test provided in Appendix 5 that produces a p-value of 0.86. Thus, in further research, other weighting scenarios which are likely to have better performance can be investigated.

The index generated from this study measures different scope from the index calculated in the CPI. In this study, price movements are calculated based on the online price that represents online transactions, whereas the CPI measures price movement for offline transactions. The index generated from this study is ultimately intended to be a complementary component to the index for mobile phone constituting the CPI. Nevertheless, some review on the methodology is still needed to achieve that long term goal. Therefore comparing the values of the two indices would be irrelevant. However, it is necessary to evaluate the characters of the resulting index in comparison to the CPI. To do that, a monthly index was generated by taking the geometric mean of the weekly index. The monthly index compared to the index for mobile phone in the CPI is presented in appendix 6.

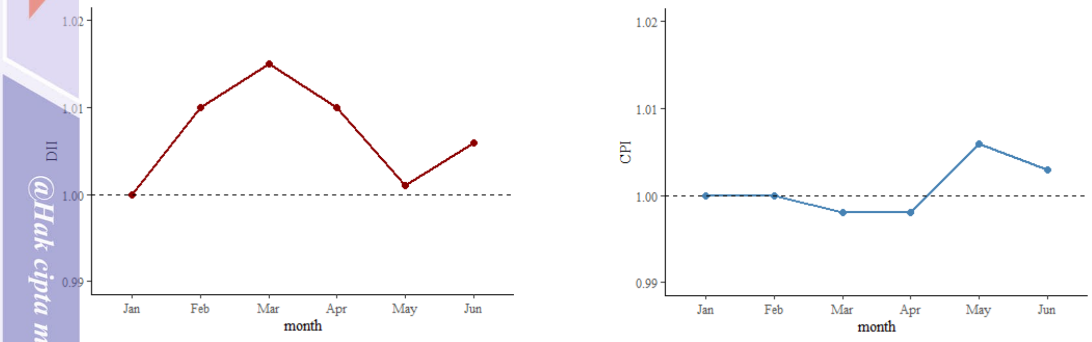


Figure 21 Monthly double imputation index and mobile phone price index from CPI with January 2020 as base period

Figure 21 shows two graphs, the left is the monthly index from the double imputation method with weighting scenario 3, and the right is the index for mobile phone in the CPI. Figure 21 delineates that the index produced in this study is more volatile than the index in the CPI. The direction of the price movement also tends to differ, even the opposite of the index on the CPI. It means that the online retail prices movement for mobile phone in this study behaves differently from the price movement of the online transaction measured in CPI.

The result of this study still needs much improvements to be utilized in the CPI. There are several aspects to be noticed regarding the index resulted. First, the index for online transactions is only available on a national scale. Meanwhile, Indonesian CPI allows for generating a local scale index. Second, to add the effect of online transactions to the CPI, it is necessary to calculate the contribution of online trade to total trade for a commodity. Third, it is crucial to ensure whether the method is in line with the method used in the CPI for the two indexes can be combined.

IV CONCLUSION AND RECOMMENDATION

This research is an empirical study to formulate a hedonic regression model for mobile phone's online retail price and used it to calculate online price indices by implementing the double imputation method. Data collection was done using web scraping techniques on a price comparison website: IPrice.co.id. Data collection was carried out weekly in 6 months from January to June 2020 to capture the price movements. Choosing a price comparison website as the data source has the advantages to gather prices information from various e-commerce sites at once. However, changes in their algorithm might affect the information displayed on their website. It would cause volatility of the scraped dataset.

The hedonic model built in this study produces 11 specifications that have significant effects on the price of a mobile phone. These specifications are battery capacity, core, screen density, internal storage capacity, RAM size, screen size, weight, brand, body material, NFC feature, and RadioFM feature. Due to the presence of outliers and high leverage points, the parameter was estimated using the MM estimation. The rise of the number of cores, density, internal storage, RAM, size and weight are significantly increasing the price of a mobile phone. Oppo, Samsung and Vivo are indicated to be the premium brands for smartphones in Indonesia. The presence of an NFC feature significantly increases the price of a mobile phone while the opposite applied for FM radio feature.

Due to the small proportion of the unmatched observation, the contribution of the hedonic regression coefficient to the double imputation index is relatively small, so that its value is determined more by the matched model index. Three weighting scenarios were carried out to overcome the absence of quantity information: weighting based on the release year, the brand, and a combination of both. It was found that the combination of brand and release year produced the most stable index. The index movement generated behaves differently from the price movements the CPI. It implies that the price movements from online transactions are not the same as offline transactions. The indices produced in this study has the potential to be taken into account in the CPI calculation even though further works are still needed.

Improvement for this research would be to develop a methodology to gather a more stable sample which contains less noise to get a more reliable prediction. Other weighting scenarios could be formulated to overcome the lack of quantity information in the index calculation. There is also a need for a profound methodological study to ensure that the index generated from online transactions can be combined with the index that is produced with the certain methodology used in calculating the CPI.

BIBLIOGRAPHY

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.

Ahmad W, Ahmed T, Ahmad B. 2019. Pricing of smartphone attributes at the retail level in a developing country: hedonic analysis. *Telecommunications Policy*. 43(4):299-309.

Bentley A, Krsinich F. 2017. *Toward Big Data CPI for New Zealand*. Wellington: Stats NZ.

Brody H. 2014. Preventing web scraping: best practices for keeping your content safe. [Retrieved 2020 September 25]. https://www.blog.hartleybrody.com /prevent-scrapers/.html

Carroll R, Ruppert D. 1988. *Monographs on Statistics & Applied Probability , Transformation and Weighting in Regression*. New York: Chapman & Hall.

Cavallo A, Rigobon R. 2016. The billion prices project: using online prices for measurement and research. *J Econ Perspect*. 30 (2): 151-78. doi: 10.1257/jep.30.2.151.

Das K, Dey AK, Hossain AM. 2015. Regression analysis for data containing outliers and high leverage points. *Ala J Math*. 39. 1-6.

De Mauro A, Greco M, Grimaldi M. 2016. A formal definition of Big Data based on its essential features. *Library Review*. 65:122-135. doi:10.1108/LR-06-2015-0061.

[DTTL] Deloitte Touche Tohmatsu Limited. 2019. Have Indonesians’ shopping patterns shifted towards digital?. [Retrieved 2020 Jun 7]. https://www2. Deloitte. com/id/en/pages/about-deloitte/articles/deloitte-indonesia-perspectives.html.

Dirgantara G. 2019. BPS ungkap e-commerce berkontribusi dua persen konsumsi rumah tangga. [Retrieved 2019 November 19]. https://www.antaranews.com/berita/1071816/bps-ungkap-e-commerce-berkontribusi-dua-persen-konsumsi-rumah-tangga.html.

Durianto, Darmadi. 2011. *Strategi Menaklukan Pasar Melalui Riset Ekuitas dan Prilaku Merek*. Jakarta : Gramedia Pustaka Utama.

Feng C, Wang H, Lu N, Tu XM. 2012. Log-transformation: applications and interpretation in biomedical research. *Stat Med*. 32:230–239. doi:10.1002/sim.5486.

Google, Temasek. 2019. e-Conomy SEA 2019. [Retrieved 2019 November 19]. https://www.thinkwithgoogle.com/_qs/documents/8600/e-Conomy_SEA_2019 _Report.pdf.

Gu G, Xu B. 2017. Housing Market Hedonic Price Study Based on Boosting Regression Tree. *J Adv Comput Intell Intell Inform*. 21 (6) 1040-1047.

Gujarati DN, Sangeetha. 2007. *Basic econometrics*. 4th ed.. New Delhi: Tata McGraw Hill Publishing Company Limited.

[IDC] International Data Corporation. 2020. IDC Indonesia: smartphone market posts a new record low in shipments impacted by COVID-19. [Retrieved 2020 September 6]. http://https://www.idc.com/getdoc.jsp?containerId=prAP46346820.

[ILO] International Labour Office; [IMF] International Monetary Fund; [OECD] Organisation for Economic Co-operation and Development; [EUROSTAT] Statistical Office of the European Communities; [UN] United Nations; World Bank. 2004. *Consumer Price Index Manual: Theory and Practice*. Geneva: ILO.

James G, Witten D, Hastie T, Tibshirani R. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Khoirunnisa. 2019. Top 5 Vendor smartphone di Indonesia Q3-2019. [Retrieved 2019 November 23]. https://selular.id/2019/11/top-5-vendor-smartphone-di-indonesia-q3-2019/.

Lancaster KJ. 1966. A new approach to consumer theory. *J Polit Econ*. 74(2):132–157.

Listianingrum T, Nefriana R. 2020. Utilizing Price Comparison Website to Produce Hedonic Price Index. *Asia-Pacific Economic Statistics Week (APES)*. United Nations ESCAP.

Loon KV, Roels D. 2018. Integrating big data in the Belgian CPI. Meeting of the Group of Experts on Consumer Price Indices. Geneva, Switzerland.

Marcel. 2020. What is a smartphone processor and what does it do?. [Retrieved 2020 October 10]. https:// www.coolblue.nl/en/advice/smartphoneprocessors.html.

Maronna RA, Martin RD, Yohai VJ, Salibian-Barrera M. 2006. *Robust Statistics Theory and Methods (with R)*. 2nd Ed. New York: John Wiley & Sons.

Martinez J, Garmendia. 2010. Application of hedonic price modelling to consumer packaged goods using store scanner data. *J Bus Res*. 63 (2010), 690-696.

Montenegro JA, Torres JL. 2016. Consumer preferences and implicit prices of smartphone characteristics. *Malaga economic theory research center working paper 2016-4*. Spain: University of Malaga.

Montgomery DC, Peck EA, Vining GG. 2012. *Introduction to Linear Regression Analysis*. New York: John Wiley & Sons.

Oheix J. 2018. An introduction to web scraping with python. [Retrieved 2020 September 25]. https://www. towardsdatascience.com/an-introduction-to-web-scraping-with-python-a2601e8619e5.html.

Paramansyah A, Ghulam D, Ernawati E. 2020. Pengaruh kesadaran merek (brand awareness) terhadap keputusan pembelian handphone samsung. Al-Kharaj: Jurnal Ekonomi, Keuangan & Bisnis Syariah. 2(1) 88-107. doi:10.47467/alkharaj.v2i1.77.

R&D Center for Post & ICT Resources, Equipment & Operation Research and Development of Human Resources Ministry of Communications and Information Technology. 2016. *2016 Households and Individuals ICT*

Indicators Infographic. Jakarta: Ministry of Communications and Information Technology.

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.

Susanti Y, Pratiwi H, Sulistijowati HS, Liana T. 2014. M estimation, S estimation, and MM estimation in robust regression. *Int J Pure Appl Math*. 91, 349–360. doi: 10.12732/ijpam.v91i3.7.

Trewin D. 2005. *The Introduction of Hedonic Price Indexes for Personal Computers*. Canberra: ABS.

Yohai, V. J. 1987. High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*. 15, 642- 656.

APPENDICES

Appendix 1 Structure of the scraped data

Column Number	Variable
1	Product
2	P_shopee1
3	P_shopee2
4	P_shopee3
5	P_shopee4
6	P_tokopedia1
7	P_tokopedia2
8	P_tokopedia3
9	P_tokopedia4
10	P_blibli1
11	P_blibli2
12	P_blibli3
13	P_blibli4
14	P_lazada1
15	P_lazada2
16	P_lazada3
17	P_lazada4
18	P_lazmall1
19	P_lazmall2
20	P_lazmall3
21	P_lazmall4
22	P_tokopda1
23	P_tokopda2
24	P_tokopda3
25	P_tokopda4
26	P_arjuna1
27	P_arjuna2
28	P_arjuna3
29	P_arjuna4
30	P_personal1
31	P_personal2
32	P_personal3
33	P_personal4
34	P_bukalapak1
35	P_bukalapak2
36	P_bukalapak3
37	P_bukalapak4
38	P_amazon1
39	P_amazon2
40	P_amazon3
41	P_amazon4
42	P_blanja1
43	P_blanja2
44	P_blanja3
45	P_blanja4
46	RAM
47	InternalStorage
48	Size
49	Resolution
50	Density
51	BackCamera
52	FrontCamera
53	Battery
54	Core
55	Weight
56	BackMaterial
57	YearReleased
58	MonthReleased
59	USBConnector
60	WifiStandard
61	SIMCard
62	ChipsetVendor
63	ScreenType
64	NetworkGeneration
65	DualSIM
66	NFC
67	DualCamera
68	Waterproof
69	ScratchResistant
70	WaterResistant
71	VirtualRealityCapable
72	HeartRateCensor
73	LEDNotificationLight
74	FaceRecognition
75	IrisScanner
76	FingerprintScanner
77	RadioFM
78	3DTouch

Appendix 2 Structure of the specification database

No	Variable
1	Type*
2	Brand
3	Size
4	Resolution
5	Density
6	MainCamera
7	SelfieCamera
8	NumberOfCameras
9	Battery
10	Core
11	Weight
12	BackMaterial
13	YearReleased
14	MonthReleased
15	USBConnector
16	WiFiStandard
17	SIMCard
18	ChipsetVendor
19	ScreenType
20	NetworkGeneration
21	Protection
22	Android Version
23	DualSIM
24	NFC
25	DualCamera
26	Waterproof
27	ScratchResistant
28	WaterResistant
29	VirtualRealityCapable
30	HeartRateCensor
31	LEDNotificationLight
32	FaceRecognition
33	IrisScanner
34	FingerprintScanner
35	RadioFM
36	3DTouch
37	Proximity
38	Compass
39	Gyro

*primary key

Appendix 3 Structure of the final dataset

No	Variable	Description	Type
1	Date	Retrieval date	Categorical
2	ID	Product identification	Categorical
3	Median	Median of the prices from all e commerces (Rupiah)	Numeric
4	RAM	Size of Random-access memory (GB)	Numeric
5	InternalStorage	Size of internal storage (GB)	Numeric
6	Brand	Brand of the phone	Categorical
7	Size	Screen size (inch)	Numeric
8	Resolution	Screen resolution (pixel)	Numeric
9	Density	Fixed number of pixels that a screen can display (ppi)	Numeric
10	MainCamera	Highest resolution of back camera (MP)	Numeric
11	SelfieCamera	Highest resolution of front camera (MP)	Numeric
12	NumberOfCameras	Numbers of camera on a mobile phone	Numeric
13	Battery	Capacity of battery (mAH)	Numeric
14	Core	Number of cores on the processor	Numeric
15	Weight	Weight of the phone (gram)	Numeric
16	BackMaterial	The material for the body	Categorical
17	YearReleased	The year of first release	Numeric
18	MonthReleased	The month of the release	Categorical
19	USBConnector	Type of USB connector used	Categorical
20	WiFiStandard	Generation of the WiFi	Categorical
21	SIMCard	Type of SIMCard inserted	Categorical
22	ChipsetVendor	The vendor producing the chipset inside the phone	Categorical
23	ScreenType	Type of screen used	Categorical
24	NetworkGeneration	Generation of the network	Categorical
25	Protection	Type of glass used for the screen	Categorical
26	Android Version	Version of android installed	Categorical
27	DualSIM	Whether a phone has a Dual SIM Card feature	Categorical
28	NFC	Whether a phone has a Near Field Communication (NFC) feature	Categorical
29	DualCamera	Whether a phone has a Dual Camera feature	Categorical
30	Waterproof	Whether a phone is waterproof	Categorical
31	ScratchResistant	Whether a phone is scratch resistant	Categorical
32	WaterResistant	Whether a phone is water resistant	Categorical
33	VirtualRealityCapable	Whether a phone is capable for virtual reality	Categorical
34	HeartRateCensor	Whether a phone has a heart rate censor	Categorical
35	LEDNotificationLight	Whether a phone has a LED notification light feature	Categorical

No	Variable	Description	Type
36	FaceRecognition	Whether a phone has a face recognition	Categorical
37	IrisScanner	Whether a phone has an iris scanner feature	Categorical
38	FingerprintScanner	Whether a phone has a fingerprint scanner	Categorical
39	RadioFM	Whether a phone has a Radio FM feature	Categorical
40	3DTouch	Whether a phone has a 3D Touch feature	Categorical
41	Proximity	Whether a phone has a proximity censor	Categorical
42	Compass	Whether a phone has a compass censor	Categorical
43	Gyro	Whether a phone has a gyroscope censor	Categorical

Appendix 4 Weekly index for the matched model, double imputation, double imputation with weighting scenario 1 (w1), double imputation with weighting scenario 2 (w2) and double imputation with weighting scenario 3 (w3) with week 1 2020 as base period

Scraping date	Week	Matched model	Double imputation unweighted	Double imputation w1	Double imputation w2	Double imputation w3
01/01/2020	1	1,000	1,000	1,000	1,000	1,000
08/01/2020	2	0,997	0,998	1,000	0,994	0,997
15/01/2020	3	0,988	0,988	0,992	0,987	0,991
22/01/2020	4	1,006	1,007	1,007	1,001	1,004
29/01/2020	5	1,001	1,001	1,000	0,995	0,996
05/02/2020	6	1,011	1,011	1,015	1,001	1,006
12/02/2020	7	1,017	1,017	1,015	1,008	1,007
19/02/2020	8	1,020	1,020	1,019	1,010	1,011
26/02/2020	9	1,006	1,007	1,011	1,002	1,008
04/03/2020	10	1,010	1,011	1,014	1,004	1,007
11/03/2020	11	1,005	1,006	1,005	1,001	1,000
18/03/2020	12	1,022	1,022	1,017	1,012	1,009
25/03/2020	13	1,044	1,045	1,044	1,035	1,035
01/04/2020	14	1,032	1,032	1,036	1,020	1,028
08/04/2020	15	1,012	1,014	1,013	1,007	1,009
15/04/2020	16	1,007	1,009	1,009	1,004	1,006
22/04/2020	17	1,009	1,010	1,007	1,004	1,003
29/04/2020	18	0,999	1,001	0,995	0,997	0,995
06/05/2020	19	0,989	0,991	0,990	0,994	0,991
13/05/2020	20	1,002	1,002	0,992	0,983	0,979
20/05/2020	21	1,010	1,010	1,005	1,019	1,011
27/05/2020	22	1,021	1,021	1,012	1,026	1,014
03/06/2020	23	1,037	1,038	1,028	1,042	1,030
10/06/2020	24	1,007	1,007	0,993	1,012	0,994
17/06/2020	25	1,004	1,005	0,994	1,012	0,998
24/06/2020	26	0,998	1,000	0,987	1,011	0,994

Appendix 5 Test for equality of variance between unweighted double imputation index and double imputation index with weighting scenario 3

Method

σ_1: standard deviation of double.imputation
σ_2: standard deviation of double.imputation.w3
Ratio: σ_1/σ_2
The Bonett and Levene's methods are valid for any continuous distribution.

Descriptive Statistics

Variable	N	StDev	Variance	95% CI for σ
double.imputation	26	0.013	0.000	(0.010, 0.020)
double.imputation.w3	26	0.012	0.000	(0.009, 0.019)

Ratio of Standard Deviations

Estimated Ratio	95% CI for Ratio using Bonett	95% CI for Ratio using Levene
1.06424	(0.605, 1.855)	(0.592, 1.812)

Test

Null hypothesis	$H_0: \sigma_1 / \sigma_2 = 1$
Alternative hypothesis	$H_1: \sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0.05$

Method	Test Statistic	DF1	DF2	P-Value
Bonett	0.06	1		0.802
Levene	0.03	1	50	0.860

Appendix 6 Monthly index for double imputation method with weighting scenario 3 and the index for mobile phone in the CPI

with January 2020 as base period

Month	Double imputation w3	CPI
January	1,000	1,000
February	1,010	1,000
March	1,015	0,998
April	1,010	0,998
May	1,001	1,006
June	1,006	1,003

BIOGRAPHY

The author was born in Boyolali, August 20th 1990 as the youngest child of Riyanto and Sawiji, In 2008, the author graduated from SMAN 4 Surakarta and continued her study to Sekolah Tinggi Ilmu Statistik (STIS) which is now renamed to Politeknik Statistika STIS majoring in Computational Statistics, The author gained his Bachelor degree in 2012, The author begin her career as a statistician at the Directorate of Census and Survey Methodology Development from January 2014, She continued his study to IPB University majoring in Applied Statistics 6 years later, in 2018.