Application of The Sequential Hot-deck Imputation Method for Identification of Indonesian Standard Classification of Business Fields (KBLI)

I J Fadillah[1], **C D Puspita**[1]

[1]BPS Statistics Indonesia, Jakarta – Indonesia

*Corresponding author's e-mail: jihadiman22@gmail.com

Abstract. The Covid-19 pandemic requires the adjustment of new habits in daily life, including in a series of data collection processes. One of the new adjustments is to use alternative types of data collection other than face-to-face, such as the telephone and the web. Information collected through telephone interviews is less accurate than the same information collected through face-to-face interviews, such as the level of non-response, consistency between entries, and outliers in the data or often identified as missing values. Missing value will be very influential on data quality when it appears on important variables. One of these variables is the Standard Classification of Business Fields (KBLI). Imputation is one method that can be used to deal with this problem. One method that is quite popular is Sequential Hot-deck Imputation. Therefore, this study aims to facilitate the identification of 5-digit KBLI by utilizing the Sequential Hot-deck Imputation method. The results of this study indicate that the use of the Sequential Hot-deck Imputation method in the KBLI identification process gives very high accuracy results. In addition, the use of this method is very efficient in the identification process, because the time required is very short, even in large datasets.

1. Introduction

Since 2020, most countries in the world, including Indonesia, have experienced disasters caused by the corona virus (Covid-19). The Covid-19 pandemic has affected various aspects of life, such as health, education, social society, and the national and global economy. Thus, various efforts and policies began to be carried out to overcome the impact of Covid-19. All efforts to adjust new habits are also carried out in daily life, such as wearing masks, washing hands, and maintaining distance. Meanwhile, various innovations and adjustments to the new normal life were also implemented by business entities, such as market digitalization and others. National statistical agencies must also make adaptations to the activities of the census and survey processes, such as finding alternative data collection modes, reducing sample sizes, modifying sample designs, reducing question items in questionnaires, or others. There is 69 percent of national statistical agencies that stopped field data collection involving face-to-face interviews [1].

Based on monitoring the state of statistical operations under the Covid-19 Pandemic [1], there are three alternative types of data collection commonly used by national statistical agencies, including phone survey, administrative data, and web surveys. Phone survey and web surveys as an alternative method is a challenge for statistical institutions, such as the National Statistical Offices in Indonesia, BPS-Statistics Indonesia. According to Biemer [2], phone survey may be affordable and timely, however, data quality for some items may be inadequate. Biemer [2] explains that information collected through

phone survey is, in some cases, less accurate than the same information collected through face-to-face interviews, such as the level of nonresponse, consistency between entries, and outliers in the data. Batista and Monard [3] explained that respondents who did not want to be interviewed or could not be found, data were not recorded due to human errors, and equipment and application failures were the causes of incompleteness or missing values in the data. In addition, according to Pearson (2005), missing values can also appear in the form of outliers or values that are inconsistent with previous values, or unreasonable entries in the data [4].

The problem of incompleteness in the observation unit can be handled by removing cases containing missing values and then performing a weighting procedure to modify the weights to adjust for nonresponse as if it were part of the data collection design. According to Little and Rubin [5], such procedures are known as weighting procedures. However, for the case of missing values that occur in question items, the handling of weighting procedures becomes less efficient because missing values only occur in some question items. Deleting all items in the observation unit, of course, will result in the loss of information that has been collected, reduce the amount of data, and make parameter estimation inefficient. According to Little and Rubin [5], when missing values occur in question items, the imputation method is a procedure that can be used to deal with this problem.

According to Biemer [2], missing values are found in almost all large-scale data collection efforts and can be a problem in conducting surveys. Missing value will be very influential on the quality of the data when it appears on important variables, such as the variables that are the basis for the sampling design or estimation later. One of the important variables in many surveys conducted by the BPS-Statistics Indonesia is the Indonesian Standard Classification of Business Fields (*Klasifikasi Baku Lapangan Usaha Indonesia*/KBLI). KBLI is one of the standard classifications published by the BPS for Indonesian economic activities. According to BPS [6], KBLI is used as a classification of economic activity according to business field groups based on an activity approach, which emphasizes the process of economic activity to produce goods/services, as well as a functional approach that looks at the function of economic actors in using inputs such as labor, capital, and goods and services to create the output of goods/services. The classification structure of economic activities at KBLI is consistent and interconnected and based on internationally agreed concepts, definitions, principles, and classification procedures. The basis for the preparation of the KBLI is the International Standard Industrial Classification of All Economic Activities (ISIC), up to 4 digits, adjusted to the ASEAN Common Industrial Classification (ACIC), and East Asia Manufacturing Statistics (EAMS), and has been developed in detail up to 5 digits according to the economic activities involved has existed in Indonesia since 1983 when Indonesian Classification of Business Fields (*Klasifikasi Lapangan Usaha Indonesia*/KLUI) published.

KBLI has an important role in grouping businesses/companies in Indonesia according to certain established rules or standards so they can be used for statistical administration, basic planning, policy evaluation, and licensing. Several publications and outputs produced by BPS-Statistics Indonesia are presented in the form of KBLI. For example, Micro and Small industry (*Industri Mikro dan Kecil*/IMK) Index data and IMK Profile Publications are presented in the form of a 2-digit KBLI taken from the first 2 digits of the 5-digit KBLI [7]. The incorrect KBLI identification will affect the data output. Doing the data pre-processing process manually is certainly very inconvenient, especially for data with many observations. In addition to taking time, of course, it can also be prone to errors. Therefore, this study aims to facilitate the identification of 5-digit KBLI by utilizing the algorithm of imputation method.

Many imputation methods exist. One method that is often used is the Hot-deck Imputation method. This method is a refinement of the previous method, namely the mean imputation, especially in the estimation of variance that is underestimated [8]. This method is more suitable for use on many types of data because it can be used on various types of data, both numeric, category, and mixed data. Several other studies discuss the use of this method. Pazanudin [9], discusses the comparison of Hot-deck and missForest methods in data imputation. Then, Fadillah and Muchlisoh [10], discussed the comparative analysis of the Hot-deck Imputation method and this method in overcoming missing values. Several previous studies tried to compare the use of the Hot-deck Imputation method on numerical data. However, the risk of missing values occurring in KBLI data (categorized data), will greatly affect the output data. Therefore, this study aimed to apply the Sequential Hot-deck Imputation method for identify

the KBLI and to analyze the accuracy and time performance generated by the Hot-deck Imputation method in imputing KBLI data.

2. Methodology

2.1. Missing Value Handling Method

Several methods have been developed to handle missing values. Missing value handling methods can be carried out starting from simple methods such as removing cases containing missing values to more advanced methods based on inference and further modeling. According to Little and Rubin [5], one method that can be used to deal with the problem of missing values is the Imputation-Based Procedures. Imputation is a term that denotes the procedure of replacing a missing value in the data with some reasonable value. A commonly used procedure for imputation is Hot-deck Imputation, in which the units recorded in the sample are used to replace the values.

Hot-deck imputation involves replacing missing values with other values based on the concept of similarity. Hot-deck imputation is one of the most popular imputation methods used. Despite their popularity in practice, there is still a lack of literature on the theoretical properties of the various methods. There are quite a few types of Hot-deck Imputation algorithms, but Andridge and Little [11], explain that there are two types of Hot-deck Imputation methods that are often used, namely Random Hot-deck Imputation and Sequential Hot-deck Imputation. This research will focus on the Sequential Hot-deck Imputation algorithm.

In determining the donor, the Sequential Hot-deck Imputation method will sort the data using a predictor variable, then the missing value will get the donor from the data of the previous or subsequent periods. The predictor variable used is the variable that is assessed to be related to the variable to be imputed [8]. Grau [12] explains that the data sorting process is selected based on the presence or absence of a relationship between variables. For example, the data containing missing values (y) that has been sorted by variable predictor (x) has a number of observations of 6. Then 3 of them are missing values, where y_1, y_4, and y_5 are valuable, while y_2, y_3, and y_6 are missing. So, y_2, y_3 are imputed by y_1, and y_6 are imputed by y_5, whereas if the missing value case only occurs in y_1 then some initial values may be needed, they can come from previous survey records, or use the later available data values, namely y_2 [13].

According to Batista and Monard [3], Hot-deck Imputation is implemented in two stages The first stage is to partition the data into clusters and the second stage is to donate values that are considered to have similar elements. The formation of clusters is done by sorting the data using predictor variables. Then, after the data is sorted, the missing value will be imputed using the donor value. Donor values are obtained based on variables that have similarities in the cluster or using data from the previous or subsequent periods. However, this method has a weakness when the number of missing values is large enough so that the value will be filled repeatedly which causes the estimation to be biased [8].

2.2. Analytical Method

The analysis used in this study was carried out in the form of a simulation. The simulation will be carried out in three stages using the help of R software. The first stage is the formation of an empty dataset which will later be identified. The formation of this dataset was carried out randomly at the levels of 10 percent, 20 percent, 30 percent, 40 percent, 50 percent, and 60 percent as research by Batista and Monard [3]. The dataset that has been formed will then proceed to the second stage for the KBLI identification process using the Sequential Hot-deck Imputation method. The third stage is data accuracy analysis. the accuracy of the analyzed data sees how accurate the identification (prediction) results are with the actual (actual) results. Hamner, et al [14] explains that accuracy is defined as the proportion of the actual element which is equal to the corresponding element in the prediction. Mathematically formulated as follows:

$$Error\ Rate = \frac{Total\ Incorrcet\ Prediction}{Total\ Prediction} \times 100\% \quad (1)$$

$$Accuration = 100\ \% - Error\ Rate \quad (2)$$

The following is a flowchart that explain overview of these stages.

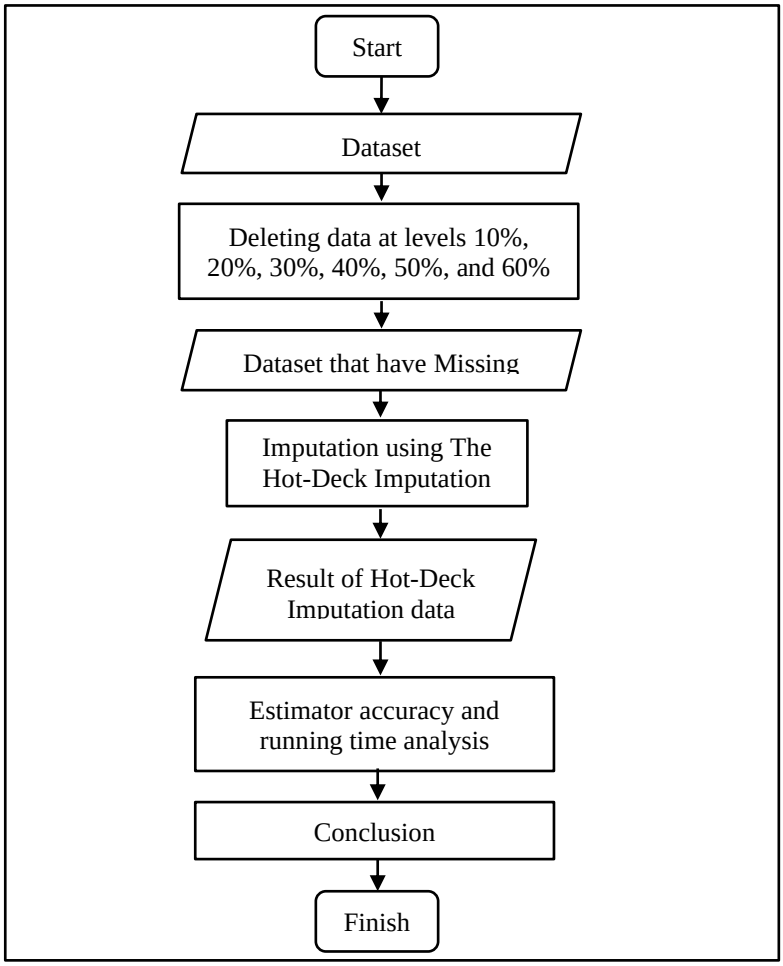

Figure 1. Flowchart Simulation of Imputation Process.

The simulation process (stage 1 until stage 3) will be repeated 20 times. Then the results of each simulation will be compared based on the accuracy of the KBLI identification and also the running time. The KBLI identification will be seen on the 2-digit KBLI and 5-digit KBLI.

2.3. Data and Sources Data

The data used in this study is secondary data that sourced from the BPS Statistics Indonesia. The data used is the 2020 Annual Micro and Small Industry Survey data from the BPS-Statistics Indonesia. As for the observations in this study is the 5-digit KBLI Manufacture Industry. The variables used in this study were 5-digit KBLI, the type of goods produced, and standard units. The 5-digit KBLI will be used as a variable to be identified (imputed), while the type of goods produced and standard units will be used as predictor variables.

3. Result and Discussion

3.1. General Overview of Micro and Small Industry Survey Data

Data analysis in this study used 5-digit KBLI data for the Manufacture Industry. The amount of data used is 76,689 records. Figure 2 describes the distribution of data according to the 2-digit KBLI. A total of 27.09 percent of the data is KBLI 10, namely the Food Industry, followed by KBLI 16 which is the Wood, Goods from Wood and Cork Industry (excluding furniture), Woven Goods from Rattan, Bamboo, and the others at 14.17 percent and KBLI 23 which is the Non-Metal Mineral Industry by 10.27 percent. Furthermore, simulations will be carried out in the KBLI identification process for 20 times repeatedly for each level of the percentage of missing data created.

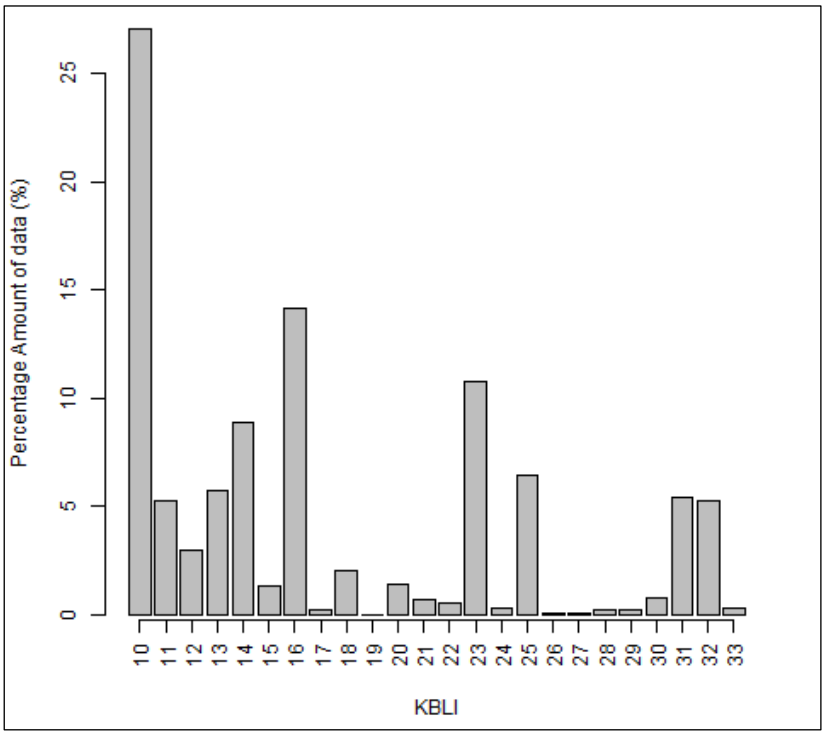

Figure 2. Percentage amount of data by 2-digit KBLI (percent).

3.2. Analysis of Imputation Results

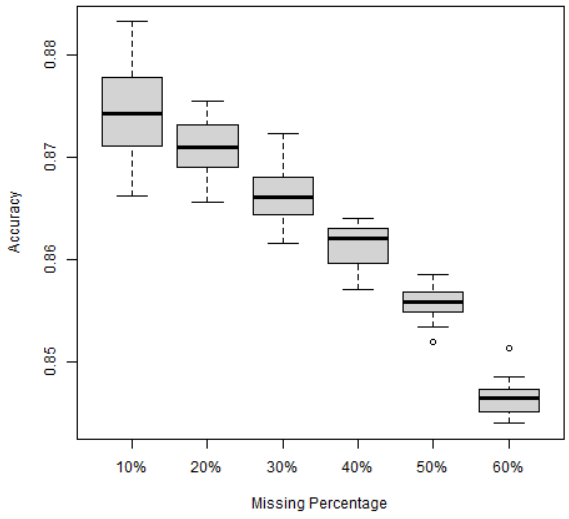

Figure 3. Accuracy of the 2-digit KBLI identification by missing data percentage.

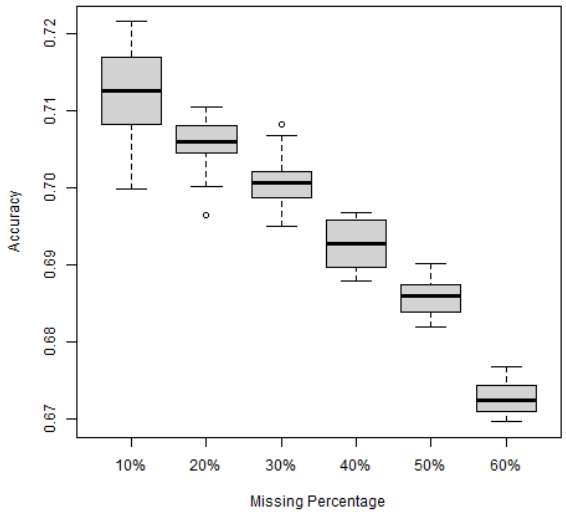

Figure 4. Accuracy of the 2-digit KBLI identification by missing data percentage.

Figure 3 shows the accuracy of the 2-digit KBLI identification result by the Sequential Hot-deck Imputation method. These results indicate that the accuracy by the Sequential Hot-deck Imputation method in identifying 2-digit KBLI is very high, which is above 84 percent. The increased number of identified data makes the accuracy level decrease for each percentage of missing data. However, even so, the decreasing accuracy in the 2-digit KBLI identification is still quite low, at an average of 0.56 percent. At the percentage of missing data on 10 percent, the average accuracy is 87.44 percent. Then the percentage of missing data is 20 percent, the accuracy decreases to 87.10 percent. At the percentage level of missing data on 10-50 percent, the average decrease in accuracy tends to be stable at around 0.47 percent. However, when the number of missing data exceeds the number of available donors (60 percent missing data), the decrease in accuracy reaches 0.93 percent.

Based on the simulation results of the 2-digit KBLI identification, the increase in the number of data identified in the dataset causes the accuracy of the Sequential Hot-deck Imputation method to decrease. This is due to the increasing number of missing values, and the decreasing number of donors. The decrease in accuracy is even more pronounced when the amount of data identified is above 50 percent or more than the data that became donors in the KBLI identification process.

Figure 4 shows the accuracy of the 5-digit KBLI identification result. The figure shows the accuracy produced by the Sequential Hot-deck Imputation method. These results indicate that the accuracy generated by the Sequential Hot-deck Imputation method in identifying 5-digit KBLI is quite high, which is above 67 percent. Similar to the results of the 2-digit KBLI identification, an increase in the number of data identified makes the accuracy level produced decreases for each percentage of missing data. However, in contrast to the 2-digit KBLI identification result, the 5-digit KBLI identification has a greater decrease in accuracy for each percentage of missing data. It can be seen that the percentage of missing data is 10 percent, the average accuracy produced is 71.23 percent. Then the percentage of missing data is 20 percent and the accuracy decreases to 70.58 percent. At the percentage level of missing data of 10-50 percent, the average decrease in accuracy tends to be stable at around 0.66 percent. However, when the number of missing data exceeds the number of available donors (60 percent missing data), the decrease in accuracy reaches 1.31 percent.

Based on the simulation results of the 5-digit KBLI identification, it can be seen that the increase in the number of data identified in the dataset causes the accuracy of the Sequential Hot-deck Imputation method to decrease. This decrease is quite large when compared to the 2-digit KBLI identification. The decrease in accuracy is even more pronounced, when the amount of data identified is above 50 percent or more than the donors data in the KBLI identification process.

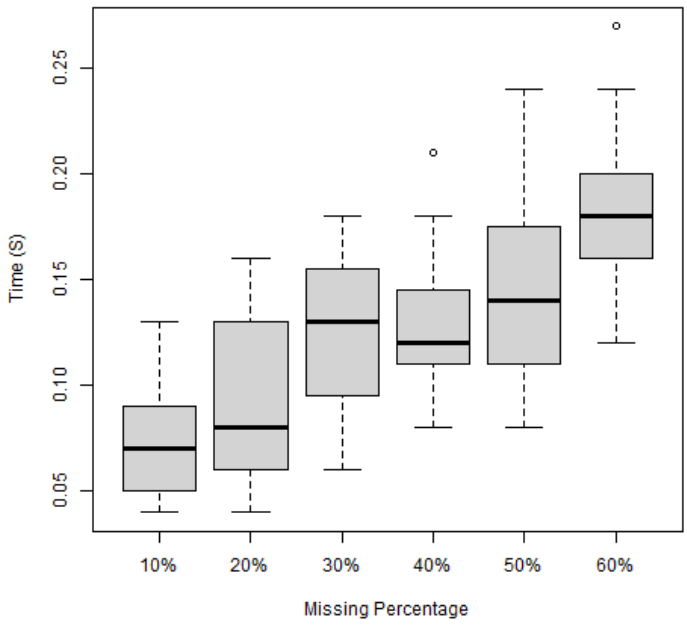

Figure 5. Imputation time by missing data percentage.

Figure 5 shows the imputation/running time required to identify the KBLI in the simulation process. Overall, it can be seen that the imputation time required in the identification process using the Sequential Hot-deck Imputation method is very fast. The time it takes to be under 1 second. Contrary to the results in the previous discussion of accuracy, the imputation time required by the Sequential Hot-deck Imputation method increases as the number of missing data increases. However, the increase that occurs is not significant and is still in a short period of time. This shows that the Sequential Hot-deck Imputation method is good for identifying large numbers of datasets and can also be done on data with a large number of missing data, of course, taking into account the number of donors, which should not be less than 50 percent of the amount of data to be identified.

Table 1. Result of the 5-digit KBLI accuracy, the 2-digit KBLI accuracy, and imputation time by missing data percentage.

Result	Persentase *missing data*					
	10%	20%	30%	40%	50%	60%
Accuracy of the 5-digit KBLI	0,712	0,706	0,701	0,693	0,686	0,673
Accuracy of the 2-digit KBLI	0,874	0,871	0,866	0,861	0,856	0,846
Imputation time	0,075	0,093	0,125	0,126	0,149	0,185

Based on the results, further analysis can be carried out regarding of using the Sequential Hot-deck Imputation method in the KBLI identification process. Table 1 shows that the Sequential Hot-deck Imputation method produces good accuracy in the 2-digit KBLI identification process The accuracy of the 5-digit KBLI identification is quite good, however, when the missing data is above 30 percent, it is below 70 percent. According to the results, the Sequential Hot-deck Imputation method is very suitable to be used for the identification of 2-digit KBLI, even though the missing data rate is quite high, of course, taking into account the number of available donors. Meanwhile, for the 5-digit KBLI identification process, the Sequential Hot-deck Imputation method is suitable for datasets with a relatively smaller number of missing data, about 30 percent below the number of donors. Then, apart from its accuracy, the Sequential Hot-deck Imputation method is very suitable to be used because the running time required to the identification process is quite short, even on datasets with large amounts of data.

4. Conclusion

Based on the simulation results and discussions that have been carried out the conclusions obtained in this study indicate that using the Sequential Hot-deck Imputation method for the identification process of KBLI is very precise, especially for the 2-digit KBLI identification that its accuracy is very high. For the 5-digit KBLI identification, the results are quite good, but it becomes less good when the amount of data identified is quite large. Determining the number of donors data is very important in the identification process using this method. In addition, the use of the Sequential Hot-deck Imputation method is very efficient for the identification process because the imputation time required is very short, even in large datasets.

The use of the Sequential Hot-deck Imputation method can be used as an alternative in the data identification process, especially in the data pre-processing stage, such as in the data imputation process and data cleaning. In addition to its high accuracy, relatively short running time even on large data, it can make the pre-processing stage of data faster and output data will be served faster.

References

[1] The Statistics Division of the United Nations Department of Economic and Social Affairs and the World Bank's Development Data Group 2020 *Monitoring the state of statistical operations under the COVID-19 Pandemic* access on 23 July 2021 via https://covid-19-response.unstatshub.org/survey/covid-19-nso-survey-report-1.pdf

[2] Biemer P P and Lyberg L E 2003 *Introduction to survey quality* (New Jersey: John Wiley & Sons,

Inc.)
[3] Batista, Gustavo E. A. P. A. Maria Carolina Monard 2002 *A Study of K-Nearest Neighbour as an Imputation Method. Second International Conference on Hybrid Intelligence*, 8
[4] Luengo J 2011 *Missing valuess in Data Mining* access on 23 July 2021 via http://sci2s.ugr.es/MVDM/index.php
[5] Little R J A and Rubin D A 2002 *Statistical Analysis with Missing Data 2 ed.* (New York: John Willey and Sons)
[6] BPS Statistics Indonesia 2020 *Indonesian Standard Classification of Business Fields (KBLI) 2020* access on 23 July 2021 via https://bps.go.id/website/fileMenu/KBLI-2020.pdf
[7] BPS Statistics Indonesia 2019 *BPS KBLI and the Boundary Case of the 2019 Quarterly Micro and Small Industry Survey* (Jakarta: BPS)
[8] Hendrawati T 2015 *Study of Imputation Methods in Handling Missing Data* Proceedings of The National Mathematics and Mathematics Education UMS Seminar (Surakarta: Muhammadiyah Surakarta University)
[9] Pazanudin A F 2017 *Study of Missing Data: Comparison of Hot-deck Method and MissForest in Data Imputation* (Jakarta: STIS)
[10] Fadillah I J and Muchlisoh S 2019 *Comparison of Hot-deck Imputation Method and KNNI Method in Handling Missing Values* Proceedings of The Official Statistics National Seminar 2019 STIS access on 23 July 2021 via https://doi.org/10.34123/semnasoffstat.v2019i1.101
[11] Andridge Rebecca R and Little Roderick J A 2010 *A Review of Hot Deck Imputation for Survey Non-response* 78(1), 40-46
[12] Grau E A, Frechtel P A, Odom D M, and Painter D 2004 *A simple evaluation of the imputation procedures used in nsduh* (Toronto: American Statistical Associatio)
[13] Fadillah I J 2019 *Comparison of Hot-deck Imputation Method and K-Nearest Neighbor Imputation Method in Handling Missing Values* (Jakarta: STIS) p 90
[14] Hamner, Ben., Frasco, Michael., and LeDell, Erin 2018 *Metrics Packages, The R Jornal* access on 7 July 2021 via https://cran.r-project.org/web/packages/Metrics/Metrics.pdf