Evaluating Age Heaping Mitigation: A Contrast between Kernel Smoothing and Local Polynomial Smoothing Methods

N A Putri[1,*], E T Astuti[1], L M A Fadila[2], S S Hafizhah[3]

[1] Politeknik Statistika STIS, Jl. Otto Iskandardinata No.64C, Jakarta, Indonesia
[2] Badan Pusat Statistik Kabupaten Lombok Barat, Jl. Soekarno Hatta, Lombok Barat, Indonesia
[3] Pusat Pelaporan dan Analisis Transaksi Keuangan, Jl. Ir. Haji Juanda No.35, Jakarta, Indonesia

* Corresponding author's e-mail: nadiaarsyta23@gmail.com

Abstract. Age data plays an important role in every aspect yet there are still found age misreporting. It involves digit preference that causes build up in a certain age. Digit preference in demography is called age heaping that often happens at age with 0 and 5 as the last digit. Age heaping induces poor data quality and data bias that could influence government policy making. Two indicators used to detect age heaping are Whipple Index (WI) and Myers Blended Index (MBI). Methods to cope with age heaping are nonparametric regression approaches which are Kernel Smoothing and Local Polynomial Smoothing. The objective of this research is to measure and elevate the quality of population age data and population mortality data in Sensus Penduduk (SP) 2020 as well as comparing methods between Kernel Smoothing and Local Polynomial Smoothing. The data being used in this paper is SP2020 which the research variables are age population, age of death, and total population. The result shows that the data quality of total population death is inaccurate compared to total population thus needs a smoothing process to improve age data to population data accuration. The method that has better accuracy is the Local Polynomial Smoothing method.

1. Introduction

Age is an essential basic information. Age data plays an important role in every aspect such as demography, health, up to government policy making-needs. Age data also has a role as the basis for making population pyramids and weighing population projections. In addition, age data can be used as a predictor variable in studies. Therefore, good quality and accurate age data is needed, which comes from a good data collection process.

Although it seems simple, the age data collecting process reported by the community is often inaccurate [1]. In addition to reporting, age data can often be obtained from documents containing birth information or date of birth. However, documents regarding birth information used as evidence of age in Indonesia are still lacking. According to the National Development Planning Agency or [2], in Indonesia only 37.8 percent or as many as one in three household members have and can show a birth certificate. This increases the chance of errors occuring in age reporting [3], [4]. This inaccuracy in reporting age data in demographics is often referred to as age misreporting [5].

Age misreporting is also related to digit preference. This term means the tendency to give numerical responses ending with a certain number, usually ending with 0 or 5 [6]. The existence of digit preference causes the accumulation and widening of the age distribution with numbers ending in 0 and 5 or in demography called age heaping. Age heaping is a problem that is often encountered in age data collection [4]. Age heaping is a situation where there is a buildup of a person choosing a certain age, usually chosen as the number with the last digit of 0 and 5 [7].

Population age data collection conducted by BPS is usually collected in the form of a single age, such as in Sensus Penduduk (SP), Sensus Pertanian (ST), Sensus Ekonomi (SE), Survei Kerja Nasional (SUPAS), Survei Sosial dan Ekonomi Nasional (Susenas), Survei Angkatan Kerja Nasional (Sakernas), and other surveys. The advantages of single age include being easier to analyze, more accurate, more detailed, and more flexible. Some statistical analyses are also more effective when using single-age data because it will provide a more accurate picture and sharpen development targets in several fields such as education, health, family planning, employment, and other fields [8]. Therefore, single age data with good quality is needed. If the single age data is incredulous due to age heaping, it will result in misrepresentation or inaccuracy of demographic estimates [5]. In addition, the problem caused by age heaping is that the data will be biased so that it is inappropriate for further analysis [4] and will interfere with the scientific process for researchers [9].

Age heaping at a certain single age will certainly make a poor-quality age data and affect the making of government policies. Good single-age quality will help in the realization of development planning for the government so that its programs and policies aim the right target.[10].

To overcome the age heaping problem, BPS performed age data smoothing with the Arriaga method [11]. This method is one of the methods used for estimating fertility, correcting age misreporting, and smoothing data [12]. However, this method has a weakness in terms of smoothing age data, which is only for age group data, causing inaccuracies in age allocation and distortions in younger age groups [13], [14]. Another weakness is that it does not take into account population variations that occur within the 5 and 10 year age groups [15]. In addition, the Arriaga method assumes that birth and mortality rates will remain constant over time [12], [16]. But in fact, based on the World Mortality Report 2019 by the World Health Organization or WHO, the mortality rate has variations in different age groups [17].

Population data based on age considered to have better data quality than mortality data based on age. This can be seen from the lack of ownership of death certificates rather than birth certificates. According to a research report conducted by [2], 43.8 percent of the Indonesian population who experienced a death event did not know what a death certificate was; 40.6 percent considered death certificates unimportant; 7.8 percent considered that remembering death was not commonplace; and 5.5 percent did not know the flow of making a death certificate. Only 2.3 percent had applied for a death certificate. Based on this data, the Indonesian population is still lacking in reporting mortality data. In conclusion, the quality data of mortality based on age has a greater chance of age heaping than the number of populations based on age.

This is supported by related research [18] which uses data on the mortality based on age from SP2010. It can be seen from Figure 1 that there is a buildup of population numbers at ages with numbers 0 and 5 as the last digit. This shows that there is clear age heaping in the data.

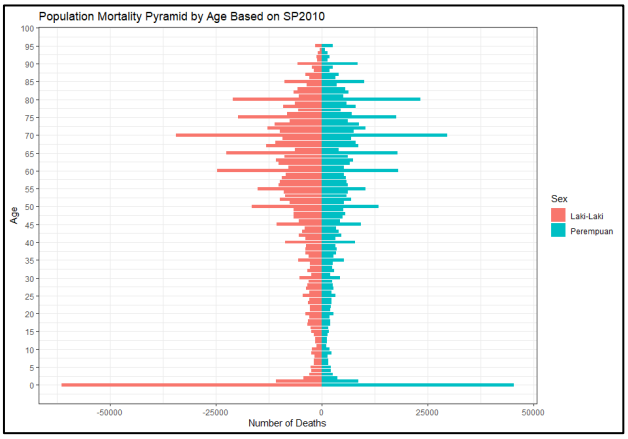

Figure 1. Population mortality pyramid based on age in SP2010
(Source: Firdaus and Astuti (2017)

Apart from the data on mortalities number based on age in SP2010, the results of Susenas in 2021 also show a buildup in single ages with the last digit of 0 and 5 (Figure 2). Therefore, it is estimated that in the data on the total population based on age from the census and other surveys conducted by BPS, there is a possibility of age heaping, just like in the SP2020 data.

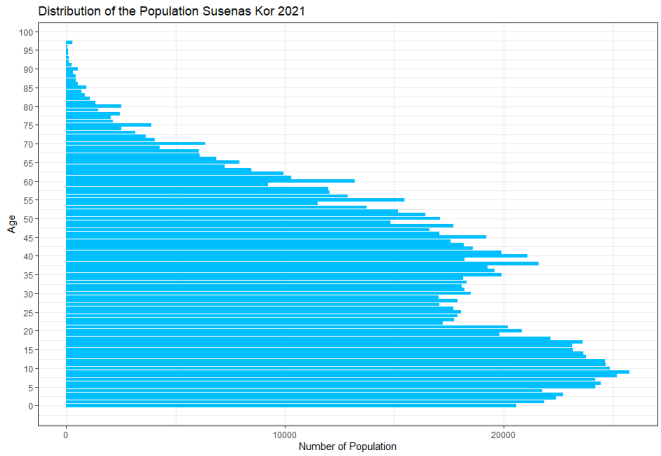

Figure 2. Population single age distribution Susenas Kor 2021

In SP2020, there was a different method of data collection compared to previous censuses or surveys. This method involved using population administration data provided by the Directorate General of Population and Civil Registration (Ditjen Dukcapil). Additionally, it was carried out in two phases, namely the short form and long form. Given the differences in data collection methods between previous surveys and censuses and the SP2020, the question arises whether there is still age heaping affecting the quality of age data in the population for the SP2020.

There are several indicators used to measure the age data quality in the distribution of which there is age heaping, including the Whipple Index (WI) and Myers Blended Index (MBI). WI is used to measure the extent to which respondents tend to report ages with 0 or 5 as the last digit [4]. Meanwhile, MBI measures based on the assumption that the population is evenly distributed in the aggregate population for each age with the last digit 0 to 9 [19].

One method that can be done to overcome age heaping is the smoothing method using nonparametric regression. Nonparametric regression has advantages in terms of flexibility and robustness because it does not require assumptions that must be met as in parametric regression [20]. Nonparametric smoothing methods are used, namely Kernel Smoothing and Local Polynomial Smoothing. The Kernel

Smoothing smoothing method was used by Vallarino in 2017 [21] who conducted forecasting of the Indonesian Composite Stock Price Index in 2017. Meanwhile, the Local Polynomial Smoothing method was used by Firdaus and Astuti [18] on handling age heaping on the data of population mortality based on age from SP2010. Research conducted by Lyons-Amos and Stones in 2017 showed the existence of age heaping in 34 Sub Saharan African countries [4]. The age heaping phenomenon was also found by Fayehun and his colleagues on the Nigerian Demographic Health Survey data in 2003, 2008, and 2013 [5].

Different methods of data smoothing will result in different data smoothing estimations. Both methods are included in nonparametric regression which does not need to fulfill any assumptions. Based on the explanation of the problem above and the advantages of using nonparametric regression in evaluating age data, we are interested in comparing the two methods to overcome age heaping in the latest census results conducted by BPS.

2. Theoretical Background

2.1. Age Heaping

The numbers 0 and 5 are more attractive than other numbers [22]. In a demographic perspective, this creates age heaping. Age heaping is the result of respondents not knowing their age [23]. Age heaping is a demographic phenomenon where respondents report their age with other numbers, but close to their actual age, such as rounding to numbers with 0 or 5 as the last digit [24]. Age data that experience age heaping can be found in population data based on age, mortality data based on age, and so on [18].

2.2. Whipple Index (WI)

This index is used to measure the quality of population age data as seen from the tendency of respondents to report age or digit preference with 0 or 5 as the last digit [25]. This index is used to measure the quality of population age data as seen from the tendency of respondents to report age or digit preference for 0 or 5 as the last digit [26].

WI is calculated using the following formula.

$$WI = \frac{\sum(P_{25}+P_{30}+P_{35}+\cdots+P_{50}+P_{55}+P_{60}+P_{65})}{\frac{1}{5}\sum(P_{23}+P_{24}+P_{25}+\cdots+P_{60}+P_{61}+P_{62})} \times 100 \quad (1)$$

Description:

WI = Whipple Index

$P_{23}, \ldots, P_{65}$ = Number of populations aged 23rd to 65t

The United Nations Statistics Division (UNSD) [27] determines the quality measure of age data reporting based on WI values with five categories as shown in Table 1.

Table 1. Age data accuracy measurement based on WI

WI	Data Quality
<105	Very accurate
105 – 109.9	Accurate
110 – 124.5	Moderately accurate
125 – 174.9	Inaccurate
>175	Very inaccurate

2.3. Myers Blended Index (MBI)

MBI measures the age data quality from the tendency to mention age with all numbers ending in 0 to 9. By including all numbers, it can be seen which numbers are most preferred and most avoided. This index shows digital preference, hence the calculation is done on a single age distribution. The "blended"

technique in this index aims to determine the proportion of the population that ends with a certain number out of the total population 10 times, by varying a certain starting age for each 10-year age group [28]. It minimizes bias in the index due to the fact that ages ending in zero will usually be longer than other terminal digits. The measure of age data quality is based on the MBI value which can be categorized into two, namely good (MBI < 10) and bad (MBI > 10) [29].

2.4. Nonparametric Regression

Regression analysis is one of the techniques in statistics that model the mathematical relationship between the response variable Y and one or more predictor or predictor variables X [30]. In a regression model, the response variable is expressed as a linear function and other variables that are more than one is called predictor variables. In such a model, it is implicitly assumed that there is a causal relationship between the response variable and the predictor variable that only flows in one direction, namely from the predictor variable to the response variable [31].

General form of classical linear regression model:

$$y_i = m(x_i) + \varepsilon_i, \ i = 1,2,\dots,n \tag{2}$$

Description:

y_i = observation value of the i-th response variable y

$m(x_i)$ = dependent regression curve of the predictor variable x

ε_i = i-th error component

Estimation for the $m(x_i)$ regression curve can be done using parametric and nonparametric approaches. Parametric regression is relatively better when viewed from the properties of the ideal estimator but requires information from the past to know the pattern and distribution of the data. Estimating a regression curve with a parametric approach also requires strict assumptions to be met regarding the shape of the regression function, whether linear, quadratic, exponential, or polynomial [32].

Meanwhile, the use of nonparametric approaches does not have a specific function pattern specification, so other estimators are used to estimate the regression function, such as kernel, local linear, spline, and so on [33]. Nonparametric is also able to minimize the assumption of the relationship pattern between the response variable and the predictor and let the data adjust the regression curve model according to its empirical conditions [34]. Nonparametric approach by applying the smoothing method because the goal is to find a relationship pattern that fits the empirical data. Techniques or methods that can be used in nonparametric regression are Local Polynomial Smoothing [35], Kernel Smoothing [36], and Spline Smoothing [37], [38].

2.5. Kernel Smoothing Method

Basically, the kernel method has similarities with other linear estimators, but the kernel method is specifically focused on using a more specific bandwidth method [32]. The kernel method uses a bandwidth, and the selection of the estimator is based on a qualitative assessment of the estimation results. Kernel smoothing method is an approach to the representation of a sequence of weights $W_{n1}(x), W_{n2}(x), \dots, W_{nn}(x)$ that aims to describe the weight function $W_{ni}(x)$ with a density function and scale parameters to measure the size and shape of the weights around x [34]. The weighting function is called the kernel function K.

Nadaraya and Watson (1964) defined a kernel regression estimator, and it is known as the Nadaraya-Watson estimator, as written in equation (3). The estimator is used to estimate the regression function m(xi) in the nonparametric regression model (Equation 2) [39].

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x-x_i)y_i}{\sum_{i=1}^{n} K_h(x-x_i)}, \ i = 1,2,\dots,n \tag{3}$$

Description:

$\hat{m}_h(x)$ = estimated regression curve of y on x values

h = bandwidth
x_i = age at the i-th observation
K = kernel function
y_i = total population at the i-th observation
n = number of observations

2.6. Local Polynomial Smoothing Method

The Local Polynomial Smoothing method is a data smoothing method with a nonparametric regression approach that uses polynomial regression locally within an area [35]. This method is referred to as a local smoothing method because the resulting estimate is done locally at points around a certain interval of values.

In the Local Polynomial method, there is a polynomial degree (p) which aims to reduce the bias between the regression curve and the empirical data. The higher the degree of polynomial used, the estimated regression curve formed will be closer to the empirical data pattern [30]. From previous researchers [35], it is recommended to use a polynomial degree that is not too high, for example p = 1 or p = 2. If p=1 is used.

$$m(x_i) = \beta_0 + \beta_1(x_i - x_0) \tag{4}$$

Equation (4) is often referred to as the local linear model. The step to obtain the regression function estimate $m(x_i)$ is equivalent to the step to obtain the parameter estimate β_j, since $\widehat{m}(x_i) = \widehat{\beta_0} + \widehat{\beta_1}\,(x_i - x_0)$. Where $\widehat{\beta_j}$ or $\widehat{m}(x_i)$ is the solution to equation 5 below.

$$\sum_{i=1}^{n}\left(y_i - m(x_i)\right)^2 K_h(x_i - x_0) \tag{5}$$

Estimation is done by the weighted least square method with kernel function weights.

2.7. Kernel Function

A kernel function is a finite, continuous, real-valued, and symmetric function of K integrated into one. There are several kernel functions that are widely used as a weight, including the Gaussian, Uniform, Epanechnikov, and Triangular kernels [34]. The Gaussian kernel function is an easier kernel function to use than other kernel functions [40]. This is because this kernel function provides an overall weight on the distribution of data in the process with a smoother graph shape and close to normal distribution. Here is the formula of the Gaussian kernel function.

$$K(x) = \left(\sqrt{2\pi}\right)^{-1} \exp\left(-\frac{x^2}{2}\right), -\infty < x < \infty \tag{6}$$

2.8. Bandwidth

In estimating the regression curve with the Kernel Smoothing or Local Polynomial Smoothing method, a smoothing parameter (bandwidth/h) is given. The selection of h is the most important thing in this method because it will determine the results of the regression curve estimator formed [35]. The selection of the optimal bandwidth is more influential than the selection of the kernel function [41].

A bandwidth that is too small ($h \to 0$) will create a complex regression model because it can determine its own data pattern (under smooth), but the variance will be large even though the bias is small. Meanwhile, a bandwidth that is too large ($h \to \infty$) will create a simple model (over smooth) that is similar to the results of parametric regression, with a small variance but a large bias. So, the optimum bandwidth value will be sought which will balance between bias and variance.

There are several efforts that can be made to help select the optimum bandwidth so that unbiased estimation results can be found, including the CV method, Silverman, AIC and BIC, and the Bootstrap method. In general, the CV method is a method that is often used for selecting the optimum bandwidth. This is because this method is quite easy to use and gives results with good performance [30]. If $\widehat{m_{-i}}(x_i)$

is the estimated regression curve without including the i-th observation, then for a certain value of h it can be calculated.

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{m}_{-i}(x_i)\right)^2 \quad (7)$$

The CV value based on the proposed *h* can determine an optimum bandwidth which is seen from the smallest CV value.

2.9. Measure of Method Accuracy

To determine which method is more accurate in smoothing age data, there are several measures that can be used such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), R-squared, Coefficient of determination, Confusion Matrix, Receiver Operating Characteristic (ROC), Area Under the ROC Curve (AUC-ROC), and so on. RMSE is one of the commonly used measures to measure the accuracy of a method. RMSE is used to see the accuracy between the smoothing results of aged data and observed data. A more accurate method is the one with the minimum RMSE.

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m(x_i)} - y_i\right)^2\right)^{\frac{1}{2}} \quad (8)$$

Description:

$\widehat{m(x_i)}$ = estimated value of the regression curve at observation i-th
y_i = the actual value of the i-th observation
n = number of x

3. Methodology

3.1. Data Collection Method

This research discusses the comparison of methods in dealing with the age heaping problem on age data. The data used in this study is SP2020 with the research variables being age at enumeration, age at death and total population. The coverage in this study is the entire territory of Indonesia. With the unit of analysis, namely individuals with an age range of 0-95 years. The number of observations in SP2020 was 275,773,770 people for the total population and 8,077,526 people for the number of population mortality.

3.2. Analysis Method

This study uses a comparative analysis method of smoothing methods on single age data with a nonparametric approach, namely the Kernel Smoothing and Local Polynomial Smoothing methods. Data processing using Open-Source Software RStudio. The packages used are Kernsmooth for Kernel Smoothing and Locpol for Local Polynomial Smoothing. The stages in the data analysis process are as follows:

1. Preparing SP2020 age data based on the total population and number of populations mortality.
2. Measuring the quality of age data with WI and MBI.
3. Smoothing process with Kernel Smoothing and Local Polynomial Smoothing methods.
4. Measuring the quality of smoothed age data.
5. Comparing the accuracy of the two methods based on the RMSE value.

4. Results and Discussion

4.1. SP2020 Population and Mortality Distribution

Figures 3 and 4 show the distribution of population and mortality by single age from SP2020. The highest age heaping is seen at the age of years, which means that the highest population based on SP2020

results is not at ages with 0 or 5 as the last digit. However, the results are different in Figure 3 where age heaping in the number distribution of SP2020 population mortality is prominent. Age heaping in the population mortality data starts at ages 40, 45, 50, 55, 60, and so on. The highest age heaping in SP2020 is at ages 0, 70, and 60.

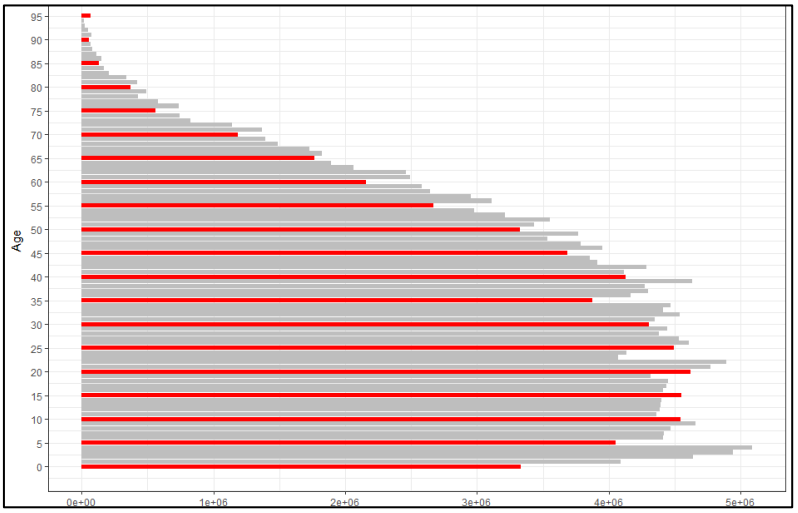

Figure 3. Population distribution based on single age from SP2020

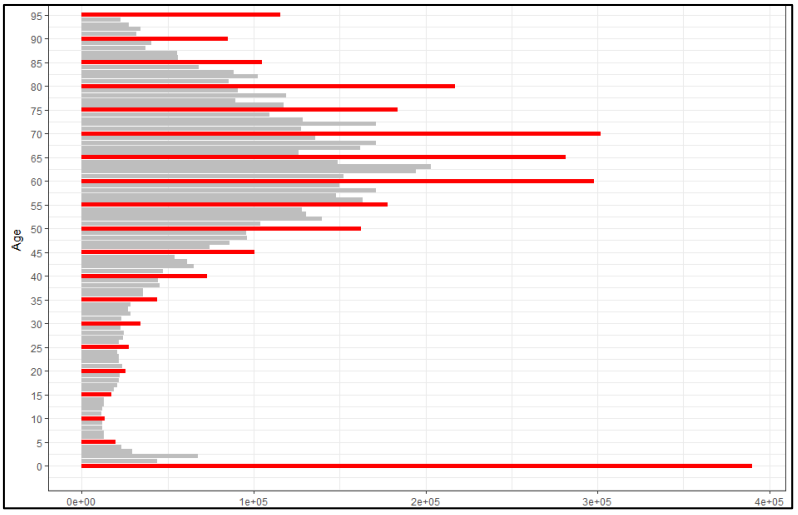

Figure 4. Population mortality distribution based on single age from SP2020

In the SP2020 population mortality count, deaths at the age of 0 years are very high. This shows that the infant mortality rate in Indonesia is still very high. In addition, most age heaping is located in the old age group. This is in line with previous research [7] which states that respondents with old age have a high potential to forget their age.

4.2. Age Data Quality Measurement Results with WI and MBI Before Smoothing Process

The quality of age data is underrepresented if only shown visually. Therefore, we measured the age data quality with WI and MBI. Table 2 shows that the age data quality of the total population is more accurate and better than the number of population mortality. Moreover, the tendency to mention age in the SP2020 population data with numbers 0 and 5 as the last digit is very minor or can be stated to be very accurate with a WI value of 100.97 and an MBI value of 3.47. For the SP2020 population mortality data, the data quality is very inaccurate and poor with a WI value of 177.20 and an MBI value of 21.37.

Table 2. WI and MBI values of SP2020 age data before smoothing process

	WI Value	Description	MBI Value	Description
Total Populations	100.97	Very Accurate	3.47	Good
Total Mortality	177.20	Very Inaccurate	21.27	Bad

4.3. Estimation of the Number of Population Mortality by Single Age in SP2020

Before smoothing the SP2020 single-age population mortality data, the optimum bandwidth was selected first. Next, the estimated SP2020 population mortality curve (Figure 5) from each smoothing method is shown. The value of h used varies, including 2, 3, 4, and 5 (Table 3). The optimum bandwidth is the bandwidth with the smallest CV value.

Table 3. CV value of SP2020 data with optimum bandwidth

WI Value	Data Quality
<105	Very Accurate
105 – 109.9	Accurate
110 – 124.5	Moderately Accurate
125 – 174.9	Inaccurate
>175	Very Inaccurate

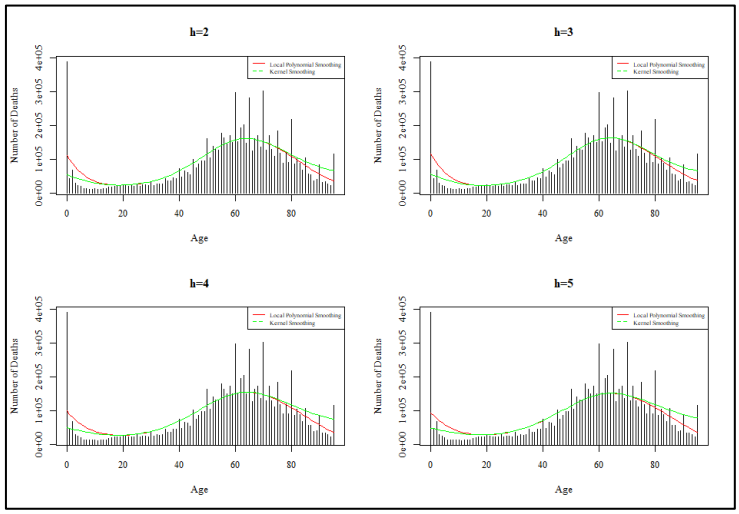

Figure 5. Estimation of the SP2020 single-age population mortality curve using Kernel Smoothing and Local Polynomial Smoothing methods and varying bandwidth values

In Figure 6, the smallest CV value is at a bandwidth of h = 3. In Figure 5, the results of the SP2020 data smoothing process, the Local Polynomial Smoothing method can cover the early age range than the Kernel Smoothing method. However, the Kernel Smoothing method is better at covering the late age range.

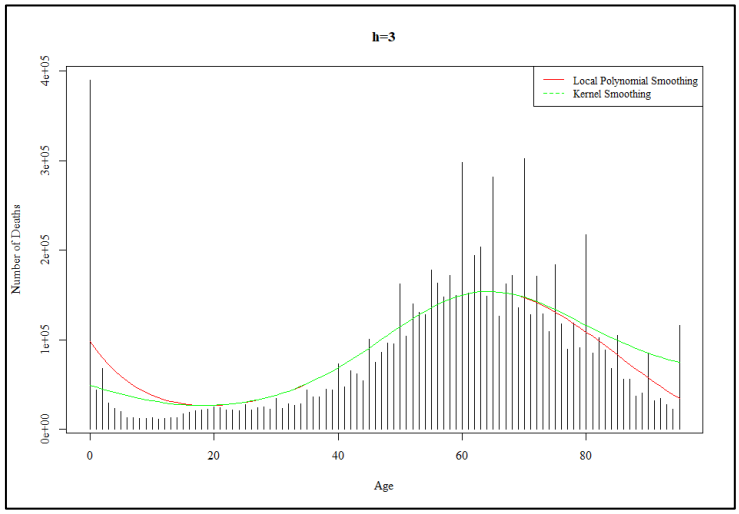

Figure 6. Optimal bandwidth for estimating the SP2020 single-age population mortality curve

4.4. Smoothing Result of Kernel Smoothing Method

Figure 7 shows the distribution of SP2020 population mortality after smoothing with the Kernel Smoothing method. The results obtained are that the curve of the population mortality number becomes smoother and there is no age heaping. Furthermore, from the two figures below, it is shown that the distribution of the mortality numbers is centered at the old age group.

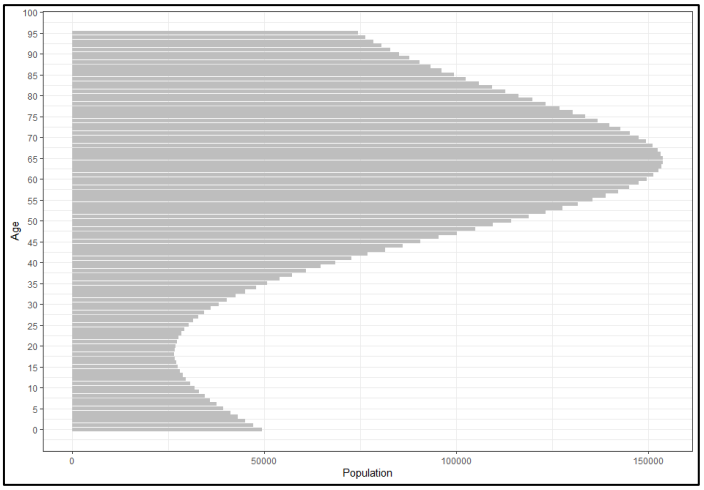

Figure 7. The population mortality distribution smoothing result based on single age from SP2020 using the Kernel Smoothing method

4.5. Smoothing Results of Local Polynomial Smoothing Method

Meanwhile, Figure 8 also shows the distribution of SP2020 population mortality that has been smoothed with the Local Polynomial Smoothing method. The results obtained are quite the same, the curve of population mortality number becomes smoother and there is no age buildup or age heaping. The Local Polynomial Smoothing method produces a smoother curve than the Kernel Smoothing method. In addition, it can also be seen that the population mortality number is centered at the old age group.

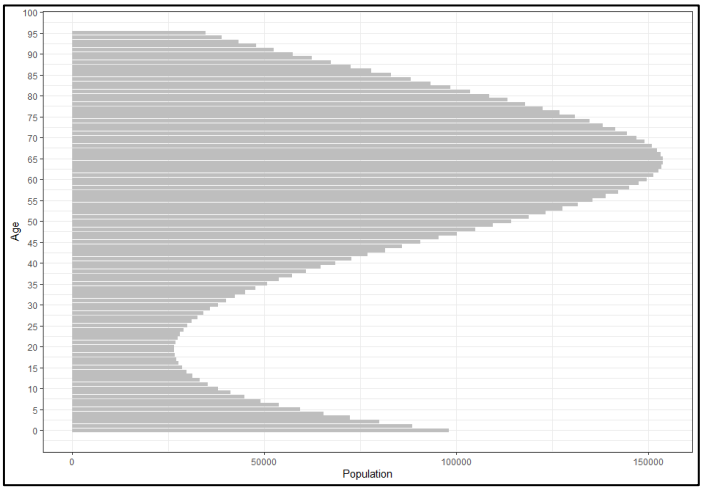

Figure 8. The population mortality distribution based on single age from SP2020 smoothing results using the Local Polynomial Smoothing method.

4.6. Results of Age Data Quality Measurement with WI and MBI After Smoothing Process

After the smoothing process, we calculate the quality of smoothed age data with both methods. Table 4 illustrates the results of measuring the age data quality based on WI and MBI. The results obtained are that the age data quality for the SP2020 population mortality counted with the Kernel Smoothing and Local Polynomial Smoothing methods has improved to be quite accurate and well. The results obtained from each data and method do not show much different numbers. Based on the results of measuring data quality with WI and MBI, it can be said that both methods are able to smooth age data and improve the quality of age data.

Table 4. Results of age data quality measurement with WI and MBI after smoothing process

Methods	WI Value	Description	MBI Value	Description
Kernel Smoothing	122.69	Moderately Accurate	2.41	Good
Local Polynomial Smoothing	122.70	Moderately Accurate	1.13	Good

Furthermore, a graph depicting the difference in MBI values for the mortality numbers in the SP2020 population (Figure 9) based on terminal digits is presented. In Figure 9, it can be observed that digits 0 and 5 have the highest MBI values compared to the other digits. This indicates the presence of age heaping in the mortality data in the SP2020 population. After undergoing smoothing processes, both with Kernel Smoothing and Local Polynomial Smoothing methods, the distribution of MBI values becomes more uniform. This suggests that age heaping in the mortality data in the SP2020 population has been addressed.

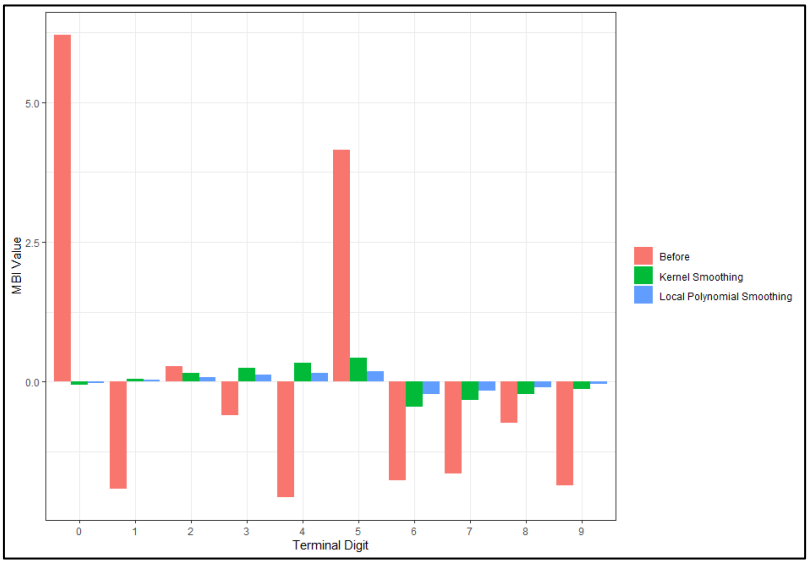

Figure 9. The MBI values based on terminal digits in the population mortality data based on single age from the SP2020 results.

4.7. Determining the Best Method

Based on Figure 10, the Local Polynomial Smoothing method exhibits a lower RMSE value compared to the Kernel Smoothing method when applied to the data of the 2020 population mortality numbers. Specifically, the Local Polynomial Smoothing method achieved an RMSE of 46,812 individuals, whereas the Kernel Smoothing method resulted in an RMSE of 50,097 individuals. This leads to the conclusion that the Local Polynomial Smoothing method offers greater accuracy in age data smoothing compared to the Kernel Smoothing method, thus enhancing the quality of age data for both the 2010 and 2020 population censuses.

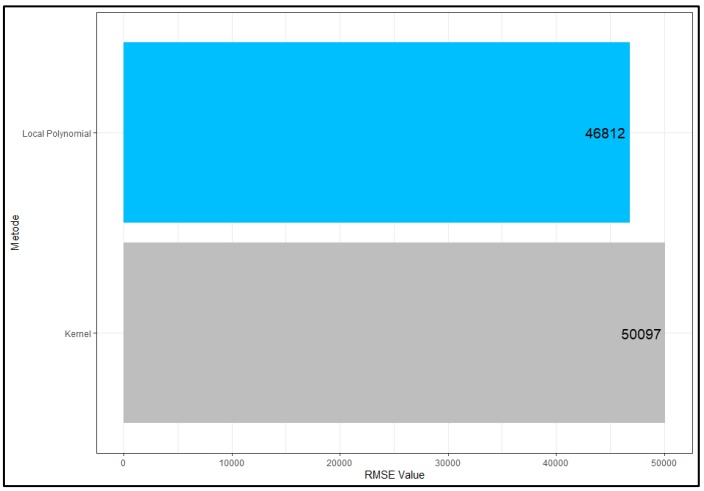

Figure 10. Comparison of Kernel Smoothing and Local Polynomial Smoothing methods in smoothing population mortality data based on single age from SP2020

5. Conclusion

Based on the results and discussion described in the previous chapter, the following conclusions can be drawn: the quality of population age data in the SP2020 results is quite accurate and good. However, the

data on the number of SP2020 population mortality fall into the category of very inaccurate and poor. The Kernel Smoothing and Local Polynomial Smoothing methods are successful in improving the quality of age data, namely based on the decrease in WI and MBI values after the smoothing process. The Local Polynomial Smoothing method produces better smoothing accuracy than the Kernel Smoothing method in the age data smoothing process. This can be seen from the RMSE value of the Local Polynomial Smoothing method which is smaller than the Kernel Smoothing method.

Meanwhile, this research has several suggestions, including for BPS to evaluate data collection and improve the quality of field officers so that the quality of age data becomes more accurate. The Local Polynomial Smoothing method can be considered for use as a smoothing method for age data. For the government to increase the availability of written information such as birth certificates and death certificates. For further research, it can compare the smoothing method from BPS, namely the Arriaga method with the Local Polynomial Smoothing method. Furthermore, if you want to measure the accuracy of age data, you can use other indices such as the bachi index. When using Kernel Smoothing or Local Polynomial Smoothing methods in the data smoothing process, other types of kernels can be chosen such as Uniform, Epanechnikov, and Triangular. For the selection of optimum bandwidth, other suggested methods include Silverman, AIC and BIC, or the Bootstrap method

References:

[1] K. O. Mason and L. G. Cope, "Sources of Age and Date-of-Birth Misreporting in the 1900 U.S. Census," 1987. [Online]. Available: http://read.dukeupress.edu/demography/article-pdf/24/4/563/905930/563mason.pdf

[2] BAPPENAS, "Menemukan , Mencatat , Melayani : Kelahiran dan Kematian di Indonesia (Studi dan Rencana Pelembagaan Identitas Hukum dan SIstem Pencatatan Sipil," Jakarta, 2016. [Online]. Available: https://sikompak.bappenas.go.id/pustaka/view/167/id/Menemukan, Mencatat, Melayani Kelahiran dan Kematian di Indonesia.pdf

[3] J. W. Budd and T. Guinnane, "Population Investigation Committee Intentional Age-Misreporting , Age-Heaping , and the 1908 Old Age Pensions Act in Ireland," vol. 45, no. 3, pp. 497–518, 1997, [Online]. Available: https://www.jstor.org/stable/2174865

[4] M. Lyons-Amos and T. Stones, "Trends in Demographic and Health Survey data quality: An analysis of age heaping over time in 34 countries in Sub Saharan Africa between 1987 and 2015," *BMC Research Notes*, vol. 10, no. 1. 2017. doi: 10.1186/s13104-017-3091-x.

[5] O. Fayehun, A. I. Ajayi, C. Onuegbu, and D. Egerson, "Age heaping among adults in Nigeria: evidence from the Nigeria Demographic and Health Surveys 2003-2013," *J. Biosoc. Sci.*, vol. 52, no. 1, pp. 132–139, 2020, doi: 10.1017/S0021932019000348.

[6] A. W. Kidane, "Digit Preference in African Survey Data and Their Impact on Parametric Estimates Digit Preference in African Survey Data and Their Impact on Parametric Estimates By Asmerom Kidane Department of Economics University of Dar es Salaam For presentation at the," no. January 2009, 2014, [Online]. Available: https://www.researchgate.net/publication/237798955_Digit_Preference_in_African_Survey_Data_and_Their_Impact_on_Parametric_Estimates

[7] B. A'Hearn, J. Baten, and D. Crayen, "Quantifying quantitative literacy: Age heaping and the history of human capital," *J. Econ. Hist.*, vol. 69, no. 3, pp. 783–808, 2009, doi: 10.1017/S0022050709001120.

[8] BAPPENAS and BPS, "Proyeksi Penduduk Indonesia Umur Tertentu dan Umur Satu Tahunan 2010-2025," 2014. [Online]. Available: https://www.ptonline.com/articles/how-to-get-better-mfi-results

[9] D. F. Heitjan and D. B. Rubin, "Inference from coarse data via multiple imputation with application to age heaping," *J. Am. Stat. Assoc.*, vol. 85, no. 410, pp. 304–314, 1990, doi: 10.1080/01621459.1990.10476202.

[10] BAPPENAS and BPS, "Proyeksi penduduk menurut umur tunggal dan umur tertentu," Jakarta, 2008.[Online].Available: https://www.bps.go.id/publication/2008/09/04/60f9ba79ac8fd1705acb3117/proyeksi-penduduk-menurut-umur-tunggal-dan-umur-tertentu-2005-2015.html
[11] BPS, *Proyeksi Penduduk Kota Bandung 2020-2035 Hasil Sensus Penduduk 2020*, no. 1. Jakarta, 2029.[Online].Available: https://www.bps.go.id/publication/2023/05/16/fad83131cd3bb9be3bb2a657/proyeksi-penduduk-indonesia-2020-2050-hasil-sensus-penduduk-2020.html
[12] E. E. Arriaga, "Population Analysis with Microcomputers.," *Present. Tech.*, vol. I, p. 398, 1994, [Online]. Available: https://www2.census.gov/software/pas/documentation/pamvi-archive.pdf
[13] T. Spoorenberg, "Reverse survival method of fertility estimation: An evaluation," *Demogr. Res.*, vol. 31, no. 1, pp. 217–246, 2014, doi: 10.4054/DemRes.2014.31.9.
[14] M. . Thwin, "Population Projections and Some Aspects on Labour Pool Population for Myanmar," 2022. [Online]. Available: https://meral.edu.mm/record/8362/file_preview/Mg Mg Thwin%2C Stats.%2C PhD. Ah 1%2C 12-9-2022.pdf?allow_aggs=True
[15] U.S. Census Bureau and USAID, "Age and Sex Structure: Smoothing Techniques to Correct for Age Misreporting," pp. 1–19, 2020, [Online]. Available: http://www.aitrs.org/sites/default/files/Age and Sex Structure.pdf
[16] J. W. Vaupel and V. C. Romo, "Decomposing change in life expectancy: A bouquet of formulas in honor of Nathan Keyfitz's 90th birthday," *Demography*, vol. 40, no. 2, pp. 201–216, 2003, doi: 10.1353/dem.2003.0018.
[17] United Nations, *World Mortality 2019*, vol. 2, no. 3. 2019. [Online]. Available: https://www.un.org/development/desa/pd/content/world-mortality-2019-data-booklet
[18] Firdaus and E. T. Astuti, "Local Polynomial Smoothing Untuk Mengatasi Masalah Age Heaping Data Jumlah Kematian Menurut Umur Hasil Sensus Penduduk 2010," *J. Apl. Stat. Komputasi Stat.*, vol. 9, no. 2, 2017, [Online]. Available: https://jurnal.stis.ac.id/index.php/jurnalasks/article/view/141
[19] R. J. Myers, "Accuracy of Age Reporting in the 1950 United States Census," *J. Am. Stat. Assoc.*, vol. 49, no. 268, pp. 826–831, 1954, doi: 10.1080/01621459.1954.10501237.
[20] A. F. Arifin, *Pemodelan Nilai Ekspor Kelapa Sawit di Indonesia Menggunakan Smoothing Kernel*. Jakarta, 2020.
[21] A. A. Vallarino, *Perbandingan Metode Smoothing Kernel dan Spline Dalam Peramalan Indeks Harga Saham Gabungan Indonesia Tahun 2017*. Jakarta, 2017.
[22] D. H. R. Spennemann, "Age Heaping among Indian Hawkers in South-eastern Australia and their source communities in the Punjab," *J. Sikh Punjab Stud.*, vol. 24, no. 1–2, pp. 149–202, 2017, [Online]. Available: http://www.giss.org/jsps_vol_24.html
[23] K. K. West, J. G. Robinson, and M. Bentley, "Did Proxy Respondents Cause Age Heaping in the Census 2000?," *Am. Stat. Assoc. Sect. Surv. Res. Methods*, pp. 3658–3665, 2000, [Online]. Available: http://www.asasrms.org/Proceedings/y2005/files/JSM2005-000443.pdf
[24] M. Zelnik, "Age Heaping in the United States Census : 1880-1950," *Wiley behalf Milbank Meml.*, vol. 39, no. 3, pp. 540–573, 1961, doi: https://doi.org/10.2307/3348729.
[25] M. Szołtysek, R. Poniat, and S. Gruber, "Age heaping patterns in Mosaic data," *Hist. Methods*, vol. 51, no. 1, pp. 13–38, 2018, doi: 10.1080/01615440.2017.1393359.
[26] A. J. Jowett and Y.-Q. Li, "Age - Heaping : Contrasting Patterns from China," *Springer*, vol. 28, no. 4, pp. 427–442, 1992, doi: https://doi.org/10.1007/BF00273112.
[27] UNSD, "Evaluation of Age and Sex Distribution Data Evaluation method of age and sex

distribution data," in *United Nations Workshop on Census Data Evaluation for English Speaking African Countries*, 2012, pp. 26–32.

[28] B. . Mukherjee and B. K. Mukhopadhyay, "A STUDY OF DIGIT PREFERENCE AND QUALITY OF AGE DATA," *Genus*, vol. 44, pp. 1–2, 1988.

[29] H. S. Shryock, J. S. Siegel, and E. A. Larmon, "The Methods and Materials of Demography," in *2*, 1973, pp. 397–398. [Online]. Available: https://books.google.com/books?hl=id&lr=&id=-wvLH72ahlUC&oi=fnd&pg=PA371&dq=Methods+and+Materials+of+Demography&ots=SyrTUNpn-Z&sig=sYN7Depv2jA0NDznvyZ77o4ahe8

[30] E. T. Astuti, "Estimator Polinomial Lokal Dalam Model Regresi Nonparametrik Untuk Data Count," Institut Teknologi Sepuluh Nopember, 2013.

[31] D. M. Gujarati, *Mechanical control of Spin States in Spin-1 molecules and the underscreened kondo effect*, vol. 328, no. 5984. 2004. doi: 10.1126/science.1186874.

[32] R. L. Eubank, *Nonparametric regression and spline smoothing*. 1999. [Online]. Available: https://www.ptonline.com/articles/how-to-get-better-mfi-results

[33] B. Lestari, Fatmawati, I. N. Budiantara, and N. Chamidah, "Smoothing parameter selection method for multiresponse nonparametric regression model using smoothing spline and Kernel estimators approaches," *J. Phys. Conf. Ser.*, vol. 1397, no. 1, 2019, doi: 10.1088/1742-6596/1397/1/012064.

[34] W. Hardle, *Applied Nonparametric Regression*. Cambridge: Cambridge University Press, 1990. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=8TprPX-7PSMC&oi=fnd&pg=PR11&dq=applied+nonparametric+regression+hardle&ots=T8Z-yMTYDq&sig=7ncK7VSaqUp2tQyWh9AIDkTnfis&redir_esc=y#v=onepage&q=applied nonparametric regression hardle&f=false

[35] J. Fan and I. Gijbels, *Local Polynomial Modelling and Its Application*, First Edit. New York: Chapman and Hall, 1996. doi: https://doi.org/10.1201/9780203748725.

[36] J. O. Ramsay, "Kernel smoothing approaches to nonparametric item characteristic curve estimation," *Psychometrika*, vol. 56, no. 4, pp. 611–630, 1991, doi: 10.1007/BF02294494.

[37] B. W. Silverman, "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting," *J. R. Stat. Soc. Ser. B*, vol. 47, no. 1, pp. 1–21, 1985, doi: https://doi.org/10.1111/j.2517-6161.1985.tb01327.x.

[38] B. A. Turlach, "Shape constrained smoothing using smoothing splines," *Comput. Stat.*, vol. 20, no. 1, pp. 81–103, 2005, doi: 10.1007/BF02736124.

[39] W. S. Hasani and A. Ubaidillah, "Pembangunan Package R untuk Small Area Estimation Pendekatan Nonparametrik Berbasis Kernel Nadaraya-Watson," *Semin. Nas. Off. Stat.*, vol. 2022, no. 1, pp. 1315–1326, 2022, doi: 10.34123/semnasoffstat.v2022i1.1545.

[40] S. Ramadhana, *Perbandingan Model Regresi Nonparametrik Menggunakan Pendekatan Truncated Spline dan Kernel Gaussian Pada Data Pencilan*. Malang, 2018.

[41] I. Puspitasari, Suparti, and Y. Wilandari, "Analisis indeks harga saham gabungan (IHSG) dengan menggunakan model regresi kernel," *J. Gaussian*, vol. 1, no. 1, pp. 93–102, 2012, [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/article/view/577/579