Mathematics and Statistics 1(2): 64-73, 2013
DOI:10.13189/ms.2013.010209 http://www.hrpub.org

Representing Data Distributions with a Nonparametric Kernel Density: The Way to Estimate the Optimal Oil Contents of Palm Mesocarp at Various Periods

Divo Dharma Silalahi[1,2,*], Putri Aulia Wahyuningsih[1], Fahri Arief Siregar[1]

[1]SMART Research Institute, PT SMART Tbk, Indonesia
[2]Institute of Statistics, University of the Philippines Los Baños, Philippines
*Corresponding Author: divosilalahi@yahoo.co.id

Abstract The most popular nonparametric density estimates is kernel density estimate. This estimate depends on the bandwidth choice which was given the optimization to kernel optimality process. We proposed Epanechnikov kernel which is the most optimal kernel in the AMISE. The resample data as replicate samples has been obtained by using bootstrap mechanism to provide the information about the sampling distribution. Then the resample data was used in Epanechnikov kernel simulation to estimate the optimal solution. This study was simulated using oil contents (%) data at various periods after pollination. The oil contents (%) were obtained by extraction of oil palm mesocarp. The result show that, Epanechnikov kernel using resamples data from bootstrap could be used for nonparametric optimization cases such as oil content (%) of oil palm mesocarp.

Keywords Nonparametric, Epanechnikov Kernel, Density Estimation, Bootstrap, Optimization, Oil Palm Mesocarp

1. Introduction

Bootstrap mechanism was introduced in 1979 as a computational statistical technique that allows making some inferences from data without making strong distributional assumptions [1]. As problem with small sample size, bootstrap mechanism can be used to resample the set of data with replacement to estimate the statistic's sampling distribution. The sampling distribution if it can be determined may then be used to estimate standard errors and confidence intervals for that particular statistic. Although the data point has been used, it is not deleted from the original data set or, using the usual terminology, which is replaced. As a result, the same observation may be included in the resample data set. In this case, we used the bootstrap mechanism to resample oil contents of palm mesocarp data at 10 periods (week: 6, 14, 16, 18, 19, 20, 21, 22, 24, 26) after pollinations. This field study was conducted in South Kalimantan, Indonesia. The data was obtained from laboratory analysis as extracted process to fresh fruit oil palm samples with variety is progeny-66 and genotype-03. We only have 3 replicates per period, and then we generated the data using bootstrap become 250 replicates.

Table 1. Descriptive statistic oil contents of %ODM

Week	ODM(%)_ Rep1	ODM(%)_ Rep2	ODM(%)_ Rep3	Mean	SD
6	2.55	2.31	2.58	2.48	0.17
14	3.86	3.54	5.05	4.15	0.22
16	5.11	10.94	8.25	8.10	4.12
18	12.56	19.74	11.53	14.61	5.08
19	29.43	26.22	32.25	29.30	2.27
20	47.51	47.50	47.72	47.58	0.01
21	59.62	63.79	64.72	62.71	2.95
22	69.03	74.99	72.80	72.27	4.22
24	72.09	76.08	73.80	73.99	2.82
26	73.93	76.21	74.38	74.84	1.61

*ODM: Oil/Dry Matter

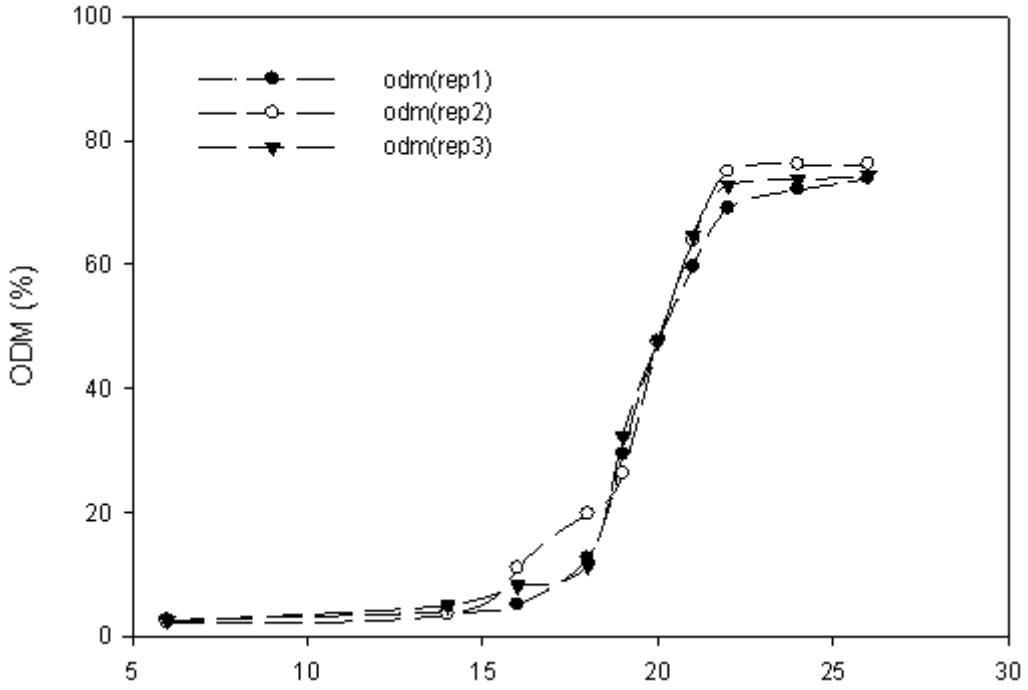

Figure 1. Oil contents of palm mesocarp in 3 replicates

A very natural use of density estimates is in the informal investigation of the properties set of resamples data. Density estimates can give valuable indication of such features as skewness and multimodality in the data. In some cases they will give conclusion that may then be regarded as self-evidently true further to data collection. This density referred to Epanechnikov kernel density estimates that used in simulation to estimate the optimal oil contents of palm mesocarp at those periods. The analysis was processed using auxiliary software such that: R version 2.15.2 and Stata 12.

2. Bootstrap

2.1. Definition

The Bootstrap is a resample mechanism designed to provide information about the sampling distribution of a functional $T(X_1, X_2, \ldots, X_n, F)$ where $X_1, X_2, \ldots, X_n$ are sample observations and F is CDF from which $X_1, X_2, \ldots, X_n$ are independent observations. The bootstrap is not limited to the iid situations. It has been studied for various kinds of dependent data and complex situations. In fact, this versatile nature of the bootstrap is the principal reason for its popularity. There are numerous texts and reviews of bootstrap theory and methodology, at varied technical level [1] [2].

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$ and $T(X_1, X_2, \ldots, X_n, F)$ is a functional,

e.g.,

$$T(X_1, X_2, \ldots, X_n, F) = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma},$$

where $\mu = E_F(X_1)$ and $\sigma^2 = Var_F(X_1)$.

In statistical problems, we frequently need to know something about the sampling distribution of T,

e.g.,

$$P_F(T(X_1, X_2, \ldots, X_n, F) \le t).$$

If we have replicated samples from the population, resulting in a series of values for the statistic T, then we could form estimates of $P_F(T \le t)$ by counting how many of the T_i's are $\le t$. But statistical sampling is not done using that way. Replicate samples were not usually obtained, but otherwise only one set of data of some size n [1].

A large sample from a finite population should be well representative of the full population itself. Suppose for some number S, we draw S resample of size n from the original sample. The resample were denoted from the original sample as

$$(X^*_{11}, X^*_{12}, \ldots, X^*_{1n}),\ (X^*_{21}, X^*_{22}, \ldots, X^*_{2n}),\ \ldots,$$
$$(X^*_{S1}, X^*_{S2}, \ldots, X^*_{Sn}),$$

with corresponding values $T^*_1, T^*_2, \ldots, T^*_S$ for the functional T, we can use simple frequency based estimates such as $\frac{\#\{j : T^*_J \le \mathrm{t}\}}{B}$ to estimate $P_F(T \le t)$ [2].

2. Bootstrap Distribution and Consistency

The formal definition of the bootstrap distribution is the following

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$ and $T(X_1, X_2, \ldots, X_n, F)$ is a given functional, the ordinary bootstrap distribution of T is defined as

$$H_{\text{Boot}}(\text{x}) = P_{Fn}(T(X_1, X_2, \ldots, X_n, F) \le x)$$

where $(X_1^*, \ldots, X_n^*)$ is an iid sample of size n from the empirical CDF F_n. P_* is common used to denote probabilities under the bootstrap distribution [4].

At first glance, the idea appears to be a bit too simple to actually work. But one has to have a definition for what one means the by the bootstrap working in a given situation. For estimating the CDF of a statistic, one should want $H_{Boot}(x)$ to be numerically close to the true CDF $H_n(x)$ of T. For a general metric ρ the definition is the following

Let F, G be two CDFs on a sample space X.

Let $\rho(F,G)$ be a metric on the space of CDFs on X.

For $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$, and a given functional $T(X_1, X_2, \ldots, X_n, F)$, let

$$H_n(x) = P_F\left(\left(X_1, X_2, \ldots, X_n \overset{iid}{\sim} F\right) \le x\right) \quad (1)$$

$$H_{Boot}(x) = P_*\left(\left(X_1^*, X_2^*, \ldots, X_n^* \overset{iid}{\sim} F\right) \le x\right) \quad (2)$$

The bootstrap is weakly consistent under ρ for T if $\rho(H_n, H_{Boot}) \overset{P}{\Rightarrow} 0$ as $n \to \infty$. Otherwise, bootstrap is strongly consistent under ρ for T if $\rho(H_n, H_{Boot}) \overset{a.s.}{\Rightarrow} 0$ [4].

2.3. Bootstrap Confidence Interval

In a series of article, Efron [5] [6] [7] [8] [9] has introduced and refined the percentile method. This was using bootstrap calculations to set approximate confidence interval for scalar parameters. These refinements of the percentile method are the bias-corrected (BC) percentile method and the accelerated bias-corrected (BC_a) percentile method. Efron's approach is to first develop these procedures in the simple context of a parametric model indexed by a scalar parameter, for which there are no nuisance parameters present, and then to adapt them for application in multiparameter families and nonparametric situations.

For a review of the percentile method in the simplest case, suppose that Let $\hat{G}$ be the cumulative distribution of $\hat{\theta}^*$. The $1-2\alpha$ percentile interval is defined by the α and $1-\alpha$ percentiles of $\hat{G}$:

$$\left(\hat{\theta}_{\%,lo}, \hat{\theta}_{\%,up}\right) = \left(\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)\right) \quad (3)$$

Since by the definition

$$\hat{G}^{-1}(\alpha) = \left(\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}\right)$$

the $100 \cdot \alpha th$ percentile of the bootstrap distribution, we can also write the percentile interval as

$$\left(\hat{\theta}_{\%,lo}, \hat{\theta}_{\%,up}\right) = \left(\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}\right) \quad (4)$$

These are the ideal bootstrap situation in which the number of bootstrap replications is infinite. In practice we must use some finite number of replications. Then the arguments in favor of the percentile interval should translate into better coverage performance.

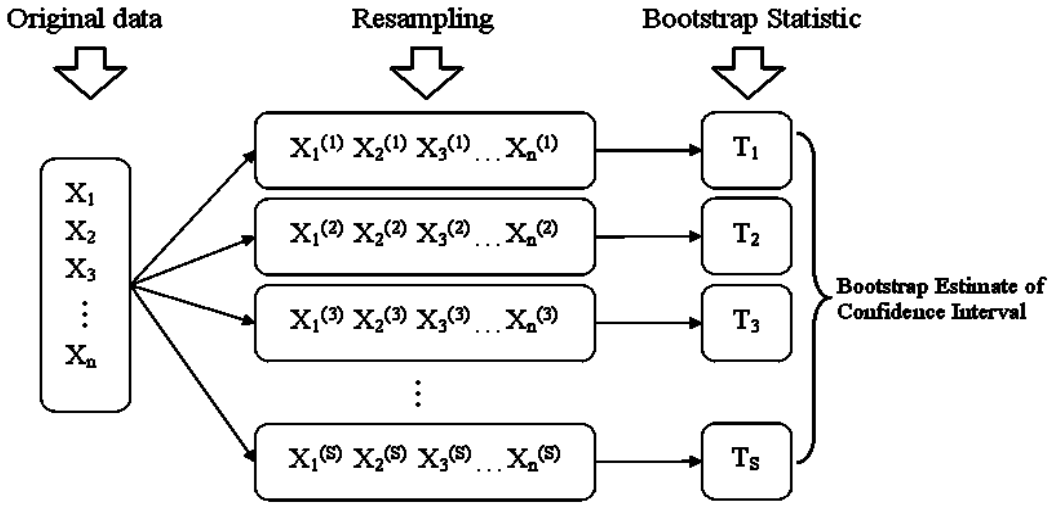

Figure 2. Bootstrap scheme of work

3. Kernel

3.1. Kernel Density Estimation

We limit the kernel overview only to the nonparametric case as this is the one primarily used in computer graphics, in other side the illumination of a scene seldom can be described adequately by a simple function. A general nonparametric density estimator is the kernel estimator. The kernel estimator approximates a density function by weighting the samples of a dataset by their distance to the position. This is done using a kernel function.

Let X is a random variable with continuous distribution $F(x)$ and density $f(x)=\frac{d}{dx}F(x)$. The objective is to estimate $f(x)$ from a random sample $\{X_1,\ldots,X_n\}$.

The distribution function $F(x)$ is naturally estimated by the Empirical Density Function $\mathrm{F(x)}=\frac{1}{\mathrm{n}}\sum_{i=1}^{n}1(X_i\le x)$. It might seem natural to estimate the density $f(x)$ as the derivative of $\hat{F}(x)$, $\frac{d}{dx}\hat{F}(x)$, but this estimator would be a set of mass points, not a density, and as such is not a useful estimate of $f(x)$ [11]. Instead, consider a discrete derivative for some small $h>0$, suppose $\hat{f}(x)=\frac{\hat{F}(x+h)-\hat{F}(x-h)}{2h}$; we can write this as

$$\frac{1}{2\mathrm{nh}}\sum_{\mathrm{i=1}}^{\mathrm{n}}1\left(\mathrm{x+h<X_i\le x+h}\right)=\frac{1}{2\mathrm{nh}}\sum_{\mathrm{i=1}}^{\mathrm{n}}1\left(\frac{\left|\mathrm{X_i-x}\right|}{\mathrm{h}}\le 1\right)$$

$$\frac{1}{\mathrm{nh}}\sum_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{k}\left(\frac{\left|\mathrm{X_i-x}\right|}{\mathrm{h}}\right)$$

$$\hat{\mathrm{f}}\left(\mathrm{x}\right)=\frac{1}{\mathrm{nh}}\sum_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{k}\left(\frac{\mathrm{X_i-x}}{\mathrm{h}}\right) \tag{4}$$

where $k(u)$ is a kernel function.

$k(u)$ is a kernel function if $k(u)=k(-u)$ symmetric about zero, $\int_{-\infty}^{\infty}k(u)du=1$ and $\int_{-\infty}^{\infty}k(u)du=0$. This will be focused only on the case where $k(u)\ge 0$ so that $k(u)$ is a symmetric density with zero mean. When $k(u)\ge 0$ it is called a second order kernel and these is the most common used in applications. The most important choice is the bandwidth $h>0$ which controls the amount of smoothing. If h is large, there is a lot of smoothing, and otherwise if h is small there is less smoothing [11].

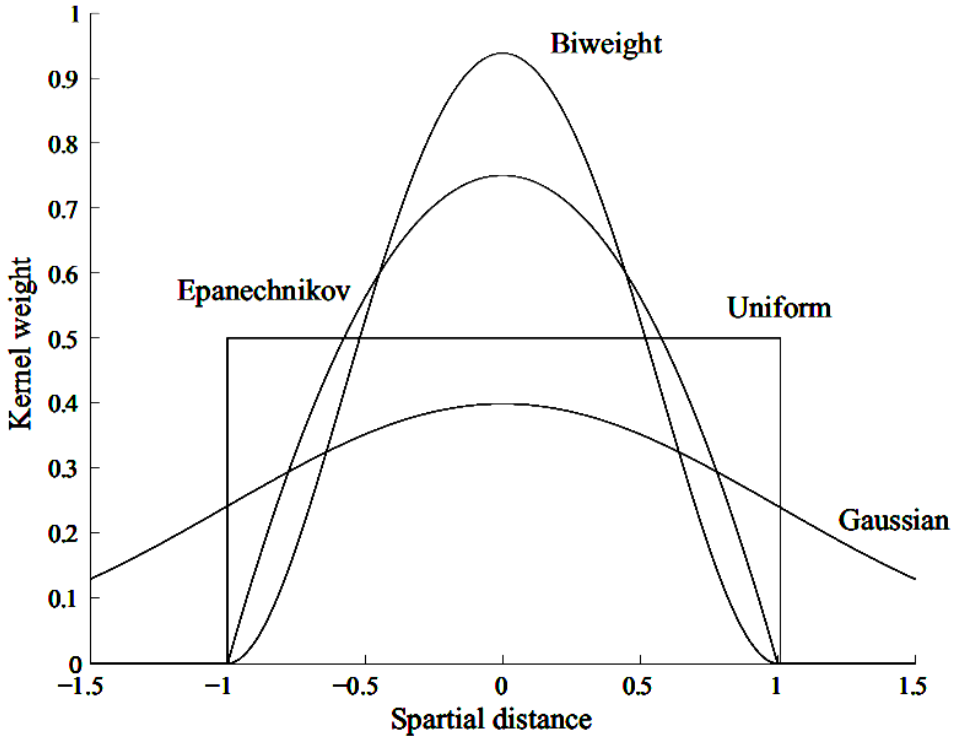

Figure 3. Several pattern of kernel density

Table 2. Several types of kernel density estimation

	$k(u)$	$\int u^2 k(u)du$	$\int k^2(u)du$
Biweight	$\mathrm{k(u)} = \frac{15}{16}(1-\mathrm{u}^2)^2 1_{\{\ \lvert u\rvert \le 1\ \}}$	$1/7$	$5/7$
Epanechnikov	$k(u) = \frac{3}{4}(1-u^2)^2 1_{\{\lvert u\rvert \le 1\}}$	$1/5$	$3/5$
Gaussian	$\mathrm{k(u)} = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}u^2}$	1	$1/2\sqrt{\pi}$
Uniform	$k(u) = \frac{1}{2} 1_{\{\lvert u\rvert \le 1\}}$	$1/3$	$1/2$

Based on some literature for kernel smoothing and optimization [10], performance of kernel is measured by Mean Integrated Squared Error (MISE) or Asymptotic MISE (AMISE). Epanechnikov kernel is the optimal kernel in the AMISE. Which is kernel efficiency is measured in comparison to Epanechnikov kernel. Based on this literature we used Epanechnikov kernel as a weighting function with

$$k(u) = \begin{cases} \frac{3}{4}(1-u^2), & |u| \le 1 \\ 0 & otherwise \end{cases} \tag{5}$$

3.2. Density Estimator

First if $k(u)$ is non-negative then it is easy to see that $\hat{f}(x) \ge 0$. However, this is not guaranteed if k is a higher order kernel. In this case it is possible that $\hat{f}(x) < 0$ for some values of x. When this happens it is prudent to zero out the negative bits and then rescale [11]:

$$\hat{f}(x) = \frac{\hat{f}(x)1(\hat{f}(x) \ge 0)}{\int_{-\infty}^{\infty} \hat{f}(x)1(\hat{f}(x) \ge 0)dx} \tag{6}$$

$\hat{f}(x)$ is non-negative yet has the same asymptotic properties as $\hat{f}(x)$. Since the integral in the denominator is not analytically available this needs to be calculated numerically.

$$\int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{X_i - x}{h}\right) dx = \int_{-\infty}^{\infty} k(u)du = 1$$

Thus

$$\begin{aligned} \int_{-\infty}^{\infty} \hat{\mathrm{f}}(\mathrm{x})\mathrm{dx} &= \int_{-\infty}^{\infty} \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{1}{\mathrm{h}} \mathrm{k}\left(\frac{\mathrm{X_i} - \mathrm{x}}{\mathrm{h}}\right) \mathrm{dx} \\ &= \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \int_{-\infty}^{\infty} \frac{1}{\mathrm{h}} \mathrm{k}\left(\frac{\mathrm{X_i} - \mathrm{x}}{\mathrm{h}}\right) \mathrm{dx} \\ &= \frac{1}{n} \sum_{i=1}^{n} 1 = 1 \end{aligned}$$

As claimed. Thus $\hat{f}(x)$ is a valid density function when k is non-negative [12].

We can also calculate the numerical moments of the density $\hat{f}(x)$. By using the change of variable $u = \frac{(X_i - x)}{h}$, the mean of the estimated density is

$$\begin{aligned} \int_{-\infty}^{\infty} \mathrm{x}\hat{\mathrm{f}}(\mathrm{x})\mathrm{dx} &= \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \int_{-\infty}^{\infty} \mathrm{x} \frac{1}{\mathrm{h}} \mathrm{k}\left(\frac{\mathrm{X_i} - \mathrm{x}}{\mathrm{h}}\right) \mathrm{dx} \\ &= \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \int_{-\infty}^{\infty} \left(\frac{\mathrm{X_i} + \mathrm{uh}}{\mathrm{h}}\right) \mathrm{k(u)du} \end{aligned} \tag{7}$$

$$\begin{aligned} &= \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i} \int_{-\infty}^{\infty} \mathrm{k(u)du} + \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{h} \int_{-\infty}^{\infty} \mathrm{u\,k(u)du} \\ &= \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i} \end{aligned} \tag{7}$$

The second moment of the estimated density is

$$\begin{aligned} \int_{-\infty}^{\infty} \mathrm{x}^2 \hat{\mathrm{f}}(\mathrm{x})\mathrm{dx} &= \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \int_{-\infty}^{\infty} \mathrm{x}^2 \frac{1}{\mathrm{h}} \mathrm{k}\left(\frac{\mathrm{X_i} - \mathrm{x}}{\mathrm{h}}\right) \mathrm{dx} \\ &= \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \int_{-\infty}^{\infty} \left(\mathrm{X_i} + \mathrm{uh}\right)^2 \mathrm{k(u)du} \\ &= \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i}^2 + \frac{2}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i h} \int_{-\infty}^{\infty} \mathrm{k(u)du} + \\ &\quad \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{h}^2 \int_{-\infty}^{\infty} \mathrm{u}^2 \mathrm{k(u)du} \\ &= \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i}^2 + \mathrm{h}^2 \mathrm{k}_2(\mathrm{k}) \end{aligned} \tag{8}$$

It follows that the variance of the density $\hat{f}(x)$ is

$$\begin{aligned} &= \int_{-\infty}^{\infty} \mathrm{x}^2 \hat{\mathrm{f}}(\mathrm{x})\mathrm{dx} - \left(\int_{-\infty}^{\infty} \mathrm{x}\,\hat{\mathrm{f}}(\mathrm{x})\mathrm{dx}\right)^2 \\ &= \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i}^2 + \mathrm{h}^2 \mathrm{k}_2 - \left(\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i}\right)^2 = \sigma^2 + \mathrm{h}^2 \mathrm{k}_2(\mathrm{k}) \end{aligned} \tag{9}$$

$\hat{\sigma}^2$ is the sample variance. Thus the density estimate inflates the sample variance by the factor $h^2 k_2(k)$.

3.3. Estimation $Bias(\hat{f}(x))$

It is useful to observe that expectations of kernel transformations, which can be written as integrals that take the form of a convolution of the kernel and the density function [12] [13] [14]:

$$E\frac{1}{h}k\left(\frac{X_i - x}{h}\right) = \int_{-\infty}^{\infty}\frac{1}{h}k\left(\frac{z-x}{h}\right)f(z)dz$$

Using the change of variables $u = (z-x)/h$ this equals to

$$\int_{-\infty}^{\infty} k(u)f(x+hu)du$$

By the linearity of the estimator we can see

$$\begin{aligned}\mathrm{E}\,\hat{\mathrm{f}}(\mathrm{x}) &= \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{E}\frac{1}{\mathrm{h}}\mathrm{k}\left(\frac{\mathrm{X_i}-\mathrm{x}}{\mathrm{h}}\right)\\ &= \int_{-\infty}^{\infty}\mathrm{k(u)f(x+h\,u)du}\end{aligned} \tag{10}$$

This integral (typically) is not analytically solvable, so we approximate it using Taylor expansion [13] [14] of $f(x+hu)$ in the argument hu, which is valid as $hu \to 0$. For a v'th order kernel we take the expansion out to the v'th term

$$f(x+hu) = f(x) + f^{(1)}(x)hu + \frac{1}{2}f^{(2)}(x)h^2u^2 + \frac{1}{3!}f^{(3)}(x)h^3u^3 + \ldots + \frac{1}{v!}f^{(v)}(x)h^vu^v + o(h^v)$$

The remainder is of smaller order than h^v as $h \to \infty$, which is written as $o(h^v)$. Integrating term by term and using $\int_{-\infty}^{\infty} k(u)du = 1$ and definition $\int_{-\infty}^{\infty} k(u)u^j du = k_j(k)$

$$\begin{aligned}\int_{-\infty}^{\infty} k(u)f(x+hu)du &= f(x) + f^{(1)}(x)hk_1(k) + \frac{1}{2}f^{(2)}(x)h^2k_2(k) + \frac{1}{3!}f^{(3)}(x)h^3k_3(k) + \ldots + \frac{1}{v!}f^{(v)}(x)h^vk_v(k) + o(h^v)\\ &= f(x) + \frac{1}{v!}f^{(v)}(x)h^vk_v(k) + o(h^v)\end{aligned}$$

This means that

$$\begin{aligned}\mathrm{E}\,\hat{\mathrm{f}}\,(\mathrm{x}) &= \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{E}\frac{1}{\mathrm{h}}\mathrm{k}\left(\frac{\mathrm{X_i}-\mathrm{x}}{\mathrm{h}}\right)\\ &= \mathrm{f(x)} + \frac{1}{\mathrm{v!}}\mathrm{f^{(v)}(x)h^vk_v(k)} + \mathrm{o(h^v)}\end{aligned}$$

The bias of $\hat{f}(x)$ is

$$\begin{aligned}Bias(\hat{f}(x)) &= E\hat{f}(x) - f(x)\\ &= \frac{1}{v!}f^{v}(x)h^vk_v(k) + o(h^v)\end{aligned} \tag{11}$$

With second order kernels, can be simplified to [12] [13] [14]

$$Bias(\hat{f}(x)) = \frac{1}{2}f^{(2)}(x)h^2k_{(2)}(k) + o(h^4) \tag{12}$$

3.4. Estimation $var\left(\hat{f}(x)\right)$

Since the kernel estimator is a linear estimator, and $k\left(\frac{X_i - x}{h}\right)$ is iid [12],

$$\begin{aligned}var\left(\hat{f}(x)\right) &= \frac{1}{nh^2}var\left(k\left(\frac{X_i-x}{h}\right)\right)\\ &= \frac{1}{nh^2}E\,k\left(\frac{X_i-x}{h}\right)^2 - \frac{1}{\mathrm{n}}\left(\frac{1}{\mathrm{h}}\mathrm{E}\;\mathrm{k}\left(\frac{\mathrm{X_i}-\mathrm{x}}{\mathrm{h}}\right)\right)^2\end{aligned} \tag{13}$$

From our analysis of bias we know that

$$\frac{1}{h}Ek\left(\frac{X_i-x}{h}\right) = f(x) + o(1)$$

so the second term is $O\left(\frac{1}{n}\right)$ [13] [14]

Taylor expansion

$$\begin{aligned}\frac{1}{h}Ek\left(\frac{X_i-x}{h}\right)^2 &= \frac{1}{h}\int_{-\infty}^{\infty}k\left(\frac{z-x}{h}\right)^2 f(z)dz\\ &= \int_{-\infty}^{\infty}k(u)^2 f(x+hu)du\\ &= \int_{-\infty}^{\infty}k(u)^2 (f(x)+O(h))du\\ &= f(x)R(k) + O(h)\end{aligned}$$

Where $R(k) = \int_{-\infty}^{\infty}k(u)^2du$ then $var\left(\hat{f}(x)\right)$ can be simplified to be

$$var\left(\hat{f}(x)\right) = \frac{f(x)R(k)}{nh} + O\left(\frac{1}{n}\right) \tag{14}$$

3.5. Mean Squared Error

A common and convenient measure of estimation precision is the mean squared error [15]

$$\begin{aligned} MSE\left(\hat{f}(x)\right) &= E\left(\hat{f}(x)-f(x)\right)^2 \\ &= Bias(\hat{f}(x))^2 + var(\hat{f}(x)) \\ &= \left(\frac{1}{v!}f^{(v)}(x)h^v k_v(k)\right)^2 + \left(\frac{f(x)R(k)}{nh}\right) \\ &= AMSE\left(\hat{f}(x)\right) \end{aligned} \tag{15}$$

A global measure of precision is the AMISE [13] [14]

$$\begin{aligned} AMISE &= \int_{-\infty}^{\infty} AMSE\,(\hat{f}(x))dx \\ &= \frac{k_v{}^2(k)}{v!}R\left(f^{(v)}\right)h^{2v} + \frac{R(k)}{nh} \end{aligned} \tag{16}$$

3.6. Optimal Bandwidth

The optimal bandwidth is the bandwidth that would minimize the mean integrated squared error. If the data were Gaussian and a Gaussian kernel was used, so it is not optimal in any global sense. In fact, for multimodal and highly skewed densities, this width is usually too wide and over smooth the density [12]. The optimal bandwidth depends on the unknown quantity $R\left(f^{(v)}\right)$. Silverman proposed to try the bandwidth computed by replacing $R\left(f^{(v)}\right)$ in the optimal formula by $R\left(g_{\hat{\sigma}}{}^{(v)}\right)$ where g_σ is a reference density a plausible candidate for f, and $\hat{\sigma}^2$ is the sample standard deviation.

The formula for the optimal bandwidth h is [12]

$$\begin{aligned} h &= \frac{0.9m}{n^{1/5}} \\ &; with\ \ m = min\left(\sqrt{Var(X)}, \frac{IQR}{1.349}\right) \end{aligned} \tag{17}$$

Where n is the number of observations on X, $var(X)$ is it is variance and $IQR(X)$ is interquartile range.

4. Result

The histogram is simple to construct and provides an impression of the density distribution of the data if an appropriate choice of classes is used. If the data are a random selection, the histogram is an estimate of the population density distribution. However, the visual impression gained from a histogram can depend to an unwelcome extent on the intervals selected for the classes (i.e., the number and midpoint of the bins). A reconstruction of the population density more consistent than the histogram would therefore be welcome.

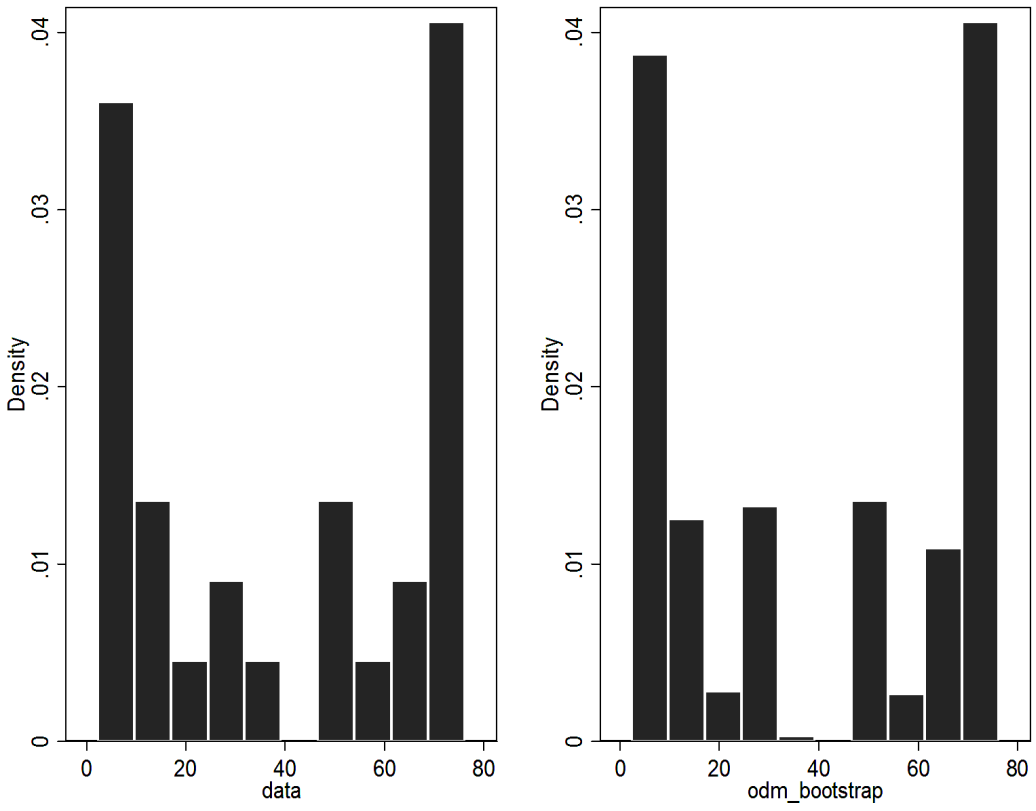

Figure 4. Histogram of data set before and after bootstrap

Based on the visual histogram (figure 4), which was used data set before and after bootstrap not really different in distribution. This visual histogram also shows that although the data point has been used, it is not deleted from the original data set in bootstrapping process. As a result, the same observation may be included in the resample data set.

The confidence interval for scalar parameters was calculated using percentile method to set approximate of bootstrap calculations. These refinements of the percentile method are the bias-corrected (BC) percentile method and the accelerated bias-corrected (BC_a) percentile method. The 250 resample data sets had been used when calculating a BCa confidence interval. As a result of not having to calculate bias correction, a smaller value, in the range of resample data can be used when using the percentile method for estimating a confidence interval. As the number of resample data sets decreases, more variability is introduced into the confidence interval estimation.

The 2.5% and 97.5% percentiles constitute the limits of the 95% confidence interval. The BCa method adjusts for bias in the bootstrapped sampling distributions relative to the actual sampling distribution, and is thus considered a substantial improvement using the percentile method. Based on this interval confidence (table 3), we known that week of 22 consistently produce the optimum oil content of palm mesocarp until week of 24 and 26. Those periods also give coverage probability ≈ 95% and small variance, even though the range of interval confidence is little bit large.

For visual plot in Figure 5, we knew that the increases of oil content production after pollination were started from week of 16 until 21. The week of 22 consistently produce the optimum oil content of palm mesocarp until week of 24 and 26.

Table 3. Bootstrap: Interval confidence, range, and coverage of probability

Week	Mean	Varian	LL	UL	Range	Coverage of probability
			0.025	0.975		
6	2.48	0.00	2.31	2.57	0.26	0.820
14	4.14	0.13	3.56	4.65	1.09	0.812
16	8.17	1.84	6.16	10.74	4.58	0.824
18	14.50	4.29	11.61	19.74	8.13	0.940
19	29.21	2.04	26.22	32.25	6.03	0.936
20	47.57	0.00	47.50	47.72	0.22	0.800
21	62.71	1.43	61.01	64.72	3.71	0.840
22	74.01	1.20	72.23	75.54	3.31	0.944
24	74.06	0.82	72.66	76.08	3.42	0.844
26	74.34	0.30	73.96	76.21	2.25	0.936

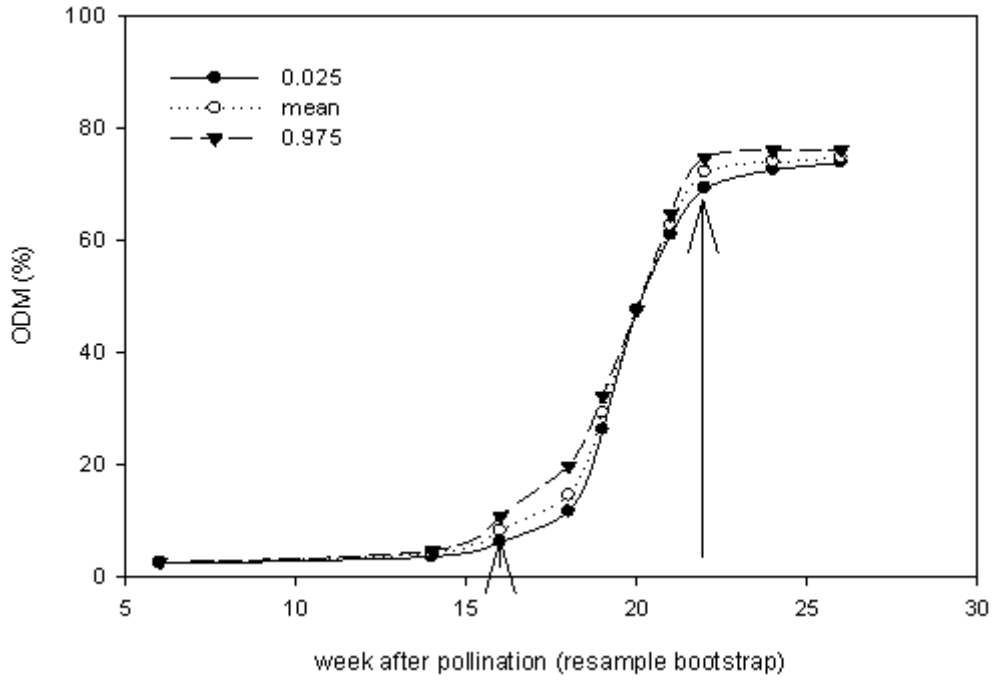

Figure 5. Oil contents of palm mesocarp in range of interval confidence (2,5% ; 97,5%)

Silverman [12] proposed to try the bandwidth computed by replacing $R\left(f^{(v)}\right)$ in the optimal formula by $R\left(g_{\hat{\sigma}}^{(v)}\right)$ where g_{σ} is a reference density a plausible candidate for f, and $\hat{\sigma}^2$ is the sample standard deviation. The Silverman formula had been done to find the optimum bandwidth of epanechnikov kernel. These bandwidths as a basic principal to build the optimum density of epanechnikov kernel distribution on resample data. As shown from the table 4 and figure 6, epanechnikov kernel has been successfully simulated the optimum estimate of oil content in palm mesocarp.

Plot estimation (figure 6) using epanechnikov kernel from resample bootstrap show that week of 22 produces the optimum oil content of palm mesocarp around 74% (ODM). This optimum consistently produces until week of 24 and 26. Based on this estimation, we can satisfy that epanechnikov kernel can be used to simulate the optimum estimate of oil content in palm mesocarp.

Table 4. Epanechnikov kernel bandwidth

Week	Bandwidth
6	0.0207
14	0.1192
16	0.4179
18	0.6593
19	0.4162
20	0.0115
21	0.3493
22	0.4637
24	0.2776
26	0.1681

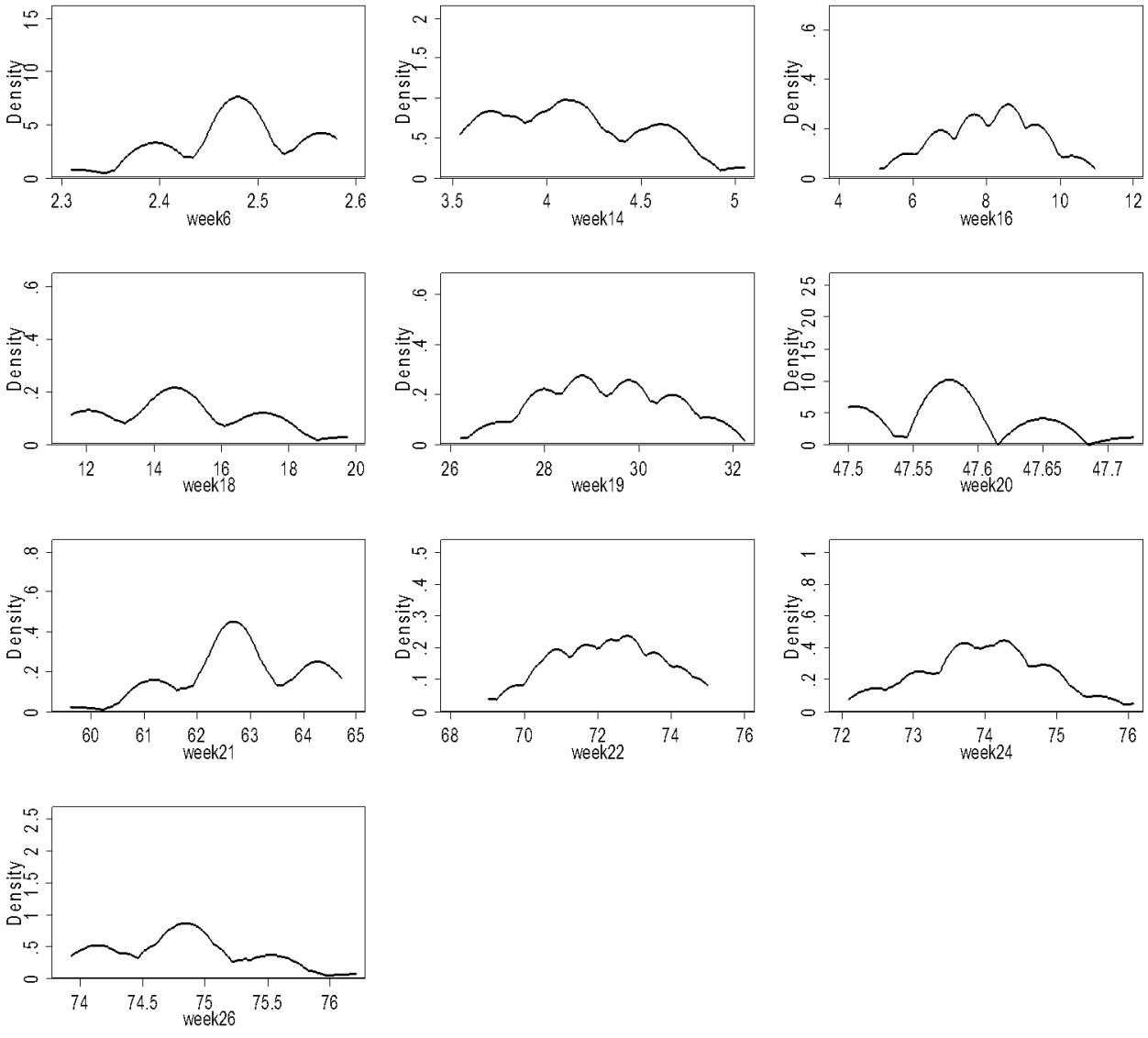

Figure 6. Plot estimation of oil contents using Epanechnikov kernel

5. Conclusions

In this paper, we have simulated the using of epanechnikov kernel as optimum kernel to estimate the oil content (%) of palm mesocarp at various periods. With respected to the method for finding optimal bandwidth. We propose Silverman [16] formula to determining the optimum bandwidth for epanechnikov kernel. The result show that epanechnikov kernel has been successfully simulated the optimum estimate of oil content (%) in oil palm mesocarp.

REFERENCES

[1] Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*, Chapman and Hall, New York. 1993.

[2] Davison, A. C. and Hinkley, D. Bootstrap *Methods and Their Application*, Cambridge University Press, Cambridge. 1997.

[3] Hall, P. *On the number of bootstrap simulations required to construct a confidence interval*, Ann.Stat.,14,4, 1453-1462. 1986.

[4] Shao, J. and Tu, D. The Jackknife and Bootstrap, Springer-Verlag, New York. 1995.

[5] Efron, B. *Bootstrap methods: another look at the jackknife*, Ann. Statist., 7, 1- 26. 1979.

[6] Efron, B. *Nonparametric standard errors and confidence intervals, with discussion*, Canad. J. Stat., 9, 2, 139 - 172. 1981.

[7] Efron, B. *The Jacknife, the Bootstrap and Other Resampling Plans*. vol 38, SIAM, Philadelphia. 1982.

[8] Efron, B. *Better bootstrap confidence intervals, with comments*, JASA, 82, 397, 171 - 200. 1987.

[9] Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*, Chapman and Hall, New York. 1994.

[10] M.P. Wand and M.C. Jones. *Kernel Smoothing*, Chapman & Hall, London. 1995.

[11] Rosenblatt, M. *Remarks on some nonparametric estimates of a density function*. Annals of Mathematical Statistics, 27, 832-837. 1956.

[12] Silverman, B.W. *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall. 1986.

[13] Wand, M.P. and Jones, M.C. *Kernel Smoothing*, London: Chapman and Hall. 1995.

[14] Wand, M.P., Marron, J.S. and Ruppert, D. *Transformations in Density Estimation (with discussion)*. Journal of the American Statistical Association, 86, 343-361. 1991.

[15] Zhang, S. and Karunamuni, R.J. *On Kernel Density Estimation Near Endpoints*. Journal of Statistical Planning and Inference, 70, 301-316. 1998.