Kernel Density Estimation of Dairy Cow Milk Production at BBPTUHPT Baturraden

Muhammad Ghifari Azzaki[1], Novita Eka Chandra[2], Indra Herdiana[3], Abdul Aziz[4], Ari Wardayani[5]
[1, 2, 3, 4, 5]Faculty of Mathematics and Natural Sciences, Universitas Jenderal Soedirman, Indonesia
Corresponding email: novita.chandra@unsoed.ac.id

Abstract

BBPTUHPT Baturraden is a dairy cattle breeding center in Central Java where daily milk production fluctuates and does not consistently follow a specific distribution pattern. Therefore, an appropriate statistical method is needed to estimate the underlying distribution of milk production data. This study aims to estimate the density function of daily dairy cow milk production in 2025 and evaluate the performance of different kernel functions and bandwidth selection methods. The Kernel Density Estimation (KDE) method was applied using Gaussian, Epanechnikov, and Cosine kernels. Optimal bandwidths were determined using Cross-Validation (CV) and Least Squares Cross-Validation (LSCV) methods. The analysis used 365 daily milk production observations. The results showed that all kernel and bandwidth combinations produced unimodal density curves, with peak densities occurring at 930–940 liters/day. Based on Mean Squared Error (MSE) evaluation, the Gaussian kernel with CV bandwidth achieved the lowest MSE value of 0.0000006495397, indicating the best estimation performance. The estimated density function showed that daily milk production was most frequently distributed within the range of 910–950 liters/day. Thus, the Gaussian kernel with CV bandwidth is considered the most suitable model for estimating the milk production distribution at BBPTUHPT Baturraden.

Keywords: Bandwidth selection; Kernel density estimation; Milk production

I. INTRODUCTION

Cattle are a strategic livestock commodity for meeting national demand for animal protein, and Indonesia continues to encourage dairy cattle development so domestic milk consumption can be supplied from local production. One expression of this effort is the network of government breeding centers for superior livestock spread across several provinces, including Central Java. The Balai Besar Pembibitan Ternak Unggul dan Hijauan Pakan Ternak (BBPTUHPT) Baturraden, located on the slopes of Mount Slamet in Banyumas Regency, is one of the region's best-known dairy breeding centers and a supplier of fresh milk to processing industries and community nutrition programs**.** However, its daily milk output is not constant. Production is dynamic and fluctuates from month to month because it is affected by biological and environmental factors such as genetics, lactation stage, availability of forage, and environmental stress (Stefani et al., 2021).

Feed quantity and milking interval are additional determinants of milk yield (Pasaribu et al., 2015). Unequal feed allocation between cows and differences in milking intervals cause the aggregate daily volume to fluctuate, so the resulting data may be asymmetric and display complex patterns. If such data are analyzed under a rigid parametric assumption, the conclusions may be biased and may not represent the actual population. A nonparametric statistical approach, which does not depend on normality or on any other theoretical distribution, therefore offers a more flexible way to describe the data structure. Nonparametric methods have proven successful in analyzing the probability structure of complex data, particularly in economics, agriculture, and agro-industry (Conover, 1999).

Nonparametric statistics is closely related to density estimation because it reconstructs the probability curve of a population in a data-driven manner, without forcing the

observations to follow a predetermined distribution. A parametric approach fixes the form of the distribution from the outset, for instance through a given mean and variance, and this assumption is often inconsistent with complex field data. In contrast, the nonparametric approach lets the empirical observations themselves define the shape of the density curve. This makes it an objective analytical instrument for highly fluctuating real data, because it can reveal skewness or multiple modes without the risk of model misspecification (Gramacki, 2018).

One approach to nonparametric density estimation is smoothing, which aims to produce a smooth and continuous estimate without assuming a particular population distribution (Cortés López & Jornet Sanz, 2020). Among smoothing techniques, the kernel method is the most widely used. The kernel estimator is mathematically tractable and flexible in shape (Härdle, 1991): it places a kernel function as a local weight on every single observation and then aggregates these weights into one continuous, adaptive, and smooth probability curve. The quality of a kernel density estimate is determined by two choices: the bandwidth and the kernel function (Santoso, 2010). The bandwidth is the smoothing parameter that controls how smooth the estimated curve is; an optimal bandwidth balances smoothness against fidelity to the true pattern. Data-driven selectors such as Least Squares Cross-Validation (LSCV), Generalized Cross-Validation (GCV), and Cross-Validation (CV) are commonly used to choose it. The kernel function also matters because it affects the shape of the density curve and hence its interpretation (Scott, 2015). Common kernels include the Gaussian, Epanechnikov, Uniform, Triangular, Biweight, Triweight, and Cosine kernels.

Several earlier studies motivate the present work. Chandra et al. (2016) carried out a simulation study of adjusted kernel density estimation that compared the Uniform, Epanechnikov, and Gaussian kernels on the basis of the MSE, using an optimal bandwidth derived theoretically by minimizing the MSE. Ramadhani et al. (2018) applied kernel density estimation to denim production data at PT Apac Inti Corpora with cross-validated bandwidths and compared several kernels, but did not evaluate the goodness of the estimates by the MSE. Harizahayu (2021) applied kernel density estimation with different bandwidths to egg production data and showed that larger bandwidths yield smoother curves that approach the normal shape, using the Gaussian, Epanechnikov, and Cosine kernels. To the best of our knowledge, kernel density estimation with data-driven optimal bandwidths has not yet been applied to daily dairy milk production at BBPTUHPT Baturraden, nor have the kernel and bandwidth choices been compared for such data with a quantitative error criterion.

The goodness of an estimate can be judged from its error: the smaller the error, the better the estimate (Gramacki, 2018). A standard measure is the Mean Squared Error (MSE), the sum of the squared bias and the variance of the estimator (Chen, 2017); a smaller MSE indicates that the estimated density represents the distribution of the data more accurately. Accordingly, this study has two objectives: (1) to determine the form of the kernel density estimate of daily milk production at BBPTUHPT Baturraden, and (2) to compare the kernel density estimates obtained from different kernel functions (Gaussian, Epanechnikov, and Cosine) and bandwidth selectors (CV and LSCV) using the MSE. The results are expected to serve as a reference for researchers applying kernel density estimation to livestock production data and as input for the center in monitoring and planning its milk production.

II. METHOD

1. *Data and research setting*

This quantitative case study used secondary data obtained from BBPTUHPT Baturraden, Banyumas Regency, Central Java. The center is a technical implementation unit of the Directorate General of Livestock and Animal Health, Ministry of Agriculture of the

Republic of Indonesia, and a national breeding center for purebred Friesian Holstein (FH) dairy cattle. The variable analyzed is the total daily milk production (liters/day) of the Tegalsari farm from 1 January to 31 December 2025, recorded by the technical staff of the center, which gives $n = 365$ observations. The data are aggregate records of all milked cows, not records of individual animals. The analysis using R through RStudio with the packages *stats*, *ks* (Duong, 2007).

2. *Kernel density estimator*

A function $f(x)$ is a probability density function if $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f\,(x)\,dx = 1$. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed observations from an unknown density f. The kernel density estimator of f with kernel K and bandwidth $h > 0$ is (Härdle, 1991)

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h\,(x - X_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where $K_h(x)$ is the kernel scaled by the bandwidth,

$$K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right). \tag{2}$$

A kernel function satisfies $K(x) \geq 0$, $\int K(x)\,dx = 1$, $\int xK(x)\,dx = 0$ (symmetry about zero), $\mu_2(K) = \int x^2K(x)\,dx < \infty$ and $\| K \|_2^2 = \int K^2(x)\,dx < \infty$. Nonnegativity and unit integral guarantee that $\hat{f}_h$ is itself a valid density; symmetry prevents systematic bias from unbalanced weighting on either side of the evaluation point; the second moment $\mu_2(K)$ governs the bias and the roughness $\| K \|_2^2$ governs the variance of the estimator (Węglarczyk, 2018). The three kernels compared in this study are the Gaussian kernel, which is infinitely differentiable and has unbounded support,

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}, \quad -\infty < u < \infty, \tag{3}$$

the Epanechnikov kernel, which minimizes the asymptotic mean integrated squared error among nonnegative kernels (Epanechnikov, 1969),

$$K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1), \tag{4}$$

and the Cosine kernel, a trigonometric kernel with the same bounded support (Węglarczyk, 2018),

$$K(u) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) I(|u| \leq 1). \tag{5}$$

For these kernels $\| K \|_2^2$ equals $1/\left(2\sqrt{\pi}\right) \approx 0.2821$, $3/5$ and $\pi^2/16 \approx 0.6169$, respectively. Because the Gaussian kernel never reaches zero, every observation contributes a nonzero weight at every evaluation point, whereas the Epanechnikov and Cosine kernels assign zero weight to observations whose normalized distance exceeds one, which makes them computationally cheaper (Chacón & Duong, 2018)

Assuming $f \in C^2$ and applying a second-order Taylor expansion of $f(x + hs)$ together with the kernel properties, the expectation and bias of the estimator are

$$E\left[\hat{f}_h(x)\right] = f(x) + \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2)$$

$$\text{Bias}\left[\hat{f}_h(x)\right] = \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2), \tag{6}$$

so the estimator is asymptotically unbiased as $h \to 0$ (Heidenreich et al., 2013). In the same way, the variance is

$$\mathrm{Var}\left[\hat{f}_h(x)\right] = \frac{\|K\|_2^2}{nh} f(x) + o\left(\frac{1}{nh}\right), \quad h \to 0,\ nh \to \infty, \tag{7}$$

which decreases as nh grows. Bias and variance therefore move in opposite directions as h changes. Combining them gives the asymptotic mean squared error

$$\mathrm{MSE}\left[\hat{f}_h(x)\right] = \frac{1}{nh} f(x) \parallel K \parallel_2^2 + \frac{h^4}{4}\left(f''(x)\mu_2(K)\right)^2 + o\left(\frac{1}{nh}\right) + o(h^4), \tag{8}$$

which converges to zero as $h \to 0$ and $nh \to \infty$; hence the kernel estimator is consistent. Setting the derivative of the leading terms of (8) with respect to h to zero yields the locally optimal bandwidth

$$h_{opt} = \left(\frac{f(x)\|K\|_2^2}{n\left(f''(x)\mu_2(K)\right)^2}\right)^{1/5}. \tag{9}$$

Equation (9) depends on the unknown f and f'' and shows that the optimal bandwidth shrinks at the rate $n^{-1/5}$, so in practice it must be estimated from the data.

3. *Bandwidth selection*

A bandwidth that is too large causes oversmoothing (high bias), so that important structure such as additional modes is flattened, whereas a bandwidth that is too small causes undersmoothing (high variance) and produces spurious peaks that merely follow individual observations (Heidenreich et al., 2013). Two data-driven selectors were used. Cross-Validation (CV) is a leave-one-out criterion that guards the estimate against overfitting (Roberts et al., 2017):

$$CV(h) = \int_{-\infty}^{\infty} \hat{f}_h^2(x)\, dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(X_i), \tag{10}$$

where $\hat{f}_{h,-i}$ is the estimate computed without the i-th observation,

$$\hat{f}_{h,-i}(X_i) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{X_i - X_j}{h}\right), \tag{11}$$

and $h_{CV} = \mathrm{argmin}_{h>0} CV(h)$. Least Squares Cross-Validation (LSCV) targets the bandwidth with the lowest integrated squared error without requiring knowledge of the true density (Bowman, 1984; Heidenreich et al., 2013) . Its objective has the same form as (10); writing the first term with the convolution $K^* = K * K$, $K^*(t) = \int K(\tau)K(t-\tau)\, d\tau$ (Węglarczyk, 2018), gives

$$LSCV(h) = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n} K^*\left(\frac{X_i - X_j}{h}\right) - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i} K\left(\frac{X_i - X_j}{h}\right), \tag{12}$$

and the minimizer of (12) minimizes the average deviation over the whole support of the density (Wu, 1997). In the computation, h_{CV} was obtained with the function *bw.ucv* of R and h_{LSCV} with the function *hlscv* of the *ks* package (Duong, 2007), the latter applied to the data with a negligible jitter (factor 0.01, seed 123) to break tied values. Both bandwidths are expressed on the scale used by the R function *density*, in which h is the standard deviation of the scaled kernel; consequently, one bandwidth value per selector can be used with all three kernels.

4. *Goodness of estimation*

For an estimator $\hat{\theta}$ of a parameter θ, $\mathrm{MSE}(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2(\hat{\theta})$ (Bain & Engelhardt, 1992). Following this decomposition and equations (7)–(8), the MSE of each kernel–bandwidth combination was computed at every observation and averaged:

$$\widehat{\mathrm{MSE}} = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\hat{f}_{h,-i}(X_i) - \phi(X_i; \bar{X}, s)\right)^2 + \frac{\|K\|_2^2}{nh}\hat{f}_{h,-i}(X_i)\right], \tag{13}$$

where the bias is measured against the reference density $\phi(X_i; \bar{X}, s)$, the normal density with the sample mean and standard deviation, which is justified when the normality test is not rejected, and the variance term uses the leave-one-out estimate in place of the unknown f. A smaller value indicates a more accurate estimate.

5. *Research procedure*

The analysis followed these steps: (1) collect the daily milk production data of BBPTUHPT Baturraden for 2025; (2) compute descriptive statistics (number of observations, mean, median, standard deviation, variance, minimum, maximum, skewness, and kurtosis) and visualize the data; (3) select the optimal bandwidth by CV and LSCV; (4) estimate the density using the three kernels with each bandwidth, giving six combinations; (5) compare the six estimates by the MSE; and (6) draw conclusions.

III. RESULTS AND DISCUSSIONS

A. Result

1. *Description of daily milk production*

Table 1 summarizes the 365 daily records of the Tegalsari farm. Mean production was 913.4392 liters/day and the median 916.5 liters/day; the closeness of the two is an early indication that the distribution does not depart far from symmetry. The standard deviation of 60.1621 liters/day (variance 3,619.486) shows considerable day-to-day variation around the mean. The minimum of 736 liters/day (9 February 2025) and the maximum of 1,078.7 liters/day (20 July 2025) give a range of 342.7 liters/day, which may reflect lactation stage, animal health, and feeding management. The skewness of 0.1334 indicates a nearly symmetric distribution with a slightly longer right tail, and the kurtosis of 2.7950, below the normal value of 3, indicates slightly lighter tails than a normal distribution.

Table 1. Descriptive statistics of daily milk production at BBPTUHPT Baturraden, 2025

Statistic	Value
Number of observations (n)	365
Mean	913.4392 liters/day
Median	916.5 liters/day
Standard deviation	60.1621 liters/day
Variance	3,619.486
Minimum	736.0 liters/day
Maximum	1,078.7 liters/day
Skewness	0.1334
Kurtosis	2.7950

Figure 1 shows that daily production fluctuated strongly throughout the year. During roughly the first 50 days, production lay mostly between 850 and 1,000 liters/day before dropping to the annual minimum on day 40. Production then recovered gradually with large daily fluctuations, reaching a first local peak of about 1,050 liters/day around days 110–120, fell back to about 830 liters/day around day 150, and rose again to the annual maximum of about 1,080 liters/day around days 190–210. These swings are consistent with the biology of dairy cows: yield rises sharply after calving, peaks around weeks 4–8 of lactation (day 35–50), and then declines by about 2.5% per week until the dry period, with an ideal lactation length of 305 days (Dyah, 2023). Because the data are herd totals rather than records of a single cow, the rises and falls are most likely the combined result of many cows at different lactation stages, together with environmental factors such as season, forage quality and quantity, and milking management. Heat stress lowers milk yield in tropical regions through genotype-by-

environment interaction (Santana et al., 2016); at the cooler highland location of Baturraden, this effect is expected to be less dominant, although alternation of rainy and dry seasons can still affect forage supply.

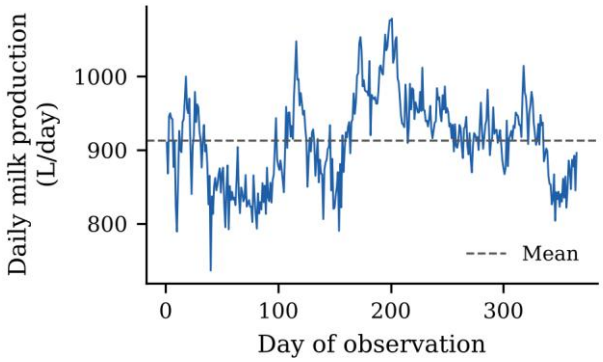

Figure 1. Daily milk production of the Tegalsari farm, BBPTUHPT Baturraden, 2025

The histogram with the empirical density curve in Figure 2 shows a main peak at about 920–940 liters/day and a shoulder at about 840–870 liters/day. The shoulder gives an early indication of possible heterogeneity of sub-groups in the population, which is common in livestock data drawn from animals at different lactation stages, ages or parities. A histogram can only show such structure coarsely, because its shape depends heavily on the choice of bins and it is discontinuous. Kernel density estimation, in contrast, produces a continuous and smooth curve without assuming a distributional form, and is therefore better able to capture local features such as shoulders or potential multimodality (Chen, 2017).

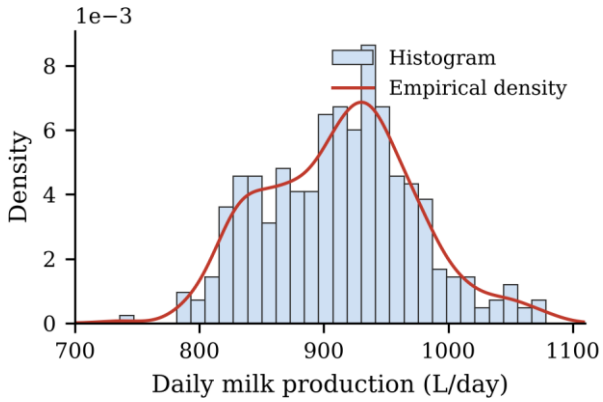

Figure 2. Histogram and empirical density of daily milk production.

2. *Kernel functions*

Figure 3 displays the three kernels in standardized form. All are unimodal and symmetric about zero, in accordance with the kernel properties. They differ in the rate at which the weight decays: the Gaussian kernel decays exponentially and never reaches exactly zero, while the Epanechnikov and Cosine kernels decrease to exactly zero at $|u| = 1$. As a numerical illustration, the Gaussian kernel gives $K(0) = 0.3989$ and $K(1) = 0.2420$; the Epanechnikov kernel gives $K(0) = 0.75, K(0.5) = 0.5625$ and $K(1) = 0$; and the Cosine kernel gives $K(0) = 0.7854, K(0.5) = 0.5554$, and $K(1) = 0$. The Cosine kernel follows a cosine curve, so its weight declines more gently near the boundary than the parabolic Epanechnikov kernel, giving a smoother transition at the edge of the support (Chacón & Duong, 2018). Applying three kernels with different decay characteristics also serves as a

sensitivity check on the choice of kernel, whose influence is theoretically much smaller than that of the bandwidth (Gramacki, 2018)

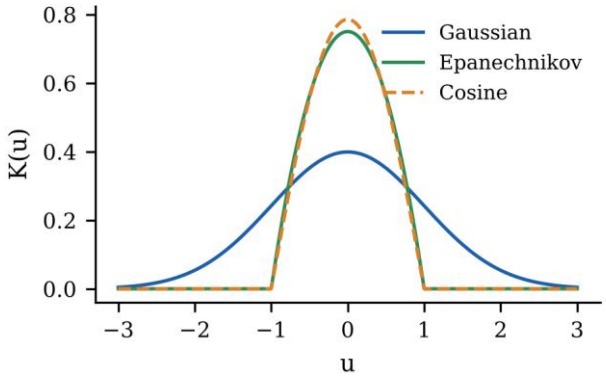

Figure 3. Gaussian, Epanechnikov and Cosine kernel functions.

3. *Optimal bandwidth*

The optimal bandwidths from the two selectors are listed in Table 2. The CV bandwidth (16.3578) is slightly larger than the LSCV bandwidth (16.32272). The difference, about 0.035 or less than 0.35% of the CV value, is very small, which is expected because both selectors minimize an unbiased estimate of the integrated squared error and differ only in numerical implementation. Relative to the scale of the data (range 342.7 liters/day, standard deviation 60.1621 liters/day), the ratio $h/s \approx 16.3578/60.1621 \approx 0.272$ shows that the bandwidth is about 27% of the standard deviation, a moderate degree of smoothing that is neither large enough to erase important features nor small enough to make the curve noisy. With $n = 365$ observations and the $n^{-1/5}$ rate implied by (9), the selected bandwidths can be expected to be reasonably close to their asymptotically optimal values.

Table 2. Optimal bandwidths selected by CV and LSCV

Kernel function	h_{CV}	h_{LSCV}
Gaussian	16.3578	16.32272
Epanechnikov	16.3578	16.32272
Cosine	16.3578	16.32272

4. *Kernel density estimates*

Using equation (1) with each kernel and each bandwidth produced the six estimates in Figure 4. All curves are unimodal with the peak at about 930–940 liters/day, consistent with the main peak of 920–940 liters/day in the exploratory histogram; this agreement indicates that the kernel estimates represent the pattern of the data well while giving a smoother and continuous picture. The shoulder seen in the histogram is still visible but appears smoother, because the moderate bandwidth tends to smooth out small-scale local detail. The differences between the kernels and between the two bandwidths are hardly visible. The Gaussian kernel, which is infinitely differentiable, is theoretically expected to produce the smoothest curve, whereas the bounded-support Epanechnikov and Cosine kernels may give slightly less smooth curves. The choice between CV and LSCV therefore does not change the visual interpretation of the distribution, although numerical differences remain and are assessed with the MSE below.

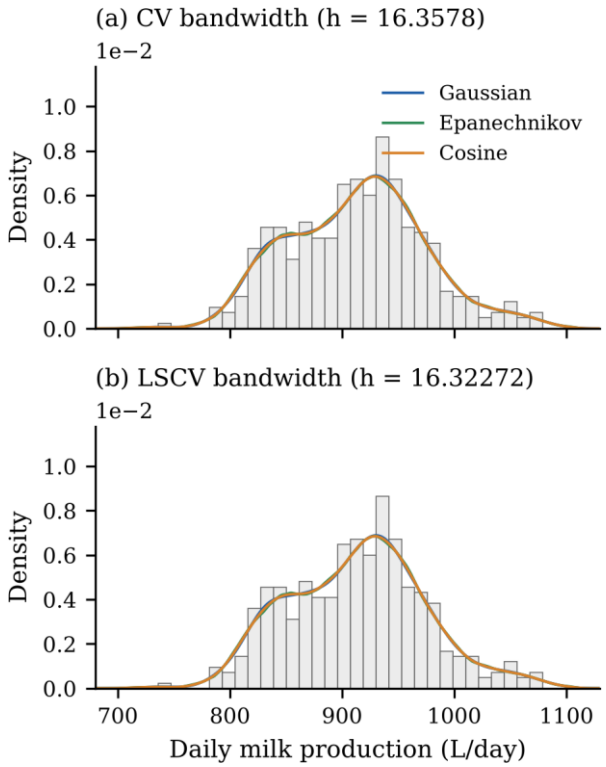

Figure 4. Kernel density estimates of the three kernels with (a) the CV bandwidth and (b) the LSCV bandwidth.

5. *Comparison of goodness of estimation*

To illustrate the bias–variance decomposition (8), consider the Gaussian kernel with the CV bandwidth at $x = 910$, where $f(910) \approx 0.006057$ and $f''(910) \approx -1.197 \times 10^{-6}$. The variance term is $0.2821 \times 0.006057/(365 \times 16.3578) \approx 2.861 \times 10^{-7}$ and the squared bias is $16.3578^4 \times (1.197 \times 10^{-6})^2/4 \approx 2.567 \times 10^{-8}$, giving an MSE of about 3.118×10^{-7} at that point. At the selected bandwidth, the variance component therefore dominates the squared bias. Averaging the pointwise MSE over all 365 observations according to (14) gives the overall values in Table 3.

Table 3. MSE of the six kernel–bandwidth combinations

Rank	Kernel function	Bandwidth selector	MSE
1	**Gaussian**	**CV**	6.495397×10^{-7}
2	Gaussian	LSCV	6.512276×10^{-7}
3	Epanechnikov	CV	8.671339×10^{-7}
4	Epanechnikov	LSCV	8.694693×10^{-7}
5	Cosine	LSCV	8.912323×10^{-7}
6	Cosine	CV	8.917977×10^{-7}

Notes: bold indicates the best combination

The Gaussian kernel with the CV bandwidth gives the smallest MSE, 6.495397×10^{-7}, and is therefore the most accurate of the six combinations, whereas the Cosine kernel with the CV bandwidth gives the largest, 8.917977×10^{-7}. All six values are of the same order of magnitude (about 6.5×10^{-7} to 8.9×10^{-7}). Regarding the bandwidth selector, CV gives a slightly smaller MSE than LSCV for the Gaussian and Epanechnikov kernels, while LSCV gives a slightly smaller MSE for the Cosine kernel; in every case the difference

is below 0.3%, so the two selectors are practically equivalent. Regarding the kernel, the ordering is the same under both selectors: Gaussian has the smallest MSE, followed by Epanechnikov and then Cosine.

6. *Best kernel–bandwidth combination*

Based on Table 3, the Gaussian kernel with the CV bandwidth was selected as the best combination. Substituting (3) into (1), the best estimate of the density of daily milk production is

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-X_i}{h}\right)^2\right], \quad n = 365,\ h = 16.3578. \tag{15}$$

For example, at $x = 726$ the estimate is $\hat{f}_h(726) = \frac{1}{365(16.3578)\sqrt{2\pi}}\sum_{i=1}^{365}\exp\left[-\frac{1}{2}\left(\frac{726-X_i}{16.3578}\right)^2\right] \approx 5.553 \times 10^{-5}$. Evaluating (15) over a fine grid gives the curve in Figure 5. The estimated density is highest over the range 910–950 liters/day, with the mode at about 930 liters/day; about 29% of the observed days fall in this range and about 75% fall between 850 and 1,000 liters/day.

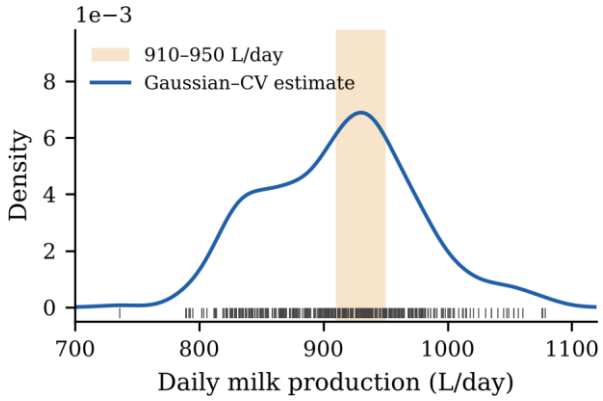

Figure 5. Best density estimate

.

B. Discussion

The first objective concerned the form of the kernel density estimate. All six estimates are unimodal with a peak at about 930–940 liters/day and a shoulder at 840–870 liters/day. The unimodal shape and the non-rejection of normality show that, at the level of the whole year, daily output is concentrated around a single typical value. The shoulder, however, is a local feature that a normal model would ignore. A plausible explanation, consistent with the time series in Figure 1, is that the herd passed through periods of lower aggregate production (for example, in February and around days 140–160 and 330–365) when a larger share of cows were in late lactation or when feed conditions changed, which adds a cluster of days with production in the 830–870 liters/day range. This illustrates the value of the kernel approach noted by Gramacki (2018) and Chen (2017): even when a global normality test is passed, a nonparametric estimate can reveal structure that a parametric fit would hide.

The second objective concerned the comparison of kernels and bandwidth selectors. The Gaussian kernel gave the smallest MSE under both selectors, and the kernel ordering (Gaussian, Epanechnikov, Cosine) was consistent. Nevertheless, the absolute differences are small, which agrees with the theory of kernel efficiency: the relative efficiencies of common kernels such as the Gaussian, Epanechnikov and Cosine are all very high, so that the shape of the kernel has only a marginal effect once the bandwidth is chosen well (Chen, 2017). Two further remarks help to interpret the ranking. First, the MSE criterion (14) measures bias

against a normal reference density; since the data were found to be consistent with normality, this reference is reasonable, but it naturally favors the smooth Gaussian kernel, so the advantage of the Gaussian kernel should be read in the context of this criterion. Second, the Epanechnikov kernel is optimal with respect to the asymptotic mean integrated squared error over all densities (Epanechnikov, 1969), so its lower ranking here does not contradict theory but reflects a finite-sample, data-specific comparison.

The reason the Gaussian kernel ranks first can be traced partly to the shape of the data. The sample skewness (0.1334) and kurtosis (2.7950) are close to those of a normal curve. Under such conditions, a kernel with smooth, unbounded tails blends neighbouring observations without introducing small kinks, which is precisely what a criterion built on squared deviations rewards. The bounded kernels behave differently: under the bandwidth convention of the R function density, the Epanechnikov kernel gives zero weight to any day lying more than about $\sqrt{5}h \approx 36.6$ liters/day from the evaluation point, so in sparsely populated regions, especially below 800 and above 1,050 liters/day, the estimate rests on only a handful of days. In addition, the larger roughness $\|K\|_2^2$ of the Epanechnikov and Cosine kernels reported in the Method section enters the variance term of (13) directly, so at a common bandwidth their variance component is larger. The computational saving of compact support Chacón & Duong (2018) matters little here, because computing time for 365 observations is negligible.

The difference in MSE between the two selectors (below 0.3%) is much smaller than the difference between kernels (about 33–37% between the Gaussian and the other kernels). This should be interpreted carefully. A distinction must be made between the bandwidth selection method (CV or LSCV) and the bandwidth value it produces. Because both selectors minimize the same integrated squared error criterion, they produced nearly identical values (16.3578 and 16.32272), so the choice of method could hardly affect the results. This does not contradict the general rule in the density estimation literature that the bandwidth value is far more important than the kernel shape (Heidenreich et al., 2013): Figure 5 shows that changing the bandwidth from 5 to 25 alters the estimate dramatically, whereas changing the kernel at a fixed bandwidth barely does. The practical implication is that future applications of kernel density estimation to similar data should concentrate effort on choosing an appropriate bandwidth rather than on the kernel shape.

The pointwise calculation at $x = 910$ liters/day also shows what limits the accuracy of the estimates. At that point the variance term (about 2.86×10^{-7}) is roughly eleven times the squared bias (about 2.57×10^{-8}), so more than ninety percent of the local error stems from sampling variability rather than from systematic smoothing bias. Two consequences follow. First, enlarging the sample, for instance by pooling several years of records, would lower the error more effectively than any further refinement of the kernel. Second, a slightly wider bandwidth would exchange a little additional bias for a lower variance, which hints that cross-validation may have undersmoothed the data to a modest degree. This behaviour agrees with the review of Heidenreich et al. (2013), who note that cross-validation selectors vary considerably between samples and frequently choose bandwidths that are too small; comparing the present results with plug-in or rule-of-thumb selectors would therefore be worthwhile. Daily totals are reported to one decimal place, so tied values occur, and the jitter added before computing the LSCV bandwidth alters the data very slightly; since the resulting bandwidth differs from the CV value by less than 0.35%, this adjustment does not affect the conclusions.

The findings complement earlier studies. Like Harizahayu (2021), this study shows that larger bandwidths produce smoother curves closer to the normal shape, and that the Gaussian, Epanechnikov, and Cosine kernels behave similarly. Unlike Ramadhani et al. (2018), the kernel–bandwidth combinations were compared with a quantitative criterion, and unlike the simulation setting of Chandra et al. (2016), the comparison was carried out on real

livestock production data with data-driven bandwidths. The result that differences between kernels are small while the bandwidth controls the shape is in line with these studies and with the reviews of Heidenreich et al. (2013) and Węglarczyk (2018).

The best estimate has practical implications for BBPTUHPT Baturraden. First, the density curve can serve as a reference for the normal range of daily production, around the peak of 910–950 liters/day, so that days with production far outside this range can be flagged as potential anomalies requiring attention to animal health, feed quality or milking management. Second, information on the shape of the production distribution can support more data-driven production planning; assuming that internal and external production conditions remain similar, daily production in the following year can be expected to fall mostly in this range. Several limitations should be noted. The daily records form a time series and are likely autocorrelated because of the continuity of lactation, whereas the Lilliefors test and the standard CV and LSCV procedures assume independent observations; cross-validation schemes designed for temporally structured data are recommended in such cases (Roberts et al., 2017). The bandwidth is global, that is, a single value is applied over the whole support, which was chosen for simplicity and consistency with the standard selectors. Finally, the data cover only one farm and one year. Further studies could therefore consider other kernels (Uniform, Triangular, Biweight, Triweight), other selectors such as Generalized Cross-Validation, multi-year or multi-center data, and multivariate kernel estimation that includes lactation stage, animal age or daily weather.

From a dairy management perspective, the yield of an individual cow depends on its genetic merit and calving history (Stefani et al., 2021), on the quantity and quality of feed and the milking routine (Pasaribu et al., 2015), and on climatic stress (Santana et al., 2016). Once the yields of many such cows are added into one daily total, these influences overlap, and the resulting distribution need not belong to any familiar parametric family. The kernel estimate accommodates this mixture without forcing it into a single shape, and the shoulder at 840–870 liters/day signals that the herd total can occasionally settle at a lower level for several consecutive weeks. Because about three quarters of the days fall between 850 and 1,000 liters/day, values outside this band may serve farm staff as a simple, data-based warning threshold. At the same time, the density only describes how often each production level occurred during 2025; it says nothing about when those levels occurred. Linking the kernel estimate with time series methods would make the analysis more useful for short-term forecasting.

IV. CONCLUSION

Kernel density estimation of the 365 daily milk production records of BBPTUHPT Baturraden in 2025 with the Gaussian, Epanechnikov, and Cosine kernels and with CV (h = 16.3578) and LSCV (h = 16.32272) bandwidths produced unimodal curves peaking at about 930–940 liters/day with a shoulder at 840–870 liters/day. The best estimate, with the smallest MSE of 6.495397×10^{-7},is the Gaussian kernel estimator with the CV bandwidth, which shows that daily production most frequently lies between 910 and 950 liters/day. The Gaussian kernel gave the smallest MSE under both selectors, followed by the Epanechnikov and Cosine kernels; CV gave a slightly smaller MSE than LSCV for the Gaussian and Epanechnikov kernels and a slightly larger one for the Cosine kernel, but all differences were small, indicating that the bandwidth value rather than the kernel shape or the selector governs the estimate.

ACKNOWLEDGEMENT

The authors thank the LPPM Universitas Jenderal Soedirman for supporting this research through the Unsoed Basic Research scheme, BLU Unsoed Research Grant in 2026, with contract number 13.404/UN23.34/PT.01.00/IV/2026, and the Balai Besar Pembibitan Ternak Unggul dan Hijauan Pakan Ternak (BBPTUHPT) Baturraden for providing the daily milk production data.

REFERENCES

Bain, L. J., & Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics* (2nd ed.). Duxbury Press Belmont, CA.

Bowman, A. W. (1984). An Alternative Method of Cross-Validation for The Smoothing of Density Estimates. *Biometrika*, *71*(2), 353–360. https://doi.org/10.1093/biomet/71.2.353

Chacón, J. E., & Duong, T. (2018). *Multivariate Kernel Smoothing and Its Applications*. CRC Press. https://doi.org/10.1201/9780429485572

Chandra, N. E., Mahyuddin, M., & Rohmaniah, S. A. (2016). Estimasi Densitas Kernel Adjusted: Studi Simulasi. *Jurnal Ilmiah Matematika Dan Pendidikan Matematika*, *8*(2), 33–40. https://doi.org/10.20884/1.jmp.2016.8.2.2885

Chen, Y.-C. (2017). A Tutorial on Kernel Density Estimation and Recent Advances. *Biostatistics & Epidemiology*, *1*(1), 161–187. https://doi.org/10.1080/24709360.2017.1396742

Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd ed.). John Wiley & Sons.

Cortés López, J. C., & Jornet Sanz, M. (2020). Improving Kernel Methods for Density Estimation in Random Differential Equations Problems. *Mathematical and Computational Applications*, *25*(2), 33–41. https://doi.org/10.3390/mca25020033

Duong, T. (2007). ks: Kernel Density Estimation and Kernel Discriminant Analysis for Multivariate Data in R. *Journal of Statistical Software*, *21*(7), 1–16. https://doi.org/10.18637/jss.v021.i07

Dyah, R. (2023). *Laktasi BBPTUHPT Baturraden*. https://bbptusapiperah.ditjenpkh.pertanian.go.id/page/laktasi

Epanechnikov, V. A. (1969). Nonparametric Estimation of A Multivariate Probability Density. *Theory of Probability & Its Applications*, *14*(1), 153–158. https://doi.org/10.1137/1114019

Gramacki, A. (2018). *Nonparametric Kernel Density Estimation and Its Computational Aspects*. Springer. https://doi.org/10.1007/978-3-319-71688-6

Härdle, W. (1991). *Smoothing Techniques: with Implementation in S*. Springer Verlag. https://doi.org/10.1007/978-1-4612-4432-5

Harizahayu. (2021). Estimasi Density Kernel dengan Bandwidth yang Berbeda. *MAp (Mathematics and Applications) Journal*, *3*(1), 35–40. https://doi.org/10.15548/map.v3i1.2636

Heidenreich, N.-B., Schindler, A., & Sperlich, S. (2013). Bandwidth Selection for Kernel Density Estimation: A Review of Fully Automatic Selectors. *Asta Advances in Statistical Analysis*, *97*(4), 403–433. https://doi.org/10.1007/s10182-013-0216-y

Pasaribu, A., Firmansyah, & Idris, N. (2015). Analisis Faktor-Faktor yang Mempengaruhi Produksi Susu Sapi Perah di Kabupaten Karo Provinsi Sumatera Utara. *Jurnal Ilmiah Ilmu-Ilmu Peternakan*, *18*(1), 28–35. https://doi.org/https://doi.org/10.22437/jiiip.v18i1.2656

Ramadhani, P., Ispriyanti, D., & Safitri, D. (2018). Kapabilitas Proses dengan Estimasi Fungsi Densitas Kernel pada Produksi Denim di PT Apac Inti Corpora. *Jurnal Gaussian*, *7*(3), 326–336. https://doi.org/10.14710/j.gauss.7.3.326-336

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Santana, M. L. J., Bignardi, A. B., Pereira, R. J., Menéndez-Buxadera, A., & El Faro, L. (2016). Random Regression Models to Account for the Effect of Genotype by Environment Interaction due to Heat Stress on the Milk Yield of Holstein Cows Under Tropical Conditions. *Journal of Applied Genetics*, *57*(1), 119–127. https://doi.org/10.1007/s13353-015-0301-x

Santoso, S. (2010). *Statistik Nonparametrik*. Elex Media Komputindo.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9780470316849

Stefani, G., Brandão Aquaroli, D., Batista Gonçalves Costa Júnior, J., Santana Júnior, M. L., Tonhati, H., Sesana, R. C., & El Faro, L. (2021). Genetic Parameters for Dystocia, Milk Yield, and Age at First Calving in Brazilian Holstein Cows. *Journal of Applied Animal Research*, *49*(1), 1–5. https://doi.org/10.1080/09712119.2020.1856115

Węglarczyk, S. (2018). Kernel Density Estimation and Its Application. *ITM Web of Conferences*, *23*, 00037. https://doi.org/10.1051/itmconf/20182300037

Wu, C. O. (1997). A Cross-Validation Bandwidth Choice for Kernel Density Estimates with Selection Biased Data. *Journal of Multivariate Analysis*, *61*(1), 38–60. https://doi.org/10.1006/jmva.1997.1659