Gaussian Process Regression with Average Hyperparameter

S Winarni\textsuperscript{1} and S W Indratno\textsuperscript{2}

\textsuperscript{1,2}Statistics Research Group, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Jalan Ganesha 10 Bandung, 40132, Indonesia

Email: sriwinarni@students.itb.ac.id

Abstract. Gaussian regression process (GPR) is a non-parametric analysis approach that is more flexible than simple linear regression analysis. Linear and non-linear patterned data can be modeled by GPR. In big data the use of GPR is constrained by the increasing size of the data. Data partitioning approach can be done as an alternative to facilitate data processing. The partitioned data then shifts one data, and so on until the last training data. In each data partition an optimum hyperparameter will be obtained. The average hyperparameter is used in the prediction process. The results obtained are that by partitioning the data and then using the average hyperparameter, the predicted results obtained are better than doing with the overall training data. In the data partition approach, small partitions provide better predictive results than large partitions. The resulting RMSE value is also getting smaller. Thus the average hyperparameter approach with small data partitions can be used as an alternative in making predictions on GPR.

1. Introduction

A Gaussian process is a stochastic process which generates samples over time \( \{X_t\}_{t \in T} \) such that no matter which finite linear combination of \( X_t \) ones takes, that linear combination will be normally distributed \([1]\). In the parametric regression analysis a regression model is formed that is used to predict the target value and the prediction results obtained in the form of a certain value without the distribution of the predicted results are known. Parametric regression can be used on data that has a linear pattern. When the data encountered is not linear and parametric regression cannot be used, Gaussian process regression can be used as an alternative for modeling. Gaussian process regression can not only determine the predicted value but can also be used to determine the distribution of the predicted value \([2]\).

The discussion on Gaussian process regression begins with a multivariate Gaussian distribution. The multivariate Gaussian distribution has a dimensionless probability density function \([3]\) given in Equation (1).

\[
P(x|\mu, \Sigma) = N(x|\mu, \Sigma) = (2\pi)^{-D/2} |\Sigma|^{-1/2} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \tag{1}
\]

Parameter \( \mu \) is an average vector with \( \mu \in \mathbb{R}^N \), and \( |\Sigma| \) is a determinant of a positive definite symmetric covariance matrix. So it can be written that \( x|\mu, \Sigma \sim \mathcal{N}(x|\mu, \Sigma) \). Vector random variable \( X = [X_1 \ \cdots \ X_n]^T \) distributed multivariate gaussian with average vectors \( \mu \) and covariance matrix \( \Sigma \). Dimension \( D \) in Equation 1 is the number of \( X \) variables found in the opportunity density function.
Covariance matrix is symmetric positive definite. Symmetric means the covariance matrix entry above the diagonal will be the same as the entry below the diagonal, or \( \Sigma = \Sigma^T \). Positive definite means that if a symmetric matrix \( \Sigma \) satisfies the square form \( z^T \Sigma z > 0 \). In the case of bivariate gaussian \( X_1 \) and \( X_2 \) which are not mutually free, the joint density function \( X_1 \) and \( X_2 \) is as follows in Equation (2).

\[
    f(x_1, x_2) = \frac{1}{2\pi \sqrt{\sigma_{11} \sigma_{22} (1 - \rho_{12}^2)}} \exp \left\{ -\frac{1}{2(1-\rho_{12}^2)} \left[ \left( \frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right)^2 - 2 \rho_{12} \frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} + \left( \frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right)^2 \right] \right\}
\]

In the case of \( X_1 \) and \( X_2 \) which independent, the joint density function \( X_1 \) and \( X_2 \) will be shaped as Equation (3).

\[
    f(x_1, x_2) = \frac{1}{2\pi \sqrt{\sigma_{11} \sigma_{22}}} \exp \left\{ -\frac{1}{2} \left( \frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right)^2 - \frac{1}{2} \left( \frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right)^2 \right\}
\]

Marginal density function \( X_2 \) can be obtained as Equation (4).

\[
    f(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 = (2\pi \sigma_{22})^{-1/2} \exp \left\{ -\frac{1}{2} \left( \frac{x_2 - \mu_2}{\sigma_{22}} \right)^2 \right\}
\]

\( X_2 \) Gaussian distribution with an average \( \mu_2 \) and variance \( \sigma_{22} \), can be written \( X_2 \sim \mathcal{N}(\mu_2, \sigma_{22}) \). Density function \( X_1 \) given \( X_2 \) can be obtained as Equation (5) [4].

\[
    f(x_1|x_2) = (2\pi \sigma_{11})(1 - \rho_{12}^2))^{-1/2} \exp \left\{ -\frac{1}{2} \left( \frac{(x_1 - a)^2}{(1-\rho_{12}^2)\sigma_{11}} \right) \right\}
\]

\( X_1 \) given \( X_2 \) Gaussian distribution with an average \( a = \mu_1 + \rho_{12} \frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}} (x_2 - \mu_2) \) and variance \( (1 - \rho_{12}^2)\sigma_{11} \). Can be written as \( X_1|X_2 \sim \mathcal{N}(\mu_1 + \rho_{12} \frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}} (x_2 - \mu_2), (1 - \rho_{12}^2)\sigma_{11}) \).

A multivariate gaussian distribution is used to model a number of up to variables. The Gaussian Process is a development of a multivariate Gaussian with an infinite number of variables. Gaussian Process is not just a random vector distribution but is a distribution of random functions. Gaussian Process is also a stochastic process with subset of random variables distributed with Gaussian multivariate.

Gaussian processes regression in the field of finance can be used to predict stock prices. The prediction is done using the optimum hyperparameter obtained from the marginal log likelihood function. In large data applications, determining hyperparameter involves large matrix inverse. This will be a problem in conducting data analysis. This research will try to estimate the optimum hyperparameter by partitioning the data window with a smaller size. This is intended to make it easier to do the analysis rather than using overall data. Each partition will get the optimum hyperparameter that matches the data partition. Prediction can be done by using the average value of the optimum hyperparameter. The purpose of this study is to compare the predicted results using an average hyperparameter partition data with an overall hyperparameter data.

2. Literature Review
Regression have the term generalized linear regression is known, which forms the average linear regression using the set of m basis functions \( \{ \Phi_i(x) \} \). The regression function will be in the form of \( y(x) = \sum_{i=1}^{m} w_i \Phi_i(x) \) for the "weight" vector \( w \). If there is a prior gaussian distribution in weights and assumed gaussian noise then there are two equivalent ways to get the regression function and interpret the Gaussian Process (GP) regression model. First is weight-space view, and second is function-space view. In function-space view, GP is defined as the distribution of functions [5].

2.1. Weight-Space View
Linear regression model has advantages in implementation and interpretation. The weakness of this model is the limitations in flexibility. If the relationship between input and output cannot be approximated by a linear function, the model will provide an incorrect predictive value. In Bayesian analysis, a linear regression model with gaussian noise is given in Equation (6) [1]
with \( \mathbf{x} \) is input vector, \( \mathbf{w} \) is weight vector (parameter) in linear model, \( f \) is function value dan \( y \) is target value. Observed value \( y \) is different from function value \( f(\mathbf{x}) \) with noise increment \( \epsilon \). Noise is assumed independent, Gaussian spreads are identical with zero mean and variance \( \sigma_n^2 \), then \( \epsilon \sim \mathcal{N}(0, \sigma_n^2) \).

This noise assumption gives a likelihood, the probability density of the given parameter observation is given in Equation (7) [6].

\[
p(y|\mathbf{x}, \mathbf{w}) = \prod_{i=1}^{n} p(y_i|x_i, \mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_n}} \exp \left( -\frac{(y_i-x_i^T \mathbf{w})^2}{2\sigma_n^2} \right) = \frac{1}{(2\pi \sigma_n^2)^{n/2}} \exp \left( -\frac{1}{2\sigma_n^2} |\mathbf{y} - \mathbf{X}^T \mathbf{w}|^2 \right) = \mathcal{N}(\mathbf{X}^T \mathbf{w}, \sigma_n^2 \mathbf{I})
\]  

with \(|z|\) is the euclidean length of the \( z \) vector. In Bayesian it is necessary to specify prior to parameters, trust in parameters before looking at observations. For weights, prior gaussian weighting with zero average and covariance matrix is taken \( \Sigma_p \). It is given in Equation (8)[7].

\[
\mathbf{w} \sim \mathcal{N}(0, \Sigma_p)
\]  

Conclusions in the Bayesian linear model are based on the posterior distribution of weights. Calculated by Bayes law given in Equation (9)[8].

\[
\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}
\]

\[
p(w|y, \mathbf{X}) = \frac{p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w})}{p(y|\mathbf{x})} = \frac{p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w})}{\int p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w})d\mathbf{w}} = \frac{\mathcal{N}(\mathbf{X}^T \mathbf{w}, \sigma_n^2 \mathbf{I}) \mathcal{N}(0, \Sigma_p)}{\int \mathcal{N}(\mathbf{X}^T \mathbf{w}, \sigma_n^2 \mathbf{I}) \mathcal{N}(0, \Sigma_p)d\mathbf{w}}
\]

\[
\propto \mathcal{N}(\mathbf{X}^T \mathbf{w}, \sigma_n^2 \mathbf{I}) \mathcal{N}(0, \Sigma_p)
\]

\[
\propto \exp \left[ -\frac{1}{2\sigma_n^2} (\mathbf{y} - \mathbf{X}^T \mathbf{w})^T (\mathbf{y} - \mathbf{X}^T \mathbf{w}) \right] \exp \left( -\frac{1}{2} \mathbf{w}^T \Sigma_p^{-1} \mathbf{w} \right)
\]

\[
\propto \exp \left[ -\frac{1}{2} (\mathbf{w} - \overline{w})^T \left( \frac{1}{\sigma_n^2} \mathbf{X}^T \Sigma_p^{-1} \mathbf{X} \right) (\mathbf{w} - \overline{w}) \right]
\]

\[
p(w|y, \mathbf{X}) \sim \mathcal{N}(\overline{w}, \mathbf{A}^{-1})
\]  

With \( \overline{w} = \sigma_n^{-2} \mathbf{A}^{-1} \mathbf{X} \mathbf{y} \) and \( \mathbf{A}^{-1} = (\sigma_n^{-2} \mathbf{X}^T \Sigma_p^{-1} \mathbf{X} + \Sigma_p^{-1})^{-1} \).

Equation (6), posterior \( p(w|y, \mathbf{X}) = \mathcal{N}(\overline{w}, \mathbf{A}^{-1}) \) is used to predict \( f \) for test point \( \mathbf{x}_o \). \( f_o = f(\mathbf{x}_o) = \mathbf{x}_o^T \mathbf{w} \). Then the predictions for \( f_o \) in \( \mathbf{x}_o \) are obtained by averaging all the outputs of a possible linear model, called posterior Gaussian given in Equation (10).

\[
p(f_o|\mathbf{x}_o, \mathbf{X}, \mathbf{y}) = \int p(f_o|\mathbf{x}_o, \mathbf{w}) p(w|\mathbf{X}, \mathbf{y}) dw = \mathcal{N} \left( \frac{1}{\sigma_n^2} \mathbf{x}_o^T \mathbf{A}^{-1} \mathbf{X} \mathbf{y}, \mathbf{x}_o^T \mathbf{A}^{-1} \mathbf{x}_o \right)
\]  

2.2. Function-Space View

In the function-space view approach, Gaussian process is used to define the distribution of functions. Gaussian process is a collection of a number of up to random variables that have a joint distribution of
gaussian. The Gaussian process is completely determined by the average function \( m(x) \), covariance function \( k(x, x') \) and real process \( f(x) \). Gaussian process can be written in Equation (11) [9].

\[
f(x) \sim \mathcal{GP}(m(x), k(x, x'))
\]

with \( m(x) = \mathbb{E}[f(x)] \) and \( k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \). In this case the random variable is the value of the function \( f(x) \) in time \( x \). Often the gaussian process is defined from time to time, with a collection of random variable indices being time. Gaussian process is defined as a collection of random variables. This definition requires consistency or known as the nature of marginalization. For example determined \( (y_1, y_2) \sim \mathcal{N}(\mu, \Sigma) \), then it must be determined \( y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}) \) with \( \Sigma_{11} \) is a submatrix that is relevant to \( \Sigma \). Consistency requirements are automatically met if the covariance function determines the covariance matrix entry.[10]

Covariance matrix elements are built by covariance functions. The covariance function is also called the kernel. One kernel that is often used is the exponential quadratic function given in Equation (12) [12].

\[
k(x_p, x_q) = \sigma_f^2 \exp \left( -\frac{1}{2\ell^2} |x_p - x_q|^2 \right)
\]

Parameters of the kernel function, i.e. \( \sigma_f^2 \) and \( \ell \) referred to as hyperparameter. \( \sigma_f^2 \) referred to as signal variance and \( \ell \) referred to as a length scale. The value can be determined freely. If \( \sigma_f^2 = 1 \) and \( \ell = 1 \) then the covariance matrix with the quadratic exponential kernel function will be shaped \( k(x_p, x_q) = \exp \left( -\frac{1}{2} |x_p - x_q|^2 \right) \). Covariance matrix are obtained in Equation (13) [11].

\[
k(x_p, x_q) = \begin{bmatrix}
k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\
k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\
\vdots & \vdots & \ddots & \vdots \\
k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n)
\end{bmatrix}
\]

Covariance matrices are symmetric, then \( k(x_2, x_1) = k(x_1, x_2) \), etc. Other covariance functions that can be used to construct covariance matrices include linear covariance functions, periodic covariance functions, wiener process covariance functions, and matern covariance functions. The covariance function and its hyperparameter will determine the shape of the Gaussian process distribution. So it is necessary to predict the optimum hyperparameter. Determination the optimum hyperparameter by maximizing the marginal log likelihood function given in Equation (14) [15].

\[
\log p(y|X) = -\frac{1}{2} y^T [K(X, X) + \sigma_n^2 I]^{-1} y - \frac{1}{2} \log |K(X, X) + \sigma_n^2 I| - \frac{n}{2} \log (2\pi)
\]

The marginal likelihood function is seen as an objective function that is influenced by hyperparameter \( \theta \) namely \( f(\theta) = -\log p(y) \), thus the optimization problem is

\[
\min_{\theta} f(\theta) \text{ to } \theta \geq 0
\]

The distribution of prior Gaussian process regression in time series data with 10 series and 20 samples, is given in Figure 1.
Figure 1. Prior distribution Gaussian process with $m(\mathbf{x}) = 0$ and $k(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}|x_p - x_q|^2\right)$. $\sigma_f^2 = 1$, $\ell = 1$.

The prior distribution is on average equal to zero.

In Gaussian process regression, the prediction model can be carried out with two approaches, namely prediction using observation data without noise and prediction approach using observation data with noise. In predictions using noiseless observations $\mathbf{X}$, values are taken outside the $\mathbf{X}$ value then predictions can be made. The shared distribution of the training data output $f$, and the output of the testing data $f^*$ according to prior is given in Equation (15) [13].

$$
\begin{bmatrix} f \\ f^* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) & K(\mathbf{X}, \mathbf{X}^*) \\ K(\mathbf{X}^*, \mathbf{X}) & K(\mathbf{X}^*, \mathbf{X}^*) \end{bmatrix}\right)
$$

(15)

If there are $n$ training data points and $n^*$ testing data points then $K(\mathbf{X}, \mathbf{X})$ is the $n \times n$, covariance matrix for all pairs of training and testing points. Likewise for $K(\mathbf{X}, X^*)$ for couples training with training, $K(X^*, X)$ for couples testing with testing, and $K(X^*, X^*)$ for couples testing with training. From Equation 10 it is found that the distribution function $(f^* | \mathbf{X}, \mathbf{X}, f)$ is given in Equation (16).

$$
(f^* | \mathbf{X}, \mathbf{X}, f) \sim \mathcal{N}\left(K(\mathbf{X}, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}f, K(\mathbf{X}, \mathbf{X}^*) - K(\mathbf{X}, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}K(\mathbf{X}, \mathbf{X}^*)\right)
$$

(16)

In the case of real data, prediction models with noise $\mathbf{y} = f(\mathbf{x}) + \mathbf{e}$ are often used. assumption of additive independent identically distributed Gaussian noise with variance $\sigma_n^2$, then $\text{cov}(\mathbf{y}) = \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}$. The distribution along with the target value and function values on the test points is given in Equation (17) [14].

$$
\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I} & \mathbf{K}(\mathbf{X}, \mathbf{X}^*) \\ \mathbf{K}(\mathbf{X}^*, \mathbf{X}) & \mathbf{K}(\mathbf{X}^*, \mathbf{X}^*) \end{bmatrix}\right)
$$

(17)

By deriving the conditional distribution we get the following prediction Equation (18).

$$
f^* | \mathbf{X}, \mathbf{y}, \mathbf{X}, \sim \mathcal{N}(\mathbf{f}_*, \text{cov}(\mathbf{f}_*))
$$

(18)

With $\mathbf{f}_* = \mathbf{K}(\mathbf{X}, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}]^{-1}\mathbf{y}$, and

$\text{cov}(\mathbf{f}_*) = \mathbf{K}(\mathbf{X}, \mathbf{X}^*) - \mathbf{K}(\mathbf{X}, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}]^{-1}\mathbf{K}(\mathbf{X}, \mathbf{X})$. 


Equation (18) produces a predictive value $\hat{f}$, with covariance $\text{cov}(\hat{f})$. The accuracy of the predicted results can be seen from the value of the root mean square error (RMSE) given in the Equation (19).

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (\hat{y}_t - y_t)^2}$$  \hspace{1cm} (19)

With $\hat{y}_t$ is a predictive results on point $X_t$, in this case it is $\hat{f}$, value of $y_t$ is a actual data on point $X_t$.

3. Method
In this study the data used is the monthly data of Bank Mandiri stock prices in January 2015-October 2019 source from yahoo finance. Data for January 2015 – September 2019 is used as training data and data from October 2019 is used as testing data. The stages of analysis conducted in this study are:

- In the training data, a window data is partitioned with 5, 10 and 20 data.
- In the first 5 data partition we estimate the optimum hyperparameter using the maximum marginal log likelihood function given in Equation 14, with kernel function is exponential quadratic.
- After obtaining the optimum hyperparameter, then the window is shifted one data.
- In the new data window, the optimum hyperparameter estimation is performed again.
- After obtaining the optimum set of hyperparameters, the average of the hyperparameters is taken.
- Predict the testing data using Equation 18. In partitioned training data, the covariance matrix $K(X, X)$ uses the average covariance matrix with each optimum hyperparameter of each partition. The matrix $K(X_*, X)$ is the covariance matrix between $X_*$, value of data to be predicted that is the 58th data with value of $X$ on the last data partition with average optimum hyperparameter. The Y value used is the Y average.
- the same process is carried out on data partitions 10 and 20.
- in the overall training data, estimation of testing data is done without data partitioning. The covariance matrix obtained is the covariance matrix obtained from the optimum hyperparameter overall data.
- Comparing the prediction results obtained with partition data and overall data. In addition, we also want to know which partitions can make the best predictions, that is, which results in the smallest RMSE given in equation 19.
- Draw conclusions related to the use of partition data to make predictions. If it turns out better then predictions can be made using only partition data.

4. Result and Discussion
Estimation of Bank Mandiri stock prices in October 2019 using training data for January 2015 - September 2019 stock prices. The training data patterns can be illustrated in Figure 2.

![Figure 2. Training data plot. There is an increasing trend pattern.](image-url)
In figure 1 it can be seen that the training data forms fluctuating patterns and is not linear. Modeling can be done with Gaussian process regression.

Modeling is done by forming data partitions of 5, 10, and 20 data. In partition 5, the first estimation of the optimum hyperparameter value is using the maximum marginal log likelihood, a hyperparameter length scale value of \( \ell = 1.8 \) is obtained. After obtaining the optimum hyperparameter, the data window shifts one data, the data to 1-5 shifts to data to 2-6; the optimum hyperparameter estimation is again carried out, and so on until the final data. The optimum hyperparameter obtained is then averaged. The optimum length scale hyperparameter values obtained from 53 data partitions 5 can be given in Figure 3.

![Figure 3](image)

**Figure 3.** Average of optimum length scale for data partition 5.

The value of the length scale is relatively large.

Figure 2 shows that the optimum length scale value on data partition 5 fluctuates and an average length scale of 2.2 is obtained (red line). The average hyperparameter for partition 5 data is worth \( \ell = 2.2 \) and \( \sigma_f^2 = 5080 \). The optimum hyperparameter estimation process is also performed on data partitions 10 and 20. In data partitions 10 and 20, the optimum length scale hyperparameter is given in Figure 4.

![Figure 4](image)

**Figure 4.** Average of optimum length scale for (a) partition data 10; (b) partition data 20.

The value of the length scale is relatively small.

For partition 10, the optimum hyperparameter results are obtained \( \ell = 1.5 \) and \( \sigma_f^2 = 4565 \) and for 20 data partitions \( \ell = 1.4 \) and \( \sigma_f^2 = 4391 \). These results indicate that the greater data partition, the smaller the hyperparameter. Likewise, hyperparameter obtained from the whole training data gives a hyperparameter value smaller than partitioned data. Hyperparameter \( \ell \) is the length parameter, the larger
bushes show a longer pattern. This means that hyperparameter $\ell$ will form a longer pattern as the data partition becomes smaller.

The result of optimum hyperparameter for overall training data is worth $\ell = 1.03$ and $\sigma^2 = 2037$. if the optimum hyperparameter is fixed to the training data, we get a prediction with the noise given in Figure 5

![Figure 5](image)

**Figure 5.** Prediction with the optimum hyperparameter for overall training data. The gray area is a 5% confidence interval. The prediction results coincide with the data.

Figure 5 it can be seen that the optimum hyperparameter forms a pattern that matches the data. Prediction on training data using optimum hyperparameter can be used to estimate the point that is in the training data. Gray areas are predictive confidence intervals. in this study predictions outside the training data were carried out to infer stock prices in October 2019.

Stock price predictions in October 2019 are done using equation 18. Equation 19 is used to estimate the RMSE used to find out the goodness of the predicted results. Prediction results for partition data and overall training data are given in Table 1.

| Parameter       | Prediction with data partition | Prediction with overall training data | Real data |
|-----------------|-------------------------------|--------------------------------------|-----------|
| Mu              | 5358                          | 4936                                 | 5107      | 4400      | 6400      |
| Confidence interval | 5358 ± 47                      | 4936 ± 949                           | 5260 ± 62 | 4400 ± 19 |
| RMSE            | 1042                          | 1464                                 | 1293      | 2000      |

In the October 2019 predicted results of Table 1, prediction results with partition data provide predictive values that are closer to real data. The RMSE value on the partition data using the average hyperparameter is smaller than using the overall training data. This means that forming a data partition and using an average hyperparameter will produce better predictive values. In the data partition, the smallest RMSE value is obtained from data partition 5. This shows that the prediction results obtained from partition 5 are closer to real data than other partitions. the prediction results are a bit far from the real data, this could be due to the improper selection of the covariance function. the next research will be conducted an experiment using another covariance function.

**5. Conclusion**

Gaussian process regression with the hyperparameter average approach to data partitioning can improve the prediction results. The RMSE value obtained is smaller than using the whole training data. In the
prediction data partition approach the results will be better by using a small data partition. An RMSE value with a small data partition will result in a smaller MSE than a larger data partition. Gaussian process regression with the average hyperparameter approach to data partitioning can be used as an alternative to predicting with better results.

6. Acknowledgment
The authors thank to the ITB’s Program of Research, Community Service and Innovation (P3MI ITB) 2019 for supporting funds.

7. References
[1] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. 2006.
[2] C. K. I. Williams, “Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond,” *Learn. Graph. Model.*, pp. 599–621, 1998.
[3] R. V. Hogg, J. W. McKean, and A. T. Craig, *Introduction to Mathematical Statistics*, Eighth ed. Pearson, 2019.
[4] R. A. JOHNSON and D. W. WICHERN, *Applied Multivariate Statistical Analysis*, Sixth edit. Prentice Hall, 2007.
[5] C. K. I. Williams and C. E. Rasmussen, “Gaussian Processes for Regression.”
[6] M. Längkvist, L. Karlsson, and A. Loutfi, “A review of unsupervised feature learning and deep learning for time-series modeling,” *Pattern Recognit. Lett.*, vol. 42, no. 1, pp. 11–24, 2014.
[7] R. Pandit and D. Infield, “Performance assessment of a wind turbine using SCADA based Gaussian Process model,” *Int. J. Progn. Heal. Manag.*, vol. 9, no. 1, pp. 1–8, 2018.
[8] J. Gonzalvez, E. Lezmi, T. Roncalli, and J. Xu, “Financial Applications of Gaussian Processes and Bayesian Optimization,” *SSRN Electron. J.*, 2019.
[9] C. E. Rasmussen, “Gaussian Processes in Machine Learning.”
[10] D. M. Bui, H. Q. Nguyen, Y. I. Yoon, S. I. Jun, M. B. Amin, and S. Lee, “Gaussian process for predicting CPU utilization and its application to energy efficiency,” *Appl. Intell.*, vol. 43, no. 4, pp. 874–891, 2015.
[11] A. Ben Abdessalem, N. Dervilis, D. J. Wagg, and K. Worden, “Automatic kernel selection for gaussian processes regression with approximate bayesian computation and sequential Monte Carlo,” *Front. Built Environ.*, vol. 3, no. August 2017, pp. 1–13, 2017.
[12] M. Blum and M. Riedmiller, “Optimization of gaussian process hyperparameters using Rprop,” *ESANN 2013 proceedings, 21st Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn.*, no. April, pp. 339–344, 2013.
[13] D. J. Leith, M. Heidl, and J. V. Ringwood, “Gaussian process prior models for electrical load forecasting,” *2004 Int. Conf. Probabilistic Methods Appl. to Power Syst.*, pp. 112–117, 2004.
[14] A. Y. Sun, D. Wang, and X. Xu, “Monthly streamflow forecasting using Gaussian process regression,” *J. Hydrol.*, vol. 511, pp. 72–81, 2014.
[15] A. Melkumyan and F. Ramos, “Multi-kernel Gaussian processes,” *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 1408–1413, 2011.