Correlation Analysis of Multivariate Regression Algorithms on RSS Data for Indoor Positioning System

Dodo ZainalAbidin[1,*], SitiNurmaini[2,*], Erwin [3], ErrissyaRasywir[4], YoviPratama[5]

[1] Doctoral Student of Departement of Engineering, UniversitasSriwijaya; dodozenalabidin@gmail.com

[2] Intelligent System Research Group, Computer ScienceFaculty, UniversitasSriwijaya; sitnurmaini@gmail.com

[3] Intelligent System Research Group, Computer ScienceFaculty, UniversitasSriwijaya; arwin@unsri.ac.id

[4] Departement of Informatics Engineering, UniversitasDinamikaBangsa; errissyarasywir@gmail.com

[5] Departement of Informatics Engineering, UniversitasDinamikaBangsa; yovipratama@gmail.com

***** Correspondence: dodozenalabidin@gmail.com; siinurmaini@gmail.com;

Abstract: In this study, we used the Received Signal Strength (RSS) dataset that we were collecting at the University ofDinamikaBangsa Jambi Building. RSS dataset is data that can be utilized in the development of signal processing technology that is very useful in various fields. Our RSS dataset has dependent and independent variables. However, we need to know whether our dataset is feasible or not to be tested with Machine Learning. Therefore, testing is needed to determine the correlation value between the dependent and independent variables in our dataset. Some of the algorithms we use in testing the correlation values of this dataset are Partial least square (PLS), Canonical Correlation Analysis (CCA), and Partial Least Square Canonical (PLSC). From the test results obtained that the correlation of multiple variables in the dataset with the highest value of r2 score is PLS regression with a value of N = 3, R2_score of 0.630453 and MSE of 44.89262. The R2_score value obtained by PLS exceeds the target value of the correlation indicator with a good value of 0.6.

Keywords: Correlation; Variable; RSS; PLS; CCA; PLSC;

1. Introduction

Simply stated, correlation can be interpreted as a relationship. However, when developed further, the correlation can not only be understood as limited to this understanding(Supriyadi, Mariani, & Sugiman, 2017). Correlation is one of the statistical analysis techniques used to find the relationship between two variables that are quantitative(Z. Zhang, Yuan, Shen, & Li, 2018). The relationship between these two variables can occur because of a causal relationship or it can also occur by chance alone. Two variables are said to correlate if changes in one variable will be followed by changes in the other variables regularly in the same direction (positive correlation) or opposite (negative correlation)(Memenuhi, Mata, Teknik, & Data, 2017). In mathematics, correlation is a measure of how closely two variables change in relation to one another (Lu, Wang, Bansal, Gimpel, & Livescu, 2015). In this study we used the Received Signal Strength (RSS) dataset that we were collecting at the University ofDinamikaBangsa Jambi Building. RSS dataset is data that can be utilized in the development of signal processing technology that is very useful in various fields. Our RSS dataset has dependent and independent variables.

However, we need to know whether our dataset is feasible or not to be tested with Machine Learning(Akhlaghi, Zhou, & Huang, 2018; Felix, Siller, & Alvarez, 2016; Liu et al., 2017) . Therefore, testing is needed to determine the correlation value between the dependent and independent variables in our dataset. Based on this, we chose several well-known algorithms that are best suited for measuring the correlation of our dataset variables. Our Algortima that we chose is a number of algorithms from the Multivariate Regression family(Supriyadi et al., 2017).

The PLS method is considered as a powerful analysis method because it is not based on many assumptions or conditions, such as normality and multicollinearity tests. The method has its own advantages, among others: data does not have to be multivariate normally distributed. Even indicators with a scale of data categories, ordinal, intervals to ratios can be used (Schmidt-Hieber, 2017a). Another advantage is that the sample size does not have to be large.

Furthermore, the canonical correlation analysis algorithm as a multiple variable statistical technique (Multivariate) investigates the close relationship between two variable groups(Lu et al., 2015). The meaning of the cluster here is group. One group of variables is identified as a group of independent variables, while the other variable groups are treated as a dependent variable group(Vu, Koo, & Choi, 2017). And through dependency between the two groups of variables can be explained the effect of one variable group on another variable group(Abidin et al., 2020). Canonical correlation analysis is one of the statistical analysis techniques used to see the relationship between a set of independent variables with a set of dependent variables (Lu et al., 2015).

In studies that have made comparisons between PLS and PLS Canonical (PLSC), it is found that the relationship between characteristics and parameters with canonical correlation analysis and partial regression analysis is least squares. The PLSC method is able to extract predictive information for latent variables more effectively than the usual PLS approach. PLSC is a simple modification of the PLS and PPLS algorithm(Shao, Wang, Zhao, su, & Cai, 2016). The purpose of testing and experimentation with this algorithm is to look for

correlations between variables that influence and variables that are affected in RSS data(Abidin et al., 2020; Sánchez-Rodríguez, Quintana-Suárez, Alonso-González, Ley-Bosch, & Sánchez-Medina, 2020). With the known correlation value between the dependent variable and the independent variable from the results of this test, the dataset we tested can find out whether our dataset is feasible or not to be subjected to task machine learning(Kumar, Kumar, & Nirmal, 2021). The algorithm used in this experiment is a family of algorithms that calculate variable correlations or what is called Cross Decomposition(Hadera et al., 2019). In our study, we used 9 independent variables and 3 dependent variables.

2.Reviewof Related Studies

2.1. Partial Least Square (PLS)

Partial least square (PLS) is a multivariate statistical technique that can handle being able to handle influencing variables and affected variables at once.

Figure 1. Partial least square (PLS) classification and function (Schmidt-Hieber, 2017a).

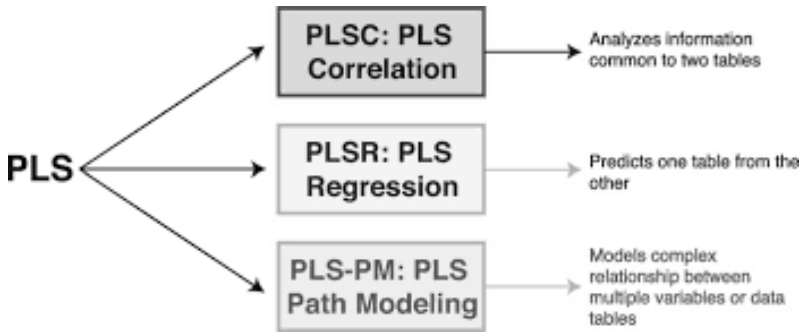


Figure 1 is a classification of PLS algorithms and their functions. PLS is a multiple regression analysis that can actually be used when there are many influential variables. Partial least square (PLS) is a type of statistical analysis whose use is similar to SEM (Structural Equation Modeling) in covariance analysis. SEM and PLS are part of a linear regression based method. So what is in linear regression, also exists in PLS. It just has a different term (Supriyadi et al., 2017).

The important advantage of Partial Least Square is that it can handle many independent variables, even if there is multicollinearity among the independent variables. Multiple regression analysis can actually still be used when there are many predictor variables (Zou, Jin, Jiang, Xie, & Spanos, 2017). However, if the number of variables is too large (for example, there are more variables than the number of observations) (S. Zhang, Choromanska, & Lecun, 2015), a model that is compatible with the sample data will be obtained, but will fail to predict for new data. This phenomenon is called overfitting. In such cases of overfitting, although there are many manifest factors, there may be only a few latent factors that can best explain variations in response (Schmidt-Hieber, 2017b). Then came the PLS idea. The general idea of PLS is to extract these latent factors, which explain as much as possible the variation of manifest factors when modeling response variables (Z. Zhang et al., 2018).

The PLS algorithm can be defined as the following analogy, for example X is a matrix with size n x p and Y is a matrix size n x q. Then the PLS procedure will extract factors from X and Y in such a row that between the extracted factors have maximum covariance (Akhlaghi et al., 2018). Partial Least Square technique will try to find a linear decomposition of X and Y. One assumption of multiple linear regression analysis is that there is no multicollinearity problem(Anagnostopoulos& Kalousis, 2019; Vu et al., 2017; Z. Zhang et al., 2018). If a multicollinearity problem occurs, the Partial Least Square (PLS) and Principal Component Regression (PCR) methods are two methods that can be used to overcome the multicollinearity problem (Y. Zhang, Zhou, Jin, Wang, & Cichocki, 2014). The method used in this study is the Partial Least Square (PLS) and Principal Component Regression (PCR) with the 2013 Central Java Regional Budget revenue data. The results of this study obtained a regression equation model with the Partial Least Square (PLS) method, namely and the equation model regression with the Principal Component Regression (PCR) method (Abdi, Guillemot, Eslami, & Beaton, 2017; Vu et al., 2017; Xie, Wang, Nallanathan, & Wang, 2016). Then the best method is chosen by using the highest value criteria and the smallest MSE. The selection of the best method is to look at the highest value and the smallest MSE.

Figure 2. Coordinate graph depicting the correlation between dependent and independent variables on PLS (Y. Zhang et al., 2014).

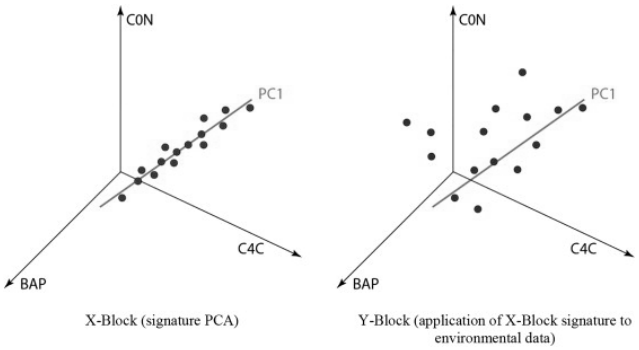


Figure 2 is a graphic coordinate drawing the correlation between the dependent and independent variables on PLS. The graph above is a visualization of the distribution of variable X data as an independent variable and Y variable as the dependent variable in the 3D coordinate plane (Supriyadi et al., 2017).

Figure 3. Linear plot of data distribution on the results of PLS processing.

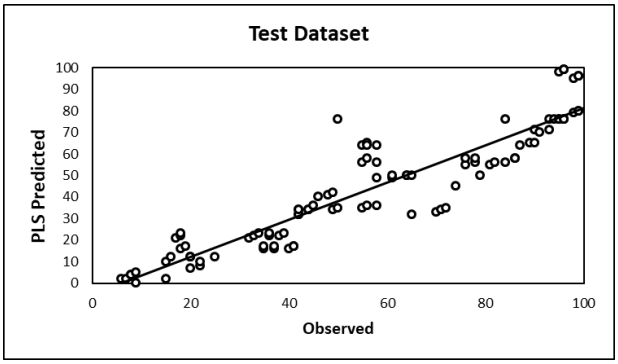


Figure 3 is a linear plot graph of the distribution of data on the results of processing with the Partial least square (PLS) algorithm.

2.2 Canonical Correlation Analysis (CCA)

Canonical correlation analysis (canonical analysis) is introduced as a multiple variable statistical technique (Multivariate) that investigates the closeness of the relationship between two variable groups (Z. Zhang et al., 2018).The meaning of the cluster here is group. One group of variables is identified as a group of independent variables, while the other variable groups are treated as a dependent variable group (Chen, Liu, Zhao, & Principe, 2017; Lu et al., 2015).

Through dependency between the two groups of variables can be explained the effect of one variable group on another variable group. Canonical correlation analysis is one of the statistical analysis techniques used to see the relationship between a set of independent variables with a set of dependent variables (Abdi et al., 2017). This analysis can measure the level of closeness of the relationship between a set of dependent variables with a set of independent variables. In addition, canonical correlation analysis is also able to describe the structure of relationships within a collection of independent variables (Lu et al., 2015).

Canonical correlation analysis focuses on the correlation between linear combinations of the dependent variable set $y' = y1, y2, ..,yp$ with the linear combination of the set of independent variables $x' = x, x2, ..., xq$. The idea of this analysis is to determine the pair of linear combinations that have the greatest correlation. Then look for pairs of linear combinations between pairs that are not correlated in the initial pair selected. This pair of linear combinations is called canonical function, and the correlation is called canonical correlation. In general, canonical correlations in the CCA algorithm can be described as follows in Figure 4 and Figure 5:

Figure 4. Visualization Chart Correlation between Independent and Dependent Variables.

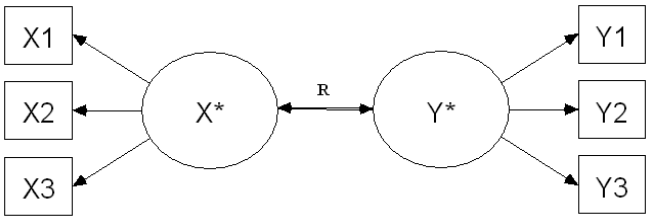


Figure 4 is a visualization of the correlation between the independent variable and the dependent variable. Line R is the value of the variable correlation generated by the variable correlation processing algorithm.

Figure 5. A chart that Describes the Name Of A Relation Contained in Variable Correlation

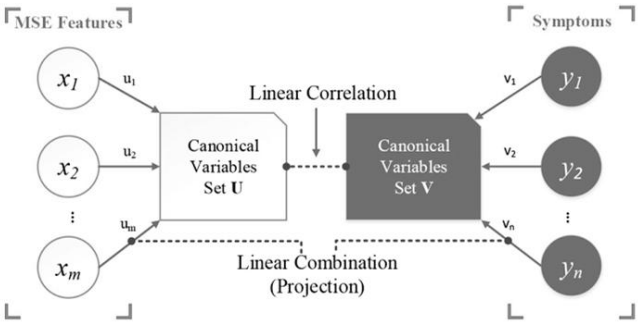
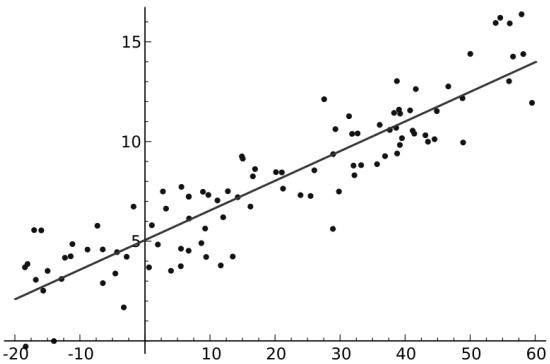


Figure 5 is a correlation visualization of the Canonical Correlation Analysis (CCA) algorithm. The set of U sets is a collection of independent variables, and the set of V is a collection of dependent variables. From this figure it can be seen that the relation sets of dependent and independent sets are multivariable.

2.3 Partial Least Square Canonical (PLSC)

Partial least square is a multivariate statistical technique that can handle multiple response variables and explanatory variables at the same time (Supriyadi et al., 2017). This analysis is a good alternative to the method of multiple regression analysis and principal component regression, because this method is more robust or invulnerable (Dwi, Rini, & Kusuma, 2012). Robust means that the parameters of the model do not change much when new samples are taken from the total population. PLSC is a predictive technique that can handle many independent variables, even if there is multicollinearity among these variables, it can be explained in Figure 6 below:

Figure 6. PLSC Visualization With Independent Variables Experiencing Multicollinearity.



In studies that have made comparisons between PLS and PLS Canocical (PLSC), it is found that the relationship between characteristics and parameters with canonical correlation analysis and partial regression analysis is least squares. Both methods find a significant relationship between parameters.Likewise, the results of the new data compression method for estimating optimal latent variables in multi-variant classification and regression problems in which more than one response variable. Latent variables were found by combining the PLS methodology and canonical correlation analysis (CCA). The PLSC method is able to extract predictive

information for latent variables more effectively than the usual PLS approach. PLSC is a simple modification of the PLS and PPLS algorithm.The following in Figure 7 is a general procedure flow chart of the PLSC algorithm, from the picture it can be seen that the PLSC algorithm is a modification of the PLS given the canonical function in its calculations in finding correlation values between variables.

Figure 7. Algorithm Syntax Least Square Canonical

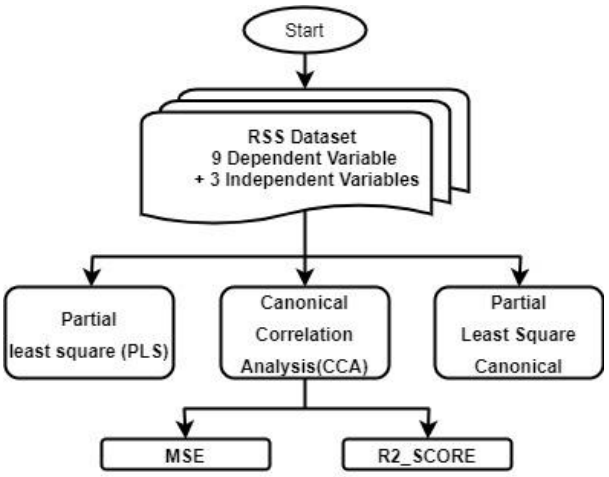
```
function PLS1(X, y, l)
    X^(0) ← X
    w^(0) ← X^T y / ||X^T y||, an initial estimate of w.
    for k = 0 to l − 1
        t^(k) ← X^(k) w^(k)
        t_k ← t^(k)^T t^(k)  (note this is a scalar)
        t^(k) ← t^(k) / t_k
        p^(k) ← X^(k)^T t^(k)
        q_k ← y^T t^(k)  (note this is a scalar)
        if q_k = 0
            l ← k, break the for loop
        if k < (l − 1)
            X^(k+1) ← X^(k) − t_k t^(k) p^(k)^T
            w^(k+1) ← X^(k+1)^T y
    end for
    define W to be the matrix with columns w^(0), w^(1), ..., w^(l−1).
    Do the same to form the P matrix and q vector.
    B ← W(P^T W)^−1 q
    B_0 ← q_0 − P^(0)^T B
    return B, B_0
```

3.Objectives Of The Study

In this section, we apply a step architecture for testing in order to find the correlation values of our variables using the three types of algorithms we have chosen, as explained in the introduction. Our test is a cross-decomposition module containing two main groups of algorithms: partial least square (PLS) and canonical correlation analysis (CCA). This set of algorithms is useful for finding linear relationships between two multivariate datasets: the X and Y arguments of the fit method are 2D arrays. The cross decomposition algorithm finds a fundamental relationship between the two matrices (X and Y).

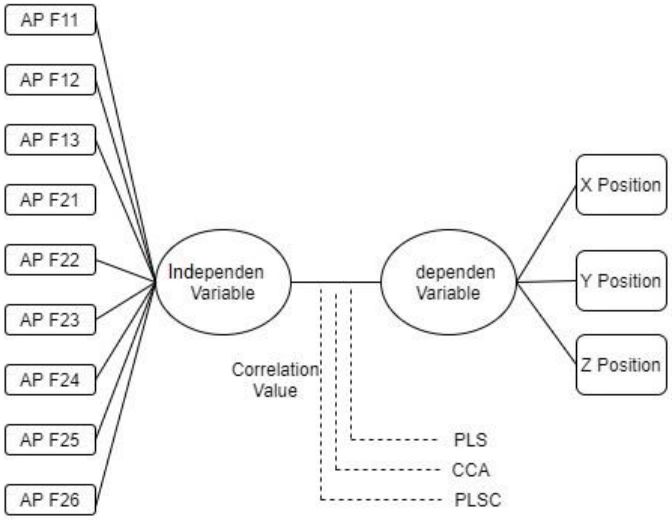
They are a latent variable approach to modeling covariance structures in these two spaces. They will try to find a multidimensional direction in X space which explains the direction of the maximum multidimensional variance in Y space. PLS-regression is very suitable when the predictor matrix has more variables than observations, and when there is multicollinearity between X values. Conversely, standard regression will fail in this case. The classes included in this module are PLS Regression PLS Canonical, and CCA which we tested in the form of the following schema in the flowchart of the experimental process:

Figure 8. General Architecture Testing The Value Of RSS Data Correlation In This Study.



The general architecture of testing the correlation value of RSS data in this study illustrated in Figure 8 above shows that the evaluation value taken is the MSE value and R2 Score.

Figure 9. Visualization Correlation of Dependent variables and Independent variables on the RSS Dataset collected from the University ofDinamikaBangsa Building.



In Figure 9 above is the Visualization of Correlation of Dependent variables and Independent variables in the RSS dataset which is collected from the University ofDinamikaBangsa Jambi Building. AP is an Access Point. AP F11 value is the 1st access point on the 1st floor, AP F12 value is the 2nd access point on the 1st floor, AP F13 value is the 3rd access point on the 1st floor. While AP F21 is the 1st access point on the floor 2, AP F22 is the 2nd access point, AP F23 is the 3rd access point on the 2nd floor on the 2nd floor, AP F24 is the 4th access point on the 2nd floor, AP F25 is the 5th access point on the 2nd floor, AP F26 is the 6th access point on the 2nd floor. Then as the dependent variable is the position estimate coordinates expressed as X, Y, Z position value. The position value expressed as the estimated variable is the dependent or affected variable. The following is a snippet code of testing program code of PLS:

```
from sklearn.cross_decomposition import PLSRegression
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
import pandas as pd
data = "./drive/My Drive/Dataset Pak Dodo/dataset.csv"
dataframe = pd.read_csv(data,sep=';',header=None,decimal=",")
dataset = dataframe.values
X = dataset[:,0:9]
Y = dataset[:,9:12]
print("n_components","r2_score","mse")
for i in range(1,4):
pls2 = PLSRegression(n_components=i)
pls2.fit(X, Y)
Y_pred = pls2.predict(X)
print(i,r2_score(Y,Y_pred),mean_squared_error(Y,Y_pred))
```

The above is the PLS algorithm program code that we are running to test the correlation values for our dataset. We use the python library for PLS functions using the syntax "from sklearn.cross_decomposition import PLSRegression", then the syntax "from sklearn.metrics import r2_score" as the R2_score evaluation metric and the "from sklearn.metrics import mean_squared_error" syntax for automatic evaluation for MSE metrics r2_score "as the R2_score evaluation metric and the" from sklearn.metrics import mean_squared_error "syntax for automatic evaluation for MSE metrics r2_score". Next, the program code of CCA algorithm:

```
from sklearn.cross_decomposition import CCA
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
import pandas as pd
data = "./drive/My Drive/Dataset Pak Dodo/dataset.csv"
dataframe = pd.read_csv(data,sep=';',header=None,decimal=",")
dataset = dataframe.values
X = dataset[:,0:9]
```

```
Y = dataset[:,9:12]
print("n_components","r2_score","mse")
for i in range(1,4):
cca = CCA(n_components=i)
cca.fit(X, Y)
 Y_pred = cca.predict(X)
print(i,r2_score(Y,Y_pred),mean_squared_error(Y,Y_pred))
```

The snippet of program code above is the CCA algorithm program code that we are running to test the correlation values of our dataset. We use the Python library for CCA functions using the syntax "from sklearn.cross_decomposition import CCA", then the syntax "from sklearn.metrics import r2_score" as the R2_score evaluation metric and the "from sklearn.metrics import mean_squared_error" syntax for automatic evaluation for the MSE metric import r2_score "as the R2_score evaluation metric and the" from sklearn.metrics import mean_squared_error "metric for automatic evaluation for MSE metrics for testing r2_score" CCA.
The following part is the snippet of program code for PLSC algorithm:

```
fromsklearn.cross_decomposition import PLSCanonical
fromsklearn.metrics import r2_score
fromsklearn.metrics import mean_squared_error
import pandas as pd
data = "./drive/My Drive/Dataset Pak Dodo/dataset.csv"
dataframe = pd.read_csv(data,sep=';',header=None,decimal=",")
dataset = dataframe.values
X = dataset[:,0:9]
Y = dataset[:,9:12]
print("n_components","r2_score","mse")
fori in range(1,10):
pls = PLSCanonical(n_components=i)
pls.fit(X, Y)
Y_pred = pls.predict(X)
print(i,r2_score(Y,Y_pred),mean_squared_error(Y,Y_pred))
```

The above part is the PLSC algorithm program code that we run to test the correlation values of our dataset. We use the Python library for PLSC functions using the syntax "from sklearn.metross_decomposition import PLSCanonical", then the syntax "from sklearn.metrics import r2_score" as the R2_score evaluation metric and the "from sklearn.metrics import mean_squared_error" syntax for automatic evaluation for the MSE metric import r2_score "as the R2_score evaluation metric and the" from sklearn.metrics import mean_squared_error "metric for automatic evaluation for the MSE metric for r2_score import testing PLSC.

4.Result and Analysis

4.1. Partial Least Square Regression (PLSR)

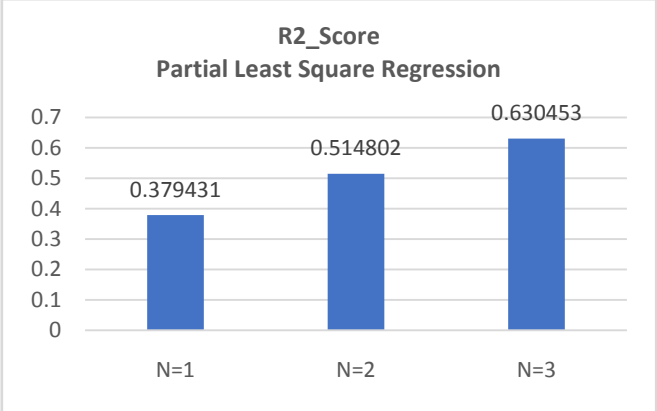
This section displays the results of testing the RSS dataset in this study using the Partial Least Square Regression algorithm. The following Table 1 is the evaluation of values from data testing using the PLS Regression algorithm.

Table 1. Evaluation Result of PLS

N Components	R2_Score	MSE
N=1	0.379431	59.09272
N=2	0.514802	47.62957
N=3	0.630453	44.89262

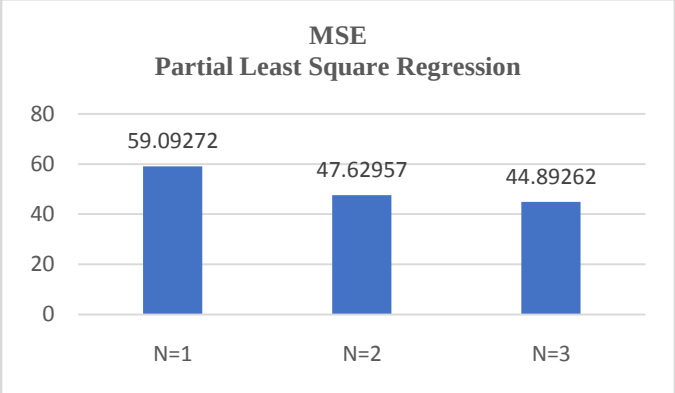
Table 1 above is a table of the results of the evaluation of values from data testing using the PLS Regression algorithm (Partial Least Square Regression). There are two evaluation values generated, namely the r2 score and the MSE value. From the 3 values of the N variable, namely 1, 2 and 3, the respective values of r2 score and MSE were obtained. Here is Figure 10 which reflects the graph of the value of the r2 score variable.

Figure 10. Visualization of R2_Score Evaluation Result using PLS for RSS.



From Figure 10 above, it can be seen that the highest r2 score is obtained with the value n = 3 and the lowest r2 score is the value n = 1. From the value of the variable r2 score, it is found that the correlation of a variable with another variable produces the best r2 score compared to the increasing number of correlated variables.

Figure 11. Visualization of MSE Evaluation Result using PLS for RSS.



Figures 11 are visualizations of table 1 about the MSE visualization data of the RSS data testing on the PLS algorithm. The highest r2_score value is obtained from the PLS test with a value of n = 3, while conversely the highest MSE value is obtained from a PLS test with a value of n = 1

4.2. Canonical Correlation Analysis (CCA)

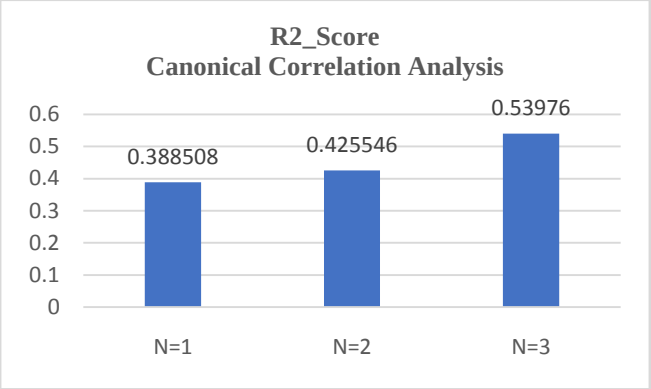
This section displays the results of testing the RSS dataset in this study using the Canonical Correlation Analysis (CCA) algorithm. The following table 2 is the evaluation of values from data testing using the CCA.

Table 2. Evaluation result of CCA

N Components	R2_Score	MSE
N=1	0.388508	56.45966
N=2	0.425546	53.70099
N=3	0.53976	46.88704

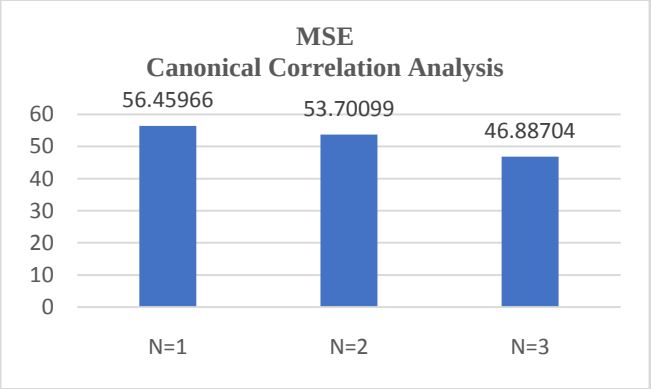
Table 2 above is a table of the results of evaluating the value of data testing using the CCA (Canonical Correlation Analysis) algorithm. There are two evaluation values generated, namely the r2 score and the MSE value. From the 3 values of the N variable, namely 1, 2 and 3, the respective values of r2 score and MSE were obtained. The following is figure 12 that reflects the graph of the variable value r2 score with the CCA (Canonical Correlation Analysis) algorithm.

Figure 12. Visualization of R2_Score Evaluation Result using CCA for RSS.



From Figure 12 above, it can be seen that the highest r2 score is obtained with the value of n = 3 and the lowest r2 score is the value of n = 1 in data correlation testing using the CCA (Canonical Correlation Analysis) algorithm. From the value of the variable r2 score, it is found that the correlation of a variable with another variable produces the best r2 score of 0.38850 compared to the increasing number of correlated variables with a value of 0.53976.

Figure 13. Visualization of MSE Evaluation Result using CCA for RSS.



Figures 13 are visualizations from table 2 of the MSE visualization data from the RSS data testing on the CCA algorithm. The highest r2_score value is obtained from the CCA test with a value of n = 3, while conversely the highest MSE value is obtained from the CCA test with a value of n = 1. The effect of MSE and r2_score on the CCA algorithm is almost similar to the PLS algorithm

4.3. Partial Least Square Canonical

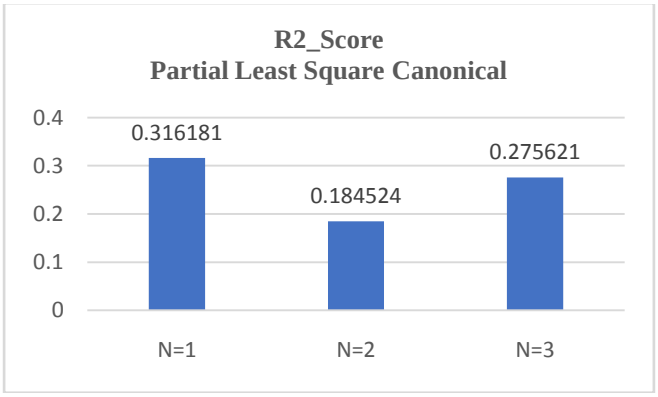
This section displays the results of testing the RSS dataset in this study using the Canonical Partial Least Square algorithm. Table 3 below is the evaluation of values from data testing using the Partial Least Square Canonical (PLSC).

Table 3. Evaluation Result of PLSC

N Components	R2_Score	MSE
N=1	0.316181	66.76277
N=2	0.184524	112.94
N=3	0.275621	111.184

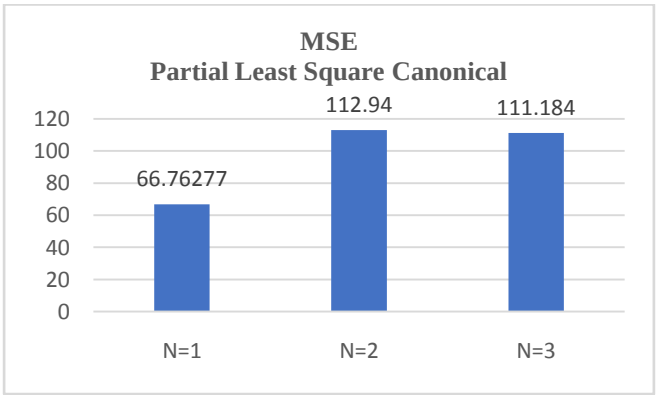
Table 3 above is a table of the results of the evaluation of values from testing data with the PLSC (Partial Least Square Canonical) algorithm. There are two evaluation values generated, namely the r2 score and the MSE value. From the 3 values of the N variable, namely 1, 2 and 3, the respective values of r2 score and MSE were obtained. The following is figure 14 which reflects the graph of the variable value r2 score with the PLSC (Partial Least Square Canonical) algorithm.

Figure 14. Visualization of R2 Score Evaluation Result using PLSC for RSS Dataset.



From Figure 14 above, it can be seen that the highest r2 score from testing with the PLSC (Partial Least Square Canonical) algorithm is obtained with the value of n = 1 and the lowest r2 score is the value of n = 2. From the value of the variable r2 score, it is found that the correlation of a variable with 2 other variables produces the best r2 score of 0.184524 compared to n=1 of correlated variables with a value of 0.316181.

Figure 15. Visualization of R2_Score Evaluation Result using PLSC for RSS Dataset.



Figures 15 are visualizations of table 3 about the MSE visualization data from testing RSS data on the PLSC algorithm. The highest r2_score value was obtained from the PLSC test with a value of n = 1, whereas the highest MSE value was obtained from a PLSC test with a value of n = 3. The effect of MSE and r2_score on the CCA algorithm is the opposite of the PLS and CCA algorithms. The following is a diagram of the r2_score value and MSE value from the average comparison of the 3 types of algorithms.

Figure 16. Value Comparison of R2_SCORE Results between PLS, CCA and PLSC algorithms.

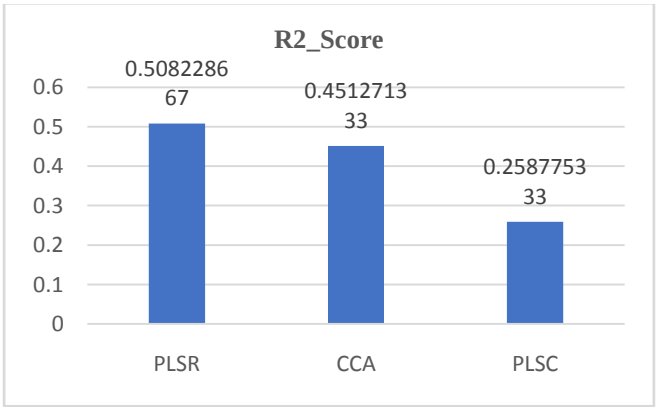


Figure 16 is a comparison of the r2_score value of the PLS, CCA and PLSC algorithms. From the comparison of the three PLS algorithms, on average, compared to CCA and PLSC techniques.

Figure 17. Value Comparison of MSE Results between PLS, CCA and PLSC algorithms.

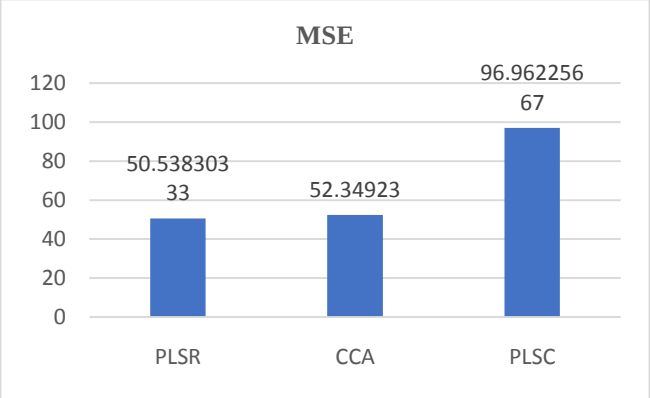


Figure 17 is a comparison of the MSE values of the PLS, CCA and PLSC algorithms. From the comparison of the three PLSC algorithms on average compared to CCA and PLS techniques. MSE values generated by PLSC on average reached 96.93. This very high MSE value indicates the high error of the validity of the variable correlation from our RSS data

5.Recommendations

From the results of the tests carried out in this study regarding the correlation between all variables in the RSS data that we took from the DinamikaBangsa University building. We obtained three types of evaluation results using different methods, namely Partial least square (PLS), Canonical Correlation Analysis (CCA), and Partial Least Square Canonical (PLSC). With the best results produced by the correlation test using Partial least squares (PLS). In comparison with the final research, besides that the dataset used is a self-crawled dataset so that comparisons with the use of correlation tests do not yet exist, but in the future it is hoped that testing on this dataset can be used with various more diverse correlation testing methods so that it can be seen that the dataset used is the more valid the eligibility. So, this RSS dataset can be tested using various existing methods according to the next technological development.

6.Conclusion

In this study we used the Received Signal Strength (RSS) dataset that we were collecting at the University ofDinamikaBangsa Jambi Building. RSS dataset is data that can be utilized in the development of signal processing technology that is very useful in various fields. Our RSS dataset has dependent and independent variables. However, we need to know whether our dataset is feasible or not to be tested with Machine Learning. Therefore, testing is needed to determine the correlation value between the dependent and independent variables in our dataset. Some algorithm that we use in testing the correlation value of this dataset are Partial least square (PLS), Canonical Correlation Analysis (CCA) and Partial Least Square Canonical (PLSC). From the test results obtained that the correlation of multiple variables in the dataset with the highest value of r squarenyo is PLS regression with a value of N = 3, R2_score of 0.630453 and MSE of 44.89262. The R2_score value obtained by PLS exceeds the target value of the correlation indicator with a good value of 0.6.

References

Abdi, H., Guillemot, V., Eslami, A., & Beaton, D. (2017). Canonical Correlation Analysis. *Encyclopedia of Social Network Analysis and Mining*. https://doi.org/10.1007/978-1-4614-7163-9

Abidin, D. Z., Nurmaini, S., Firsandava Malik, R., Erwin, Rasywir, E., & Pratama, Y. (2020). RSSI Data Preparation for Machine Learning. *Proceedings - 2nd International Conference on Informatics, Multimedia, Cyber, and Information System, ICIMCIS 2020*, 284–289. https://doi.org/10.1109/ICIMCIS51567.2020.9354273

Akhlaghi, S., Zhou, N., & Huang, Z. (2018). Adaptive adjustment of noise covariance in Kalman filter for dynamic state estimation. *IEEE Power and Energy Society General Meeting*, *2018-Janua*, 1–5. https://doi.org/10.1109/PESGM.2017.8273755

Anagnostopoulos, G. G., & Kalousis, A. (2019). A reproducible analysis of RSSI fingerprinting for outdoor localization using sigfox: Preprocessing and hyperparameter tuning. *2019 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2019*, 1–8. https://doi.org/10.1109/IPIN.2019.8911792

Chen, B., Liu, X., Zhao, H., & Principe, J. C. (2017). Maximum correntropy Kalman filter. *Automatica*, *76*, 70–77. https://doi.org/10.1016/j.automatica.2016.10.004

Dwi, R., Rini, K., & Kusuma, H. (2012). Pengenalan Wajah Dengan Algoritma Canonical Correlation Analysis (CCA). *JURNAL TEKNIK ITS*, *1*.

Felix, G., Siller, M., & Alvarez, E. N. (2016). A fingerprinting indoor localization algorithm based deep learning. *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*, 1006–1011. https://doi.org/10.1109/ICUFN.2016.7536949

Hadera, H., Ekström, J., Sand, G., Mäntysaari, J., Harjunkoski, I., & Engell, S. (2019). Integration of production scheduling and energy-cost optimization using Mean Value Cross Decomposition. *Computers and Chemical Engineering*, *129*. https://doi.org/10.1016/j.compchemeng.2019.05.002

Kumar, S., Kumar, V., & Nirmal, S. J. (2021). *Mask Detection Turkish Journal of Computer and Mathematics Education Research Article*. *12*(12), 1541–1546.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*(December 2016), 11–26. https://doi.org/10.1016/j.neucom.2016.12.038

Lu, A., Wang, W., Bansal, M., Gimpel, K., & Livescu, K. (2015). Deep multilingual correlation for improved word embeddings. *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 250–256. https://doi.org/10.3115/v1/n15-1028

Memenuhi, U., Mata, T., Teknik, K., & Data, A. (2017). *" Cannonical Correlation Analysis (CCA)."*

Sánchez-Rodríguez, D., Quintana-Suárez, M. A., Alonso-González, I., Ley-Bosch, C., & Sánchez-Medina, J. J. (2020). Fusion of channel state information and received signal strength for indoor localization using a single access point. *Remote Sensing*, *12*(12). https://doi.org/10.3390/rs12121995

Schmidt-Hieber, J. (2017a). Nonparametric regression using deep neural networks with ReLU activation function. *Arxiv Analytics Statistic*. Retrieved from http://arxiv.org/abs/1708.06633

Schmidt-Hieber, J. (2017b). Nonparametric regression using deep neural networks with ReLU activation function. *Arxiv Analytics Statistic*.

Shao, J., Wang, L., Zhao, Z., su, F., & Cai, A. (2016). Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval. *Neurocomputing*, *214*, 618–628. https://doi.org/10.1016/j.neucom.2016.06.047

Supriyadi, E., Mariani, S., & Sugiman. (2017). Perbandingan Metode Partial Least Square (Pls) Dan Principal Component Refression (Pcr) Untuk Mengatasi Multikolinearitas Pada Model Regresi Berganda. *UNNES Journal of Mathematics*, *6*(2), 117–128.

Vu, H., Koo, B., & Choi, S. (2017). Frequency detection for SSVEP-based BCI using deep canonical correlation analysis. *2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings*, 1983–1987. https://doi.org/10.1109/SMC.2016.7844531

Xie, Y., Wang, Y., Nallanathan, A., & Wang, L. (2016). An Improved K-Nearest-Neighbor Indoor Localization Method Based on Spearman Distance. *IEEE Signal Processing Letters*, *23*(3), 351–355. https://doi.org/10.1109/LSP.2016.2519607

Zhang, S., Choromanska, A., & Lecun, Y. (2015). Deep learning with elastic averaging SGD. *Advances in Neural Information Processing Systems*, *2015-Janua*, 685–693.

Zhang, Y., Zhou, G., Jin, J., Wang, X., & Cichocki, A. (2014). Frequency recognition in ssvep-based BCI using multiset canonical correlation analysis. *International Journal of Neural Systems*, *24*(4). https://doi.org/10.1142/S0129065714500130

Zhang, Z., Yuan, Y. H., Shen, X. B., & Li, Y. (2018). Low Resolution Face Recognition and Reconstruction Via Deep Canonical Correlation Analysis. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2018-April*, 2951–2955. https://doi.org/10.1109/ICASSP.2018.8461985

Zou, H., Jin, M., Jiang, H., Xie, L., & Spanos, C. J. (2017). WinIPS: WiFi-based non-intrusive indoor positioning system with online radio map construction and adaptation. *IEEE Transactions on Wireless Communications*, *16*(12), 8118–8130. https://doi.org/10.1109/TWC.2017.2757472