ICIC Express Letters
Part B: Applications
Volume **13**, Number **10**, October **2022**

ICIC International ©2022 ISSN 2185-2766
pp. 1101–1108

DEVELOPMENT OF GUI R-SHINY FOR CLUSTERING RIVER WATER QUALITY INDEX

Budi Warsito[1,3,*], Hasbi Yasin[1], Sri Sumiyati[2] and Ali Mahmudan[1]

[1]Department of Statistics
Faculty of Sciences and Mathematics
[2]Department of Environmental Engineering
Faculty of Engineering
[3]School of Postgraduate Studies
Diponegoro University
Prof. Soedarto, SH. Street, Tembalang, Semarang 50275, Indonesia
{ hasbiyasin; srisumiyati; alimahmudan }@live.undip.ac.id
*Corresponding author: budiwarsito@live.undip.ac.id

Received January 2022; accepted April 2022

Abstract. *The use of applications that are open source makes it easier for users to get sources because they always develop quickly following scientific developments and are free to access. However, most open-source software has limitations in terms of writing programming scripts which are time-consuming and look complicated. Interface development is needed so that it is easier for users to implement program code that is made according to the required application. This paper discusses the preparation of an R-Shiny-based interface with the application of clustering rivers in Central Java Province based on water quality. The clustering method used is Hierarchical Agglomerative Clustering. The resulting interface has made it easier for users with various scientific backgrounds to operationalize the program code. The resulting interface program has also succeeded in grouping the objects in question based on well-defined variables.*

Keywords: Interface, R-Shiny, Clustering, Hierarchical agglomerative, Water quality, Open-source

1. **Introduction.** The development of statistical models equipped with computer-based programming has been widely applied in various fields. Interface-based applications make it easier for users to run the computing program. The use of open-source software to build application interfaces has become a necessity. This is very necessary considering that free software makes it easier for all parties to take advantage of it. One useful tool is R-Shiny which is an application based on RStudio. Several kinds of literature related to the use of the R language for programming statistical models have been carried out, including in clustering [1], classification [2], machine learning regression [3], and time series [4]. Meanwhile, the use of R-Shiny for interface creation has also begun to be developed as in [5-7] and has become a positive competitor to the use of another language such as GUI Matlab [8,9]. In this study, the development of the R-Shiny GUI was applied to clustering rivers in Central Java, Indonesia based on their water quality. The clustering method used is Hierarchical Agglomerative Clustering. It is important to develop this research, considering that the river is one of the sources of water that has an important role in human life.

Rivers have many functions, both for human life, the natural environment, restoring water quality, distributing floods, and the main generator of flora and fauna ecosystems [10]. However, the problem that arises at this time is the quality of river water which is still not clean due to pollution. The water quality monitoring in Indonesia did not meet the

DOI: 10.24507/icicelb.13.10.1101

requirements for drinking water [11]. Research conducted by Ramadhani et al. [12] who reviewed the Analysis of Bengawan Solo River Water Quality Pollution Due to Industrial Waste in Kebakkramat District, Karanganyar Regency stated that the distribution of water pollution in the Bengawan Solo River showed a fluctuating trend in the degree of sewage contamination. This makes it difficult to access proper drinking water. Currently, the State of Indonesia has inequality in access to safe drinking water. Areas in the center of the capital city will receive more clean water than remote areas. Therefore, it is necessary to group rivers based on water quality parameters. Several parameters that can be used to measure water quality are pH, DS, TS, BOD, COD, DO, Fecal Coliform, Sulfide, Phenol, and Atrazine [13]. Therefore, it is important to develop an R-Shiny application for clustering rivers based on water quality which is the goal of this research. The major contribution and significance of this research is the development of a graphical interface based on the R-Shiny program for clustering rivers in Central Java, Indonesia, according to their water quality. The developed interface design can make it easier for users to operate the program, and it is simpler than using the source code directly. In order to make this writing more structured, the organization of this paper is summarized as follows. Section 2 describes the proposed methodology and Section 3 discusses the main experimental results. Finally, Section 4 presents the conclusions.

2. **Methods.** Hierarchical Agglomerative Clustering algorithm is a clustering method that starts from combining two objects with the closest distance into only one cluster and then will continue until finally a cluster is formed consisting of all objects. In hierarchical clustering, the data set is divided into a sequence of nested partitions [14]. The agglomerative method has several techniques in calculating the closest distance, namely Single Linkage, Complete Linkage, Average Linkage, Centroid Method, and Ward's Method. In general, the difference between these methods is based on the calculation of the distance. The Single Linkage method performs grouping based on the closest distance between objects [15,16], while Complete Linkage performs grouping based on the maximum distance or complete linkage based on the maximum distance [17,18]. The distance calculation on the Average Linkage is carried out by calculating the distance between two clusters which is called the average distance calculated for each cluster [19] while the determination of the distance between the two clusters formed by the Centroid Method is based on the proximity of the cluster's centroid distance [20]. On the other hand, Ward's Method is based on the Sum Square Error (SSE) criteria with a measure of homogeneity between two objects based on the most minimal SSE [21].

The selection of the optimal number of clusters can be done using the Calinski-Harabasz Pseudo F-Statistic method. Pseudo F-Statistic is one of the commonly used methods to determine the optimum number of groups by looking at the highest Pseudo F value. The higher the Pseudo F value, the better the cluster formed [22]. The Pseudo F formula is written in the following equation:

$$\text{Pseudo F} = \frac{\left(\dfrac{R^2}{c-1}\right)}{\left(\dfrac{1-R^2}{n-c}\right)}$$

$$R^2 = \frac{\text{SST} - \text{SSW}}{\text{SST}}$$

$$\text{SST} = \sum_{i=1}^{n_c} \sum_{j=1}^{c} \sum_{k=1}^{p} \left(x_{ij}^k - \bar{x}^k\right)^2$$

$$\text{SSW} = \sum_{i=1}^{n_c} \sum_{j=1}^{c} \sum_{k=1}^{p} \left(x_{ij}^k - \bar{x}_j^k \right)^2$$

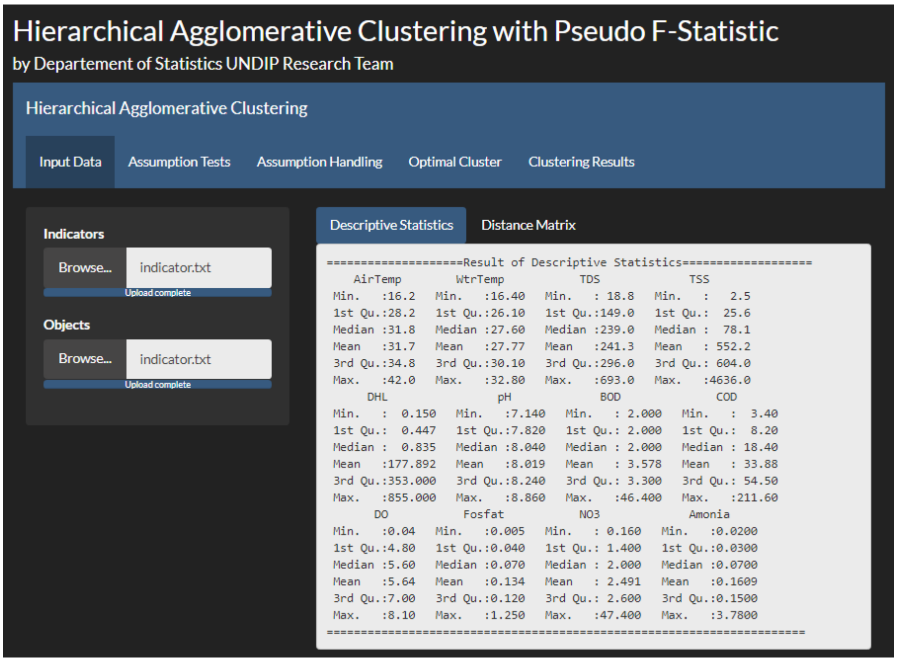
where R^2 is the proportion of the sum of the squares of the distances between the centers of the group by the number of squares of the sample to the overall average, SST is the total sum of the squares of the distances to the overall average, SSW is the total sum of squares of the sample distances to the group mean, n is the sample size, c is the number of groups, n_c is the number of data in the i-th group, p is the number of variables, x_{ij}^k is the i-th sample in the j-th group and the k-th variable, $\bar{x}^k$ is the sample mean on the k-variable, and $\bar{x}_j^k$ is the sample mean in the j-th group & the k-th variable.

The program code used for clustering is R, which is a statistical programming language created by Ross Ihaka and Robert Gentleman from the Department of Statistics, University of Auckland, New Zealand [23]. It can be used for the analysis and manipulation of statistical data or statistical modeling and graphs. One application in R that can create a web-based User Interface menu is R-Shiny. Shiny is an R package that makes it easy to build interactive web applications (`apps`) straight from R. Shiny allows users to build interactive web apps. Shiny is available at RStudio, an add-on to R. RStudio allows the user to run R in a more user-friendly environment. It is open-source (i.e., free) and available at http://www.rstudio.com/. Shiny combines the computational power of statistics in R and its interaction with the modern web.

Components of R-Shiny are divided into two groups, namely User Interface (UI) and Server (server). User Interface (UI) is a function that defines the web appearance of our application. Its function contains all inputs and outputs that will be displayed in the app. The UI contains a control panel to control input in the form of data, variables, models, depending on the complexity of the module. Data input is done by command on the function that has been given an identity or id (`input$id<-()` on the server). After the data is input, the input data will be processed as needed and the results will be defined using the command `output$(output id)<-()`. Output id is the identity used to call the processed output to the UI function. The server is a function that defines the analysis work logic from the server-side of the application. This section is the brain of the program in charge of performing simulations, analyzing various data according to the user's choice, and then sending the results to the output section. Both UI and Server components are called to run via the ShinyApp (`shinyapp`) application function.

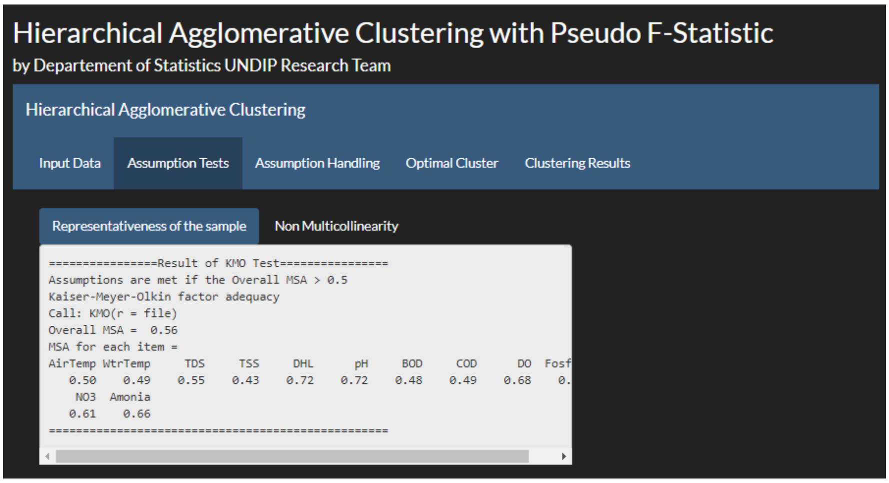
3. **Main Results.** In Graphical User Interface (GUI) development, the R-Shiny program was used. Views in GUI applications are set on the ShinyApp function by using UI (User Interface) objects. The display structure in the UI includes placing commands for input and displaying output using the `ui<-fluidPage()` command. The `fluidPage()` function that defines the basic GUI display window is written with the command `ui<-fluidPage (navbarPage())`. In this study, the GUI is designed into five main views or called tab panels. The first tab is "Input Data" which is designed to be two main sides. The first is the sidebar (`sidebarPanel()`), which houses the `fileInput()` function. This panel is used for data input purposes. Meanwhile, the second side is the main panel `mainPanel()`, where descriptive statistics and distance matrix outputs are displayed.

The second tab is "Assumption Test" which is designed to be a two-part tab panel. This tab is organized using `tabsetPanel()` which is the panel tab for representativeness of the sample and non-multicollinearity. This section is used to check assumptions on the data used in the study. Sample representativeness assumptions were checked through the overall MSA. If the overall MSA value is more than 0.5 then the assumption is fulfilled. The assumption of non-multicollinearity is checked through the VIF value. If the VIF value is more than 10 then the assumption is violated and we can solve it by using PCA.



```
===================Result of Descriptive Statistics===================
    AirTemp         WtrTemp           TDS              TSS
 Min.   :16.2   Min.   :16.40   Min.   : 18.8   Min.   :   2.5
 1st Qu.:28.2   1st Qu.:26.10   1st Qu.:149.0   1st Qu.:  25.6
 Median :31.8   Median :27.60   Median :239.0   Median :  78.1
 Mean   :31.7   Mean   :27.77   Mean   :241.3   Mean   : 552.2
 3rd Qu.:34.8   3rd Qu.:30.10   3rd Qu.:296.0   3rd Qu.: 604.0
 Max.   :42.0   Max.   :32.80   Max.   :693.0   Max.   :4636.0
      DHL                pH             BOD              COD
 Min.   :  0.150   Min.   :7.140   Min.   : 2.000   Min.   :  3.40
 1st Qu.:  0.447   1st Qu.:7.820   1st Qu.: 2.000   1st Qu.:  8.20
 Median :  0.835   Median :8.040   Median : 2.000   Median : 18.40
 Mean   :177.892   Mean   :8.019   Mean   : 3.578   Mean   : 33.88
 3rd Qu.:353.000   3rd Qu.:8.240   3rd Qu.: 3.300   3rd Qu.: 54.50
 Max.   :855.000   Max.   :8.860   Max.   :46.400   Max.   :211.60
       DO            Fosfat             NO3              Amonia
 Min.   :0.04   Min.   :0.005   Min.   : 0.160   Min.   :0.0200
 1st Qu.:4.80   1st Qu.:0.040   1st Qu.: 1.400   1st Qu.:0.0300
 Median :5.60   Median :0.070   Median : 2.000   Median :0.0700
 Mean   :5.64   Mean   :0.134   Mean   : 2.491   Mean   :0.1609
 3rd Qu.:7.00   3rd Qu.:0.120   3rd Qu.: 2.600   3rd Qu.:0.1500
 Max.   :8.10   Max.   :1.250   Max.   :47.400   Max.   :3.7800
======================================================================
```

Figure 1. GUI for hierarchical clustering of river water quality using R-Shiny

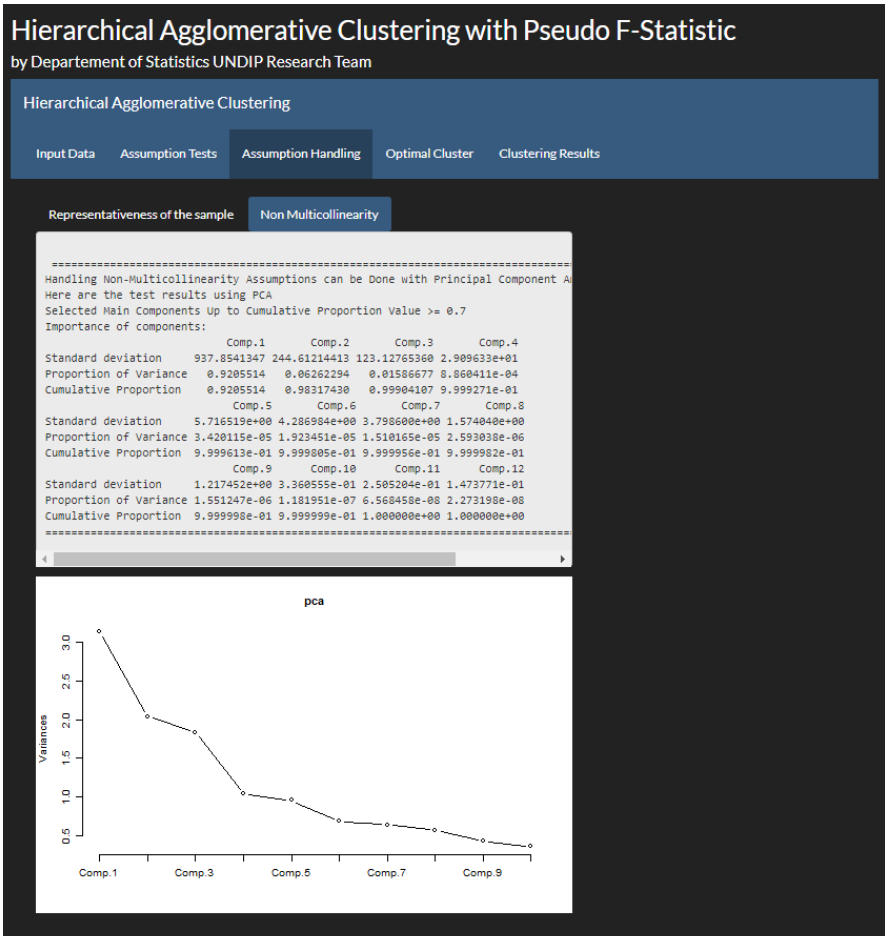

Hierarchical Agglomerative Clustering with Pseudo F-Statistic
by Departement of Statistics UNDIP Research Team
Hierarchical Agglomerative Clustering
Input Data
Assumption Tests
Assumption Handling
Optimal Cluster
Clustering Results
Representativeness of the sample
Non Multicollinearity

```
================Result of KMO Test================
Assumptions are met if the Overall MSA > 0.5
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = file)
Overall MSA =  0.56
MSA for each item =
AirTemp WtrTemp     TDS     TSS     DHL      pH     BOD     COD      DO  Fosf
   0.50    0.49    0.55    0.43    0.72    0.72    0.48    0.49    0.68    0.
    NO3  Amonia
   0.61    0.66
==================================================
```

Figure 2. Assumption tests for water quality index using GUI R-Shiny

The third tab is "Assumption Handling" with a display window design that is compiled using `tabsetPanel()` and is divided into two-panel tabs, namely the panel tab for the assumptions of Representativeness of the Sample and Non Multicollinearity. This section is used to handle assumptions if the assumptions are violated. The assumption of representativeness of the sample is handled by adding variables while the non-multicollinearity assumption is handled using PCA.

The fourth tab is "Optimal Cluster" with the display window divided into two main sides, namely the side panel and the main panel. The side panel (`sidebarPanel()`) is the location of the function for selecting the closest distance determination method (`selectInput()`), maximum input of the desired number of clusters (`textInput()`), and input of the number of main components (`textInput()`). Meanwhile, the main panel (`mainPanel()`) is a place for output in the form of Pseudo F-Statistics. This value can be obtained by clicking "Run" which functions as `actionButton()`. This tab section will display the results of Pseudo F values based on the original variable and Pseudo F values based on the principal component variables obtained from PCA.

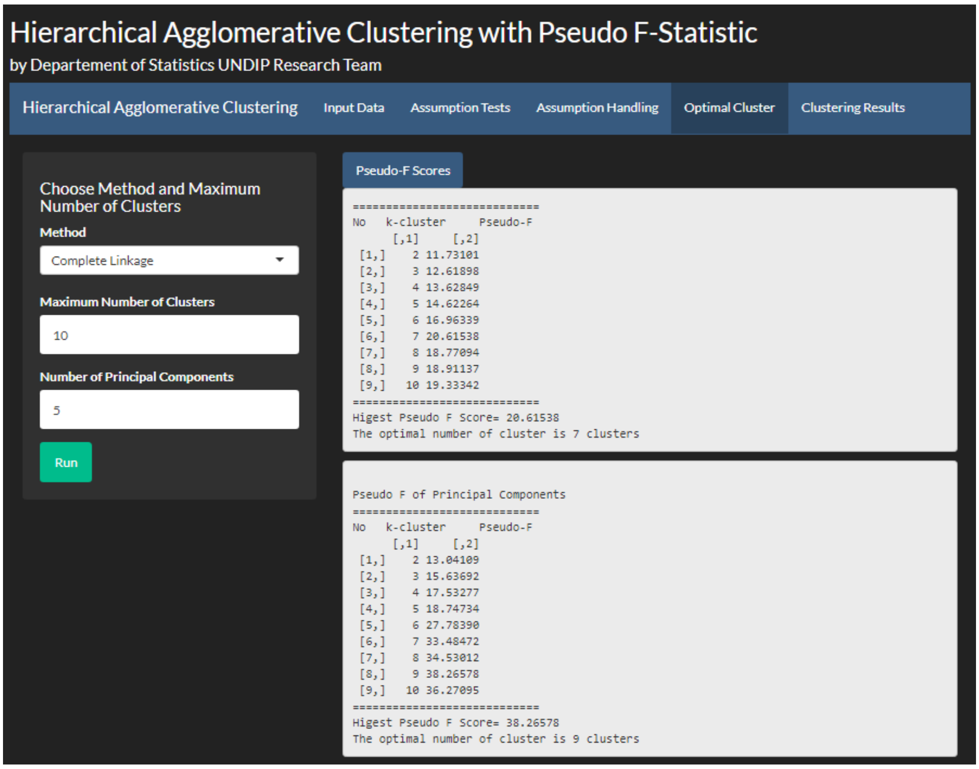


Hierarchical Agglomerative Clustering with Pseudo F-Statistic

by Departement of Statistics UNDIP Research Team

```
 ================================================================================
Handling Non-Multicollinearity Assumptions can be Done with Principal Component A
Here are the test results using PCA
Selected Main Components Up to Cumulative Proportion Value >= 0.7
Importance of components:
                          Comp.1       Comp.2       Comp.3       Comp.4
Standard deviation     937.8541347 244.61214413 123.12765360 2.909633e+01
Proportion of Variance   0.9205514   0.06262294   0.01586677 8.860411e-04
Cumulative Proportion    0.9205514   0.98317430   0.99904107 9.999271e-01
                           Comp.5       Comp.6       Comp.7       Comp.8
Standard deviation     5.716519e+00 4.286984e+00 3.798600e+00 1.574040e+00
Proportion of Variance 3.420115e-05 1.923451e-05 1.510165e-05 2.593038e-06
Cumulative Proportion  9.999613e-01 9.999805e-01 9.999956e-01 9.999982e-01
                           Comp.9      Comp.10      Comp.11      Comp.12
Standard deviation     1.217452e+00 3.360555e-01 2.505204e-01 1.473771e-01
Proportion of Variance 1.551247e-06 1.181951e-07 6.568458e-08 2.273198e-08
Cumulative Proportion  9.999998e-01 9.999999e-01 1.000000e+00 1.000000e+00
================================================================================
```

FIGURE 3. Assumption handling for water quality index using GUI R-Shiny



Hierarchical Agglomerative Clustering with Pseudo F-Statistic

by Departement of Statistics UNDIP Research Team

Choose Method and Maximum Number of Clusters

Method: Complete Linkage

Maximum Number of Clusters: 10

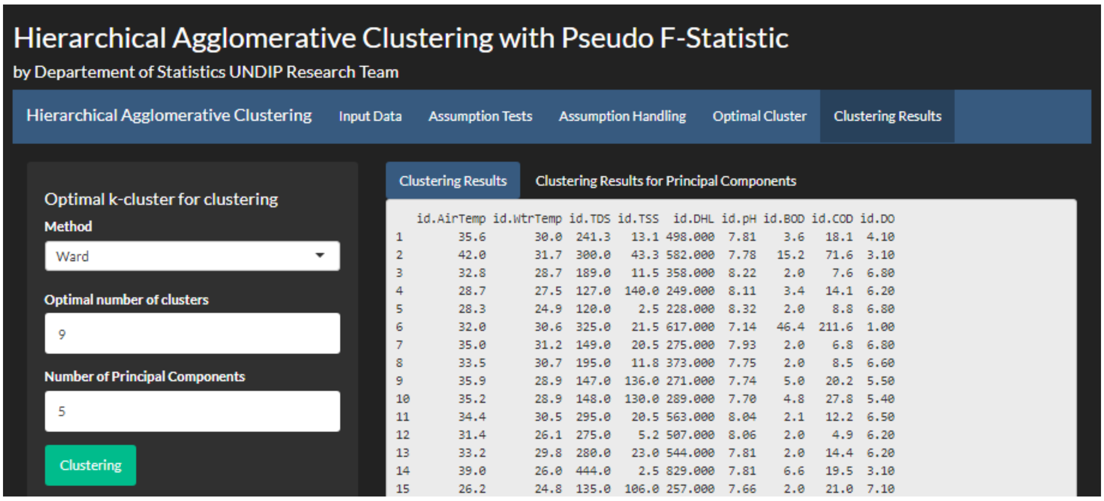
Number of Principal Components: 5

```
============================
No   k-cluster     Pseudo-F
      [,1]     [,2]
 [1,]    2 11.73101
 [2,]    3 12.61898
 [3,]    4 13.62849
 [4,]    5 14.62264
 [5,]    6 16.96339
 [6,]    7 20.61538
 [7,]    8 18.77094
 [8,]    9 18.91137
 [9,]   10 19.33342
============================
Higest Pseudo F Score= 20.61538
The optimal number of cluster is 7 clusters
```

```
Pseudo F of Principal Components
============================
No   k-cluster     Pseudo-F
      [,1]     [,2]
 [1,]    2 13.04109
 [2,]    3 15.63692
 [3,]    4 17.53277
 [4,]    5 18.74734
 [5,]    6 27.78390
 [6,]    7 33.48472
 [7,]    8 34.53012
 [8,]    9 38.26578
 [9,]   10 36.27095
============================
Higest Pseudo F Score= 38.26578
The optimal number of cluster is 9 clusters
```

FIGURE 4. Optimal cluster based on Pseudo F

The fifth tab is "Clustering Results" with the display window divided into two main sides, namely the side panel and the main panel. The side panel (`sidebarPanel()`) is a place for selecting the closest distance determination method (`selectInput()`), input the optimal number of clusters (`textInput()`), and input the number of main components (`textInput()`). Meanwhile, the main panel (`mainPanel()`) is used as a place to display the output of the clustering results, the average indicator variables for each cluster, and the dendrogram or agglomeration. These results can be obtained by clicking "Clustering" which functions as `actionButton()`. This tab section will display the clustering results based on the original variables and the clustering results based on the main component variables obtained from PCA.


Hierarchical Agglomerative Clustering with Pseudo F-Statistic
by Departement of Statistics UNDIP Research Team
Hierarchical Agglomerative Clustering
Input Data
Assumption Tests
Assumption Handling
Optimal Cluster
Clustering Results
Optimal k-cluster for clustering
Clustering
Clustering Results
Clustering Results for Principal Components

Method: Ward

Optimal number of clusters: 9

Number of Principal Components: 5

```
   id.AirTemp id.WtrTemp id.TDS id.TSS  id.DHL id.pH id.BOD id.COD id.DO
1        35.6       30.0  241.3   13.1 498.000  7.81    3.6   18.1  4.10
2        42.0       31.7  300.0   43.3 582.000  7.78   15.2   71.6  3.10
3        32.8       28.7  189.0   11.5 358.000  8.22    2.0    7.6  6.80
4        28.7       27.5  127.0  140.0 249.000  8.11    3.4   14.1  6.20
5        28.3       24.9  120.0    2.5 228.000  8.32    2.0    8.8  6.80
6        32.0       30.6  325.0   21.5 617.000  7.14   46.4  211.6  1.00
7        35.0       31.2  149.0   20.5 275.000  7.93    2.0    6.8  6.80
8        33.5       30.7  195.0   11.8 373.000  7.75    2.0    8.5  6.60
9        35.9       28.9  147.0  136.0 271.000  7.74    5.0   20.2  5.50
10       35.2       28.9  148.0  130.0 289.000  7.70    4.8   27.8  5.40
11       34.4       30.5  295.0   20.5 563.000  8.04    2.1   12.2  6.50
12       31.4       26.1  275.0    5.2 507.000  8.06    2.0    4.9  6.20
13       33.2       29.8  280.0   23.0 544.000  7.81    2.0   14.4  6.20
14       39.0       26.0  444.0    2.5 829.000  7.81    6.6   19.5  3.10
15       26.2       24.8  135.0  106.0 257.000  7.66    2.0   21.0  7.10
```

FIGURE 5. Clustering results for water quality index using GUI R-Shiny

KMO test is used for assumption test of representativeness of the sample. The sample has represented the population if the KMO value is > 0.5. From the output results, the obtained KMO value is equal to 0.56 which is greater than 0.5. It means that the sample used is representative of the population. Meanwhile, the non-multicollinearity assumption is met because all variables have a VIF value less than 10. The optimal number of clusters is determined to choose which method can form the optimal cluster. The purpose of the optimal cluster is that the cluster that has been formed has high diversity between its clusters and has high similarity between its cluster members. Determination of the optimal number of clusters is carried out using the Calinski-Harabasz Pseudo F-Statistic method where this method will produce an output in the form of a Pseudo F value. The highest

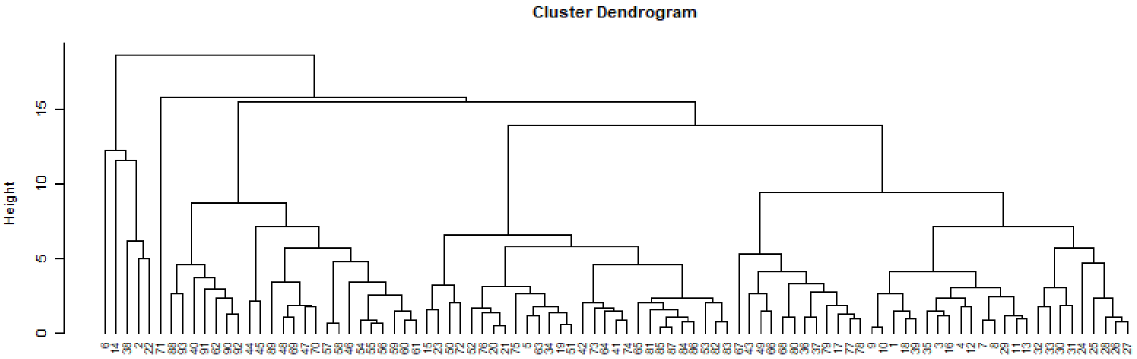


FIGURE 6. Dendrogram of hierarchical agglomerative clustering for first period monitoring

Table 1. Pseudo F values

Number of clusters	Pseudo F				
	Single linkage	Complete linkage	Average linkage	Centroid method	Ward's method
2	11.731	11.731	11.731	11.731	16.966
3	12.619	12.619	12.619	12.619	16.679
4	13.628	13.628	13.628	13.628	18.136
5	12.379	14.623	14.623	14.189	19.341
6	11.772	16.963	12.425	11.772	20.139
7	10.720	20.615	11.212	10.720	21.207
8	9.943	18.771	9.943	9.943	21.296
9	9.155	18.911	10.487	9.991	21.464
10	8.486	19.333	9.817	9.392	21.002

Pseudo F value will be selected as the optimal number of clusters because the higher the Pseudo F value, the better the cluster formed. Pseudo F values are presented in Table 1.

Based on Table 1, the highest Pseudo F value is 21.464 which is the Pseudo F value in the Ward's method with the number of clusters 9. This means that the Ward's method with the number of clusters 9 is the optimal number of clusters formed in clustering using Hierarchical Agglomerative Clustering. After knowing the optimal number of clusters, the last step in clustering rivers in Central Java Province based on water quality is to interpret the optimal number of clusters. Based on the determination of the optimal number of clusters using the Pseudo F value, the optimal number of clusters is 9 clusters using the Ward's method. The method gives the result that cluster 1 consists of 24 rivers, cluster 2 consists of 3 rivers, cluster 3 consists of 28 rivers, cluster 4 consists of 1 river, cluster 5 consists of 1 river, cluster 6 consists of 12 rivers, cluster 7 consists of 7 rivers, cluster 8 consists of 16 rivers, and cluster 9 consists of 1 river. Based on the clustering results, it appears that the Hierarchical Agglomerative Clustering algorithm method has succeeded in clustering objects of rivers in Central Java Province based on water quality variables. The interface that has been obtained has also been arranged interactively and is easy to operate.

4. **Conclusions.** The R-Shiny-based programming interface for clustering rivers based on water quality using the Hierarchical Agglomerative Clustering method has been developed. The developed interface has made it easier for users to operate programs and process data. Interface development for other models and applications in other fields can be developed for future research. The use of the features provided by R-Shiny can be further optimized to produce a better display.

Acknowledgment. This work is supported by the Directorate of Research and Community Service of the Ministry of Education, Culture, Research and Technology with contract No: 187-05/UN7.6.1/PP/2021.

REFERENCES

[1] S. Sivaarunagirinathan, B. A. Bala, S. Fairooz, G. Sasi, H. N. Upadhyay and V. Elamaran, Lossy data compression using k-means clustering on retinal images using RStudio, *2021 3rd International Conference on Advances in Computing Communication Control and Networking (ICAC3N)*, Greater Noida, India, pp.1772-1776, 2021.

[2] M. Hu and Y. Huang, atakrig: An R package for multivariate area-to-area and area-to-point kriging predictions, *Computers and Geosciences*, vol.139, 104471, 2020.

[3] R. M. Khalifa, S. Yacout and S. Bassetto, Developing machine-learning regression model with logical analysis of data (LAD), *Computers and Industrial Engineering*, vol.151, 106947, 2021.

[4] C. Tian, M. Theophile, J. Qian, Y. Zhang and Z. Hu, Prediction of time series analysis of power usage based on RStudio, *International Journal of Machine Learning and Computing*, vol.12, no.2, pp.68-72, 2022.

[5] Y. Li, Towards fast prototyping of cloud-based environmental decision support systems for environmental scientists using R Shiny and Docker, *Environmental Modelling and Software*, vol.132, 104797, 2020.

[6] R. Jelle, D. Nick and T. Lennert, The tale of the river Scheldt as told by historic maps – Building an RShiny 'side-by-side viewer' to visualize 16th-20th century maps, *Book of Abstracts*, p.136, 2019.

[7] E. A. Goto, K. Clarke and E. Keller, A tool to compute the landslide degree of risk using R-Studio and R-Shiny, *Proc. of the 4th ACM SIGSPATIAL International Workshop on Safety and Resilience*, New York, pp.1-7, 2018.

[8] S. He and P. Li, A MATLAB based graphical user interface (GUI) for quickly producing widely used hydrogeochemical diagrams, *Geochemistry*, vol.80, no.4, 125550, 2020.

[9] R. Santoso, B. Warsito and H. Yasin, Graphical interface of genetic optimization in neural network modelling for time series, *ICIC Express Letters, Part B: Applications*, vol.12, no.3, pp.301-306, 2021.

[10] B. Rahman, Correlation of cultural activity of river bank to tidal river transportation function, *International Conference on Coastal and Delta Areas*, Semarang, vol.3, pp.617-624, 2017.

[11] D. Ratnaningsih, Implementation of the STORET method on river water quality data in Indonesia, *Ecolab*, vol.4, no.1, pp.1-11, 2010.

[12] E. Ramadhani, A. N. Anna and M. Cholil, *Analysis of Bengawan Solo River Water Quality Pollution Due to Industrial Waste in Kebakkramat District, Karanganyar Regency*, Ph.D. Thesis, Universitas Muhammadiyah Surakarta, 2016.

[13] M. J. Alam, M. R. Islam, Z. Muyen, M. Mamun and S. Islam, Water quality parameters along rivers, *International Journal of Environmental Science and Technology*, vol.4, no.1, pp.159-167, 2007.

[14] Z. Nazari, M. Nazari and D. Kang, A bottom-up hierarchical clustering algorithm with intersection points, *International Journal of Innovative Computing, Information and Control*, vol.15, no.1, pp.291-304, 2019.

[15] F. Ros and S. Guillaume, A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise, *Expert Systems with Applications*, vol.128, pp.96-108, 2019.

[16] A. M. Jarman, *Hierarchical Cluster Analysis: Comparison of Single Linkage, Complete Linkage, Average Linkage and Centroid Linkage Method*, Department of Computer Science, Georgia Southern University, 2020.

[17] S. Sharma and N. Batra, Comparative study of single linkage, complete linkage, and Ward method of agglomerative clustering, *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, pp.568-573, 2019.

[18] B. Biggio, S. R. Bulò, I. Pillai, M. Mura, E. Z. Mequanint, M. Pelillo and F. Roli, Poisoning complete-linkage hierarchical clustering, *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, Joensuu, Finland, pp.42-52, 2014.

[19] T. Lavastida, K. Lu, B. Moseley and Y. Wang, Scaling average-linkage via sparse cluster embeddings, *Proc. of the 13th Asian Conference on Machine Learning, PMLR 157*, pp.1429-1444, 2021.

[20] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman and B. D. Satoto, Integration k-means clustering method and elbow method for identification of the best customer profile cluster, *IOP Conference Series: Materials Science and Engineering*, Surabaya, Indonesia, vol.336, no.1, 012017, 2018.

[21] Ö. Akay and G. Yüksel, Clustering the mixed panel dataset using Gower's distance and k-prototypes algorithms, *Communications in Statistics-Simulation and Computation*, vol.47, no.10, pp.3031-3041, 2018.

[22] M. C. L. Bue and S. Klasen, Identifying synergies and complementarities between MDGs: Results from cluster analysis, *Social Indicators Research*, vol.113, no.2, pp.647-670, 2013.

[23] R. Ihaka, R: Past and future history, *Computing Science and Statistics*, vol.392396, 1998.