Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

The ensemble distance on model-based clustering for regions clustering based on rainfall: The case of rainfall in West Java Indonesia

Triyani Hendrawati[a*], Aji Hamim Wigena[b], I Made Sumertajaya[b], Bagus Sartono[b], Anindya Apriliyanti Pravitasari[a] and Mohammad Hamid Asnawi[a]

[a]*Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Bandung, Indonesia*
[b]*Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia*

C H R O N I C L E

Article history:
Received: August 15, 2023
Received in revised format: October 25, 2023
Accepted: November 21, 2023
Available online: November 21, 2023

Keywords:
Time series clustering
Distance
Rainfall data
Ensemble distance

A B S T R A C T

Time series data clusters are being researched thoroughly. The distance metric drives the development of the clustering time series. The ARIMA model is one of the models that can be employed in model-based clustering, although differing model selection criteria can lead to uncertainty in the model. In this investigation, we created a technique for ensemble distance-based time series data clustering. To express the distance between two series, five distances based on the five model selection criteria are utilized. The average of the five distances reflects the distance of two time series data. According to the simulation results, the ensemble distance method could boost clustering accuracy by more than 11%. Based on the pattern of rainfall levels, we applied our methods to find clusters of locations in the Province of West Java (Indonesia). The findings indicate that the rainfall pattern in the same cluster is similar. The cluster model is effective and feasible for representing individual models in a cluster.

1. Introduction

Cluster analysis, a technique aimed at grouping similar objects together while keeping dissimilar ones apart (Gan et al., 2007), has emerged as a pivotal tool in the realm of data analysis and pattern recognition. In recent years, the application of cluster analysis has extended its reach to encompass a diverse range of data types. Notably, the utilization of time series data—a category of dynamic data where observations evolve sequentially—has gained prominence in various domains. Time series data clustering have the capacity to offer valuable insights into the temporal evolution of phenomena, and their analysis plays a pivotal role in a multitude of applications, including but not limited to the detection of data relationships, predictive modeling, recommendation systems, and the discovery of intricate data patterns (Aghabozorgi et al., 2015). researchers have applied time-series data clustering techniques across an array of fields. Gullo *et al.* (2012), for example, used the Dynamic Time Warping distance metric when combined with k-means to cluster the healthcare data. Meanwhile, Caiado *et al.* (2006) explored time series clustering in the economic domain, specifically in the context of industrial production index data. Corduas and Piccolo (2008) extended the utility of time series clustering to domains as diverse as economics and medicine.

In the pursuit of refining the art of clustering time series data, several researchers have ventured into this multifaceted domain. Studies by Aghabozorgi *et al.* (2015), Liao (2005), Rani and Sikka (2012), and Ergüner Özkoç (2021) have undertaken comprehensive reviews of the existing body of research on time series data clustering. These endeavors have unveiled the remarkable diversity of approaches and techniques employed in the field. For instance, Javed *et al.* (2020) undertook a comparative analysis of clustering algorithms, categorizing them into three distinct categories—partition, hierarchy, and density-based—while also evaluating the suitability of various distance measures, including the Euclidean, Dynamic Time Warping (DTW), and shape-based measures. The complexity of time series data presents a common challenge that has motivated researchers

* Corresponding author.
E-mail address: triyani.hendrawati@unpad.ac.id (T. hendrawati)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2023.11.015

to seek innovative solutions. Most time series data, due to their unique temporal structure and high dimensionality, defy the straightforward application of conventional clustering algorithms. Consequently, researchers in the field have directed their efforts towards the development of novel measures of similarity and dissimilarity. This endeavor is necessitated by the intrinsic characteristics of time-series data, where temporal dependencies and large dimensions add layers of intricacy (Keogh & Kasetty, 2003; Rani & Sikka, 2012). As a result, the search for effective distance metrics and the creation of customized similarity measures have become cornerstones of time series clustering research. To address the inherent challenges posed by time series data, clustering methods have undergone significant adaptations. Traditional algorithms of clustering have been expanded to fit the time-series format, or data from time-series has been turned into a more amenable structure, allowing the application of normal clustering techniques (Liao, 2005). The crucial role of similarity and dissimilarity metrics is underscored by Keogh and Kasetty (2003), emphasizing that these measures lie at the heart of clustering algorithms, shaping the outcomes and patterns that emerge from the data. Moreover, there have been instances where experts in specific domains have crafted custom distance metrics, such as Biabiany *et al.* (2020), who devised an expert distance metric for climate clustering.

Time series data clustering approaches offer a diverse array of techniques, each specifically designed to address the special features of time-series datasets. Typically, these approaches can be categorized into three primary methods: clustering based on raw data, feature based, and model based of model (Liao 2005). The selection of the methodology depends on the characteristics of the data and the goals of the analysis. Model-based clustering has gained traction due to its ability to accommodate time series with varying observation periods. In the realm of model-based time-series data clustering, one prominent model utilized is the Auto Regressive Integrated Moving Average (ARIMA). The application of ARIMA-based clustering has been employed by several researchers as an effective means to measure similarity between time series data, including Piccolo (1990), Piccolo (2010), Maharaj (2000), Kalpakis *et al.* (2001), Corduas and Piccolo (2008), and Triacca (2016).

One central challenge in employing ARIMA models for time series clustering is the selection of the most appropriate model from the plethora of possibilities. From a given set of time-series data, it is possible to derive multiple ARIMA models, each potentially suited to the data but distinct in terms of complexity and representation. The crux of the issue lies in the diverse selection criteria employed to identify the optimal model, including well-known measures such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and various others. The selection of a different model based on varying criteria has the potential to yield diverse results, casting a shadow of uncertainty over which model to choose for the analysis. This quandary holds profound implications for time series clustering. Specifically, it can lead to markedly different cluster results, as the choice of model plays a pivotal role in shaping the clustering outcomes. This issue is at the heart of the problem that the present study seeks to address: devising a robust and consistent approach to model-based clustering for rainfall patterns in West Java, Indonesia, amidst the inherent variability introduced by model selection criteria.

While existing approaches have explored clustering time-series data using various methods, this research charts a novel path with distinct objectives. Notably, Hendrawati *et al.* (2020) developed a method for clustering time-series data by harnessing the ensemble parameters of the ARIMA model. In contrast, the primary aim of this study is to forge a method for time-series data clustering employing the concept of ensemble distance. The ensemble distance method leverages the strengths of multiple models, each selected through diverse model selection criteria, rather than adhering to a single model choice. This approach offers an innovative solution to circumvent the limitations associated with rigid model selection, which can potentially yield inaccurate or unreliable clustering results. As we delve deeper into this research, we explore the intricate details of the ensemble distance method, its application to rainfall pattern clustering in West Java, Indonesia, and the invaluable insights it offers in unveiling the complex temporal dynamics of this region's meteorological data.

2. Materials and Method

2.1 ARIMA Model

Rainfall data, which is time series data, can be modeled using ARIMA. Here is the formulation of the model. The Autoregressive Model, denoted as AR (p), is mathematically represented as follows:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \tag{1}$$

The Autoregressive Moving Average model represented as ARMA (p, q), is formulated as follows: (p, q))

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{2}$$

The Autoregressive Integrated Moving Average, abbreviated as ARIMA (p, d, q), is expressed as follows:

$$\phi(B)(1-B)^d Y_t = \theta(B) e_t \tag{3}$$

Where: $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$; $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$; $e_t \sim$ N $(0, \sigma_e^2)$; p is order of AR, d is differencing, and q is order of MA (Wei 2006; Hendrawati *et al.* 2020).

2.2 Model Selection Criteria

There are several criteria for selecting the best model: Akaike's Information Criterion (AIC); Bayesian Information Criterion (BIC); Akaike's Information Criterion Bias Corected (AICc); Mean Absolute Percentage Error(MAPE); and Root Mean Squared Error(RMSE).

$$AIC = -2\log(\hat{\sigma}_e^2) + 2k \tag{4}$$

$$BIC = -2\log(\hat{\sigma}_e^2) + k\log(n) \tag{5}$$

$$AICc = AIC + \frac{2(k+1)(k+2)}{n-k-2} \tag{6}$$

$$\text{MAPE} = \frac{\sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \text{x}100 \right|}{n} \tag{7}$$

$$\text{RMSE} = \left(\frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n} \right)^{1/2} \tag{8}$$

where $\hat{\sigma}_e^2$ denotes the estimator for the variance of error; k denotes the number of parameters; n denotes the number of observations; y_t denotes the observation's value at time t; and $\hat{y}_t$ denotes the fit observation's value at time t (Cryer & Chan, 2008; Montgomery *et al.,* 2008).

2.3 Clustering Method

2.3.1 Ward Method

The Ward Method aims to join objects into groups where the variance within groups is minimized. The pairings of objects with the least increase in the Error Sum of Squares (ESS) are combined at each phase of Ward methods.

$$\text{ESS} = \sum_{j=1}^{n} (\boldsymbol{x}_j - \overline{\boldsymbol{x}})'(\boldsymbol{x}_j - \overline{\boldsymbol{x}}) \tag{9}$$

where $\boldsymbol{x}_j$ represents *a j-th* object and $\overline{\boldsymbol{x}}$ represents an average of objects (Eszergár-Kiss & Caesar, 2017; Jain & Dubes, 1988; Everitt et al., 2011; Kaufman & Rousseeuw, 1990; Murtagh & Legendre, 2014; Murtagh & Legendre, 2014).

2.3.2 Optimization of clusters number

To ascertain optimal number of clusters, we employ average of silhouette coefficient (SC). The SC represents averaging of distance between a data point and the others within its designated cluster, as well as all data points belonging to the nearest neighboring cluster.

$$S(i) = \frac{y(i) - x(i)}{\max\{x(i); y(i)\}} \tag{10}$$

$x(i)$ denotes the distance between i-th object to its own cluster; $y(i)$ denotes the distance between i-th object to its nearest neighbor cluster (Everitt *et al.* 2011; Jain & Dubes 1988; Kaufman & Rousseeuw 1990).

2.4 Cluster Model

Each cluster is represented by a time series data. In this instance, the prototype—a time series data set—is derived from the cluster's average value of data.

$$prototype = \text{average}\{x_{1,i}, x_{2,i}, \ldots, x_{c,i}\} \quad \forall\ c \in C \tag{11}$$

where $i = 1, 2, \ldots, n$ (Aghabozorgi et. al., 2015; Hendrawati, 2020).

The assessment of the cluster method's accuracy occurs following the formation of clusters. Each cluster created is subsequently compared to the reference or ground-truth data. Clusters that originate from the same underlying data will be

consolidated into the same group. In cases where data points from the same underlying group are assigned to different clusters, this indicates an error in the clustering process, referred to as a misclassification error. The accuracy is computed using the following formula:

$$\text{Accuracy} = \frac{n_o - n_{miss}}{n_o} \times 100\% \tag{12}$$

where n_o represent the number of object and n_{miss} represent the number of objects that misclassified.

2.5 Distances for Time-series Data

In 1990, Piccolo introduced a method for quantifying the similarity between two time-series data using a distance formula, as presented in Eq. (13). A collection of time-series data can be represented through various ARIMA models, with the model selection based on specific criteria. Choosing a model, for instance, based on the smallest AIC or BIC. The Piccolo method employs the ARIMA model approach to compute the similarity between two time-series datasets. Based on the invertible ARIMA process, X_t and Y_t can be expressed by the Autoregressive model (AR (∞)). The distance between ARIMA processes X_t and Y_t is distance between the coefficients $\hat{\pi}$ in AR (∞) using the equation:

$$d(X_t, Y_t) = \sqrt{\sum_{i=1}^{\infty} \left(\hat{\pi}_{i,x} - \hat{\pi}_{i,y} \right)^2} \tag{13}$$

which $\hat{\pi}_{i,x}$ and $\hat{\pi}_{i,y}$ represent coefficient of AR model for time-series data X_t and Y_t. Distance $d(X_t, Y_t)$ serves as a metric for quantifying the structural similarity between two ARIMA processes (Corduas & Piccolo, 2008; Piccolo, 1990, 2010).

2.6 Proposed Method

Debates exist regarding the most suitable criteria for choosing a model. According to some academics, choosing the best model should not solely rely on specific information criteria but should also consider alternative approaches (Anderson & Burnham, 2002; Brewer et al., 2016; Burnham & Anderson, 2004). On the other hand, Claeskens (2016) argued that if multiple estimators of model parameters are derived from the same population, combining these estimators may yield a superior model parameter estimator. Several researchers advocate for the use of multiple models (multimodels) rather than exclusively relying on a single model, considering it as an alternative to becoming constrained by a potentially incorrect model (Burnham and Anderson 2004). An approach for selecting a model from a collection of models is the averaging method (Claeskens, 2016).

In this research, the chosen methodology is the ensemble distance, specifically the average distance. Applying different model selection criteria, notably AIC, BIC, AICc, RMSE, and MAPE, five different distances are determined. For instance, Model A is selected based on the AIC criterion, while Models B, C, D, and E are determined using the BIC, AICc, RMSE, and MAPE criteria, respectively. The distance associated with each model measures the dissimilarity between two time series, aligning with the corresponding model selection criterion.

Subsequently, to represent these five distances collectively, an average distance is employed, as illustrated in the following equation:

$$\bar{d}_{x,y} = \frac{1}{5} \left(d_A + d_B + d_C + d_D + d_E \right) \tag{14}$$

where d_A, d_B, d_C, d_D, d_E are distance associate to Eq. (13) which based on model selection with criteria of AIC, BIC, AICc, RMSE, and MAPE respectively.

2.7 Simulation

The simulation involves the generation of data from three distinct clusters, each of which adheres to an Autoregressive (AR (2)) model. The parameter values for model A, model B and Model C are (0.2 , 0.1), (0.4 , 0.5), and (0.6 , 0.2), respectively. Each cluster generates a total of 10 series, resulting in 30 series of generated data. The observation period (t) for the generated data varies across six different values: 50, 75, 100, 150, 200, and 300 (Kumar dan Patel 2008; Hendrawati *et al.* 2020).

The generated data were first organized into clusters using the Piccolo distance method (Hendrawati *et al.* 2020) and the ensemble distance approach. Subsequently, the data were modeled using the Autoregressive (AR (p)) method, with p ranging from 1 to 15. Model selection was based on the minimization of the AIC criterion. The distances between time-series were computed using the distance formula, as defined in Equation (1). This process was repeated; the least BIC, AICc, RMSE, and MAPE were the new model selection criteria used in this process (Cryer & Chan, 2008; Montgomery DC, Jennings CL, 2008). The average of the five distances obtained from the previous steps was then calculated using a formula like Eq. (2).

Subsequently, clusters were determined using the Ward method (Eszergár-Kiss & Caesar, 2017; Everitt et al. , 2011; Jain & Dubes, 1988; Murtagh & Legendre, 2014; Kaufman & Rousseeuw, 1990). The quality of the clustering results was evaluated by calculating the percentage of correct cluster membership across 100 replications (Hendrawati *et al.,* 2020).
The simulation procedure is outlined as follows:

1. Generate a time series dataset consisting of three clusters according to the specified rules.
2. Modelling the generation time series data using the AR (p) model approach, where $p = 1, 2, \ldots, 15$. The optimal model is selected based on the minimization of Akaike's Information Criterion (AIC).
3. Compute the distances between time series using the distance formula provided in Eq. (13).
4. Repeat steps two and three, but this time utilize distinct standards for selecting models, including the smallest BIC , AICc , RMSE , and MAPE .
5. Calculate the average of the five distances obtained from the previous steps, using a formula in Equation (14).
6. Determine the clustering of the time series data using Ward's method.
7. Evaluate the accuracy of the clustering results by employing a formula, often detailed in Equation (12).
8. Repeat steps one to seven a total of 100 times.

3. Results

3.1 Results

Table 1 displays the percentage of correct cluster membership using the ensemble distance and Piccolo method. In the Piccolo distance method, various criteria were applied, including AIC, BIC, AICc, RMSE, and MAPE. When the observation period length (t) was set to 50, the lowest correct cluster membership percentage, at 72.47%, was observed in the Piccolo method when using the RMSE criterion. Conversely, for t = 75, 100, 150, 200, and 300, the Piccolo method with the MAPE criterion consistently exhibited the lowest correct cluster membership percentage among the different criteria. However, as illustrated in Fig. 1, it is evident that the ensemble distance method consistently yielded higher correct cluster membership percentages when compared to the Piccolo method.

Table 1
Percentage of correct cluster membership using the ensemble distance and Piccolo method

The length of observation period	The Piccolo method					The ensemble distance method
	AIC	BIC	AICc	RMSE	MAPE	
50	78,03%	83,8%	79,53%	**72,47%**	74,27%	**87,63%**
75	82,67%	86,4%	83%	79,97%	**78,23%**	**90,77%**
100	90,27%	88,1%	90,67%	89,13%	**81,43%**	**93,77%**
150	94,77%	93,33%	94,93%	94,5%	**85,3%**	**96,93%**
200	96,5%	97,03%	96,33%	96,8%	**86,27%**	**97,73%**
300	98,87%	98,93%	99%	98,57%	**87,8%**	**99,43%**

The ensemble distance method compared to Piccolo with the RMSE criterion for the length of the observation period (t) = 50, significantly increased the correct cluster membership, resulting in an increase of 15.16%. Furthermore, for t = 75, 100, 150, 200, and 300, the ensemble distance method achieved substantial enhancements in correct cluster membership. Specifically, it raised the correct cluster membership by 12.54%, 12.34%, 11.63%, 11.46%, and 11.63%, respectively, in comparison to the Piccolo method with the MAPE criterion.

3.2 Application for Rainfall Data

In this study, secondary data was employed, specifically monthly rainfall data (in millimeters), sourced from 26 rainfall monitoring stations located in West Java. The data spanned the years 2000 to 2009 and were acquired from the meteorology, climatology, and geophysics agency (BMKG) in Indonesia.

The rainfall data was processed using the R programming and clustered using the ensemble distance method. Initially, the rainfall data was segregated into two categories: training data and testing data. The testing data was employed for model evaluation, whereas the training data was utilized for clustering and modeling purposes. Specifically, there were 24 data points designated for testing and 96 data points allocated for training.

Table 2
Monthly rainfall characteristics and Geographical location of the rainfall monitoring stations in West Java

No	Rainfall station	Minimum (mm)	Maximum (mm)	Average (mm)	Longitude	Latitude
1	Bekasi	0	637	137.0339	106.98	-6.28
2	Bojong picung	0	618	195.025	107.27	-6.85
3	Bondan	0	551	142.6122	108.3	-6.61
4	Ciawi	0	976	318.4274	106.85	-6.66
5	Cibeureum	0	706	171.7083	107.49	-7.04
6	Cibukamana	0	836	242.0403	107.51	-6.55
7	Cisalak	0	789	209.2955	106.66	-6.84
8	Cisondari	0	502	144.5958	107.48	-7.09
9	Dempet	0	828	117.0044	108.25	-6.35
10	Depok	0	1344	233.7902	106.76	-6.41
11	Darmaga	8	839	321.1146	106.74	-6.56
12	Emp Agra	0	1544	202.3611	106.99	-7.41
13	Empang	14	810	335.0964	106.8	-6.61
14	Gunung Mas	0	1038	236.7527	106.97	-6.71
15	Indramayu	0	979	139.9946	108.32	-6.34
16	Juntinyuat	0	710	128.0509	108.44	-6.43
17	Kebun raya	0	859	322.2708	106.8	-6.59
18	Krangkeng	0	556	115.8148	108.48	-6.5
19	Leles	0	735	176.9861	107.9	-7.19
20	Losarang	0	700	119.4113	108.15	-6.41
21	Pacet	0	1040	279.0672	107.01	-6.71
22	Pegaden	0	846	166.0507	107.81	-6.55
23	Rajamandala	0	824	201.8718	107.35	-6.84
24	Stageof cemara	0	610	178.3452	107.62	-6.81
25	Sukadana	0	526	135.0625	108.32	-6.55
26	Wanayasa	0	1436	362.0538	107.55	-6.68

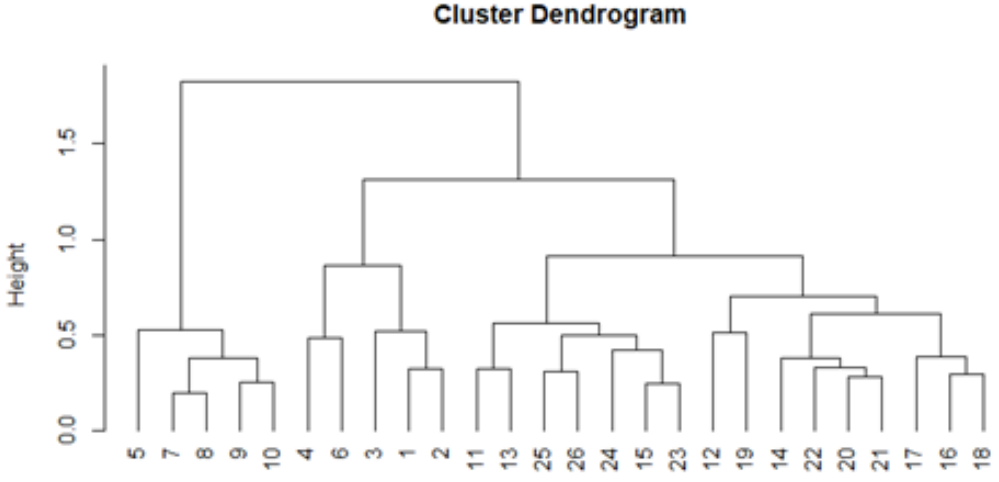


Fig. 1. The dendrogram's cluster of precipitation data based on ensemble distance

The dendrogram that illustrates the clustering process is shown in Fig. 1. The silhouette index approach used to find the optimal number of clusters (Kaufman & Rousseeuw, 1990), was found to be three. The outcomes of the clusters and their respective members are detailed in Table 3. This table illustrates the formation of three distinct clusters, denoted as Cluster A, B, and C. Cluster A consists of 16 member stations, Cluster B comprises five members, and Cluster C includes five members as well. Each cluster is characterized by a prototype, represented by its average value. A visual representation of the regional clustering can be observed in Fig. 2.

Table 3
Clusters and the names of their member rainfall stations

Cluster	Rainfall stations
A	Dramaga, Emp agra, Empang, Gunung mas, Indramayu, Juntinyuat, Kebun Raya, Krangkeng, Leles, Losarang, Pacet, Pegaden, Rajamandala, Stageof Cemara, Sukadana, and Wanayasa
B	Bekasi, Bojong picung, Bondan, Ciawi, and Cibukamana
C	Cibeureum, Cisalak, Cisondari, Dempet, and Depok

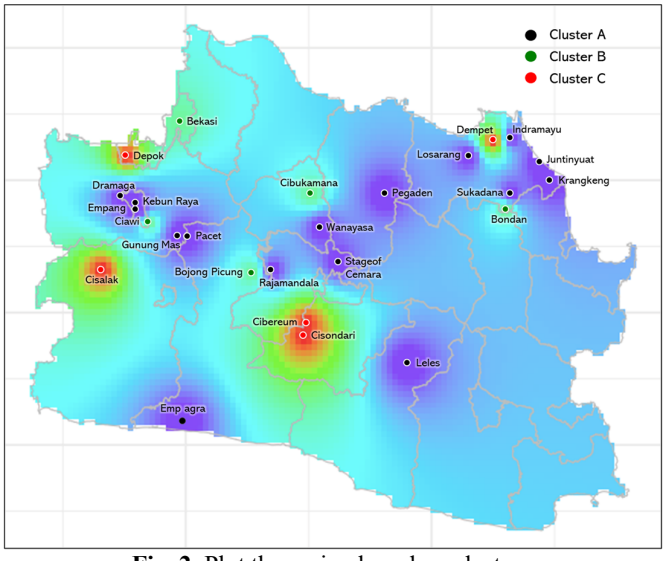
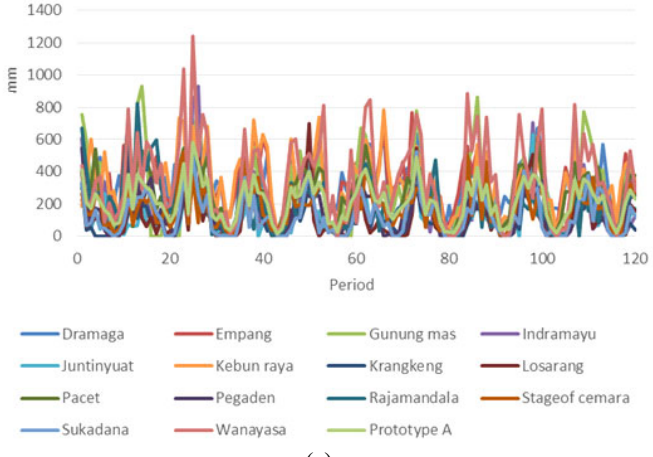


Fig. 2. Plot the region based on clusters.

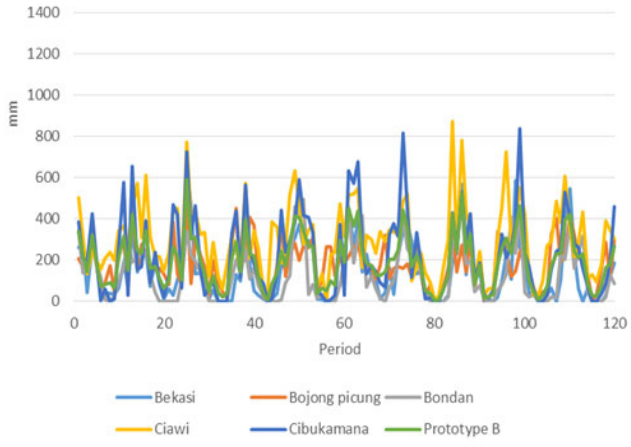
Fig. 3 presents the prototype plot along with the members of each cluster. Within each cluster, there is a noticeable similarity in the pattern of rainfall. Specifically:

- In Cluster A, frequent rainfall events are observed, characterized by very high intensity (> 500 mm), high intensity (300-500 mm), medium intensity (100-300 mm), and low intensity (0-100 mm). This cluster experiences a wide range of rainfall intensities.
- Cluster B exhibits regular rainfall patterns with occurrences of high, medium, and low-intensity rainfall. However, rainfall events with very high intensity are relatively rare in this cluster.
- Cluster C often experiences rainfall events with low, medium, and high intensity, but very high-intensity rainfall is infrequent in this cluster.

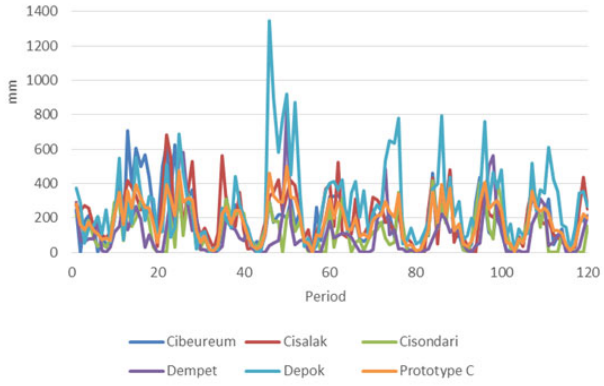
These findings suggest that the clusters are defined by their distinctive rainfall patterns, with varying levels of intensity and frequency.



(a)



(b)



(c)

Fig. 3. Prototype plots and clusters of rainfall data in West Java using the ensemble distance method: cluster A (a), cluster B (b), and cluster C (c)

The cluster models are represented by their respective prototype ARIMA models, as outlined in Table 4. The ARIMA model for cluster A (0, 0, 0) (1, 1, 0) ; cluster B has an ARIMA model (1, 0, 2) (1, 1, 0) ; and cluster C has an ARIMA model (1, 0, 0) (1, 1, 0). These ARIMA models provide a statistical framework for understanding and forecasting the rainfall patterns within each cluster.

Table 4
Cluster models and parameters

Cluster	Model	Structure of Model
A	ARIMA (0,0,0) (1,1,0)	$(1 + 0.5908\, B^{12})(1 - B^{12})\, Y_t = e_t$
B	ARIMA (1,0,2) (1,1,0)	$(1 - 0.7778\ \ B)(1 + 0.5089\, B^{12})$ $(1 - B^{12})\, Y_t = (1 + 0.9940B - 0.3814\, B^2)e_t$
C	ARIMA (1,0,0) (1,1,0)	$(1 - 0.3992\ \ B)(1 + 0.623\, B^{12})$ $(1 - B^{12})\, Y_t = e_t$

By contrasting the RMSE values for models with and without clustering, the clustering results were evaluated. To conduct this comparison, a mean difference test was employed at a significance level (α) of 0.05. The RMSE values were derived from both the model predictions using the training data and the forecasting RMSE calculated from the model predictions using the testing data. The results of this analysis indicate that the p-value for the predicted RMSE is 0.525, and the p-value for the forecasting RMSE is 0.464. All these p-values exceed the significant level of 0.05. Therefore, there is insufficient evidence to reject the null hypothesis (H_0). Stated differently, the model with and without clustering do not differ statistically significantly. This implies that individual models inside a cluster can be accurately represented by the cluster model.

4. Discussion

In this paper, to determine which ARIMA model is the best, several model selection criteria are applied. Table 1 shows the time series data clustering using the Piccolo method with different model selection criteria. The performance evaluation of clustering with different model selection criteria is done by simulation. The simulation results show that the accuracy of

findings obtained from clustering using the AIC, BIC, AICc, RMSE, and MAPE criteria varies. Clustering with BIC or AICc criteria shows better results compared to AIC, RMSE, and MAPE. Clustering with the MAPE criterion shows the lowest accuracy results. The longer the observation period (t), the AIC, BIC, AICc, and RMSE criteria show similar clustering accuracy results, but MAPE shows different results from the others. Based on the simulation, it can be concluded that the accuracy of clustering results using the Piccolo method (Piccolo, 1990; Piccolo, 2010) is influenced by the model selection criteria and the length of the observation period (t). The results are in line with Rahkmawati et al. (2019) where the AIC, BIC, and AICC criteria have similar accuracy patterns. RMSE has a pattern that is quite similar to AIC, BIC, and AICC while MAPE does not have a similar pattern with other criteria.

A parameter ensemble method was created by Hendrawati *et al.*, (2020), and in comparison to the Piccolo(Piccolo, 1990; Piccolo, 2010), this method was able to increase the percentage of clustering accuracy by more than 10%. In this research develops the Piccolo method by using the average distance. This method is called the ensemble distance method. Based on the simulation results shown in Table 1, it is found that the ensemble distance method is better than the Piccolo method. When compared to the Piccolo method, the ensemble distance method is able to increase the clustering accuracy percentage by over 11%. In other words, this method is better than the Piccolo method (Piccolo, 1990; Piccolo, 2010) and the ensemble parameter method (Hendrawati et al., 2020).

This research uses the ARIMA model with model selection criteria AIC, BIC, AICc, RMSE, and MAPE. There are many criteria that can be used to determine the goodness of a model. As a suggestion, the next research needs to try with various models and the latest model selection criteria.

5. Conclusion

This study focuses on the development of a clustering method for time-series data using the ensemble distance approach. The simulation results demonstrate the superiority of the ensemble distance method, which utilizes the average distance of five models, compared to the Piccolo method that relies on a single model. The percentage of clustering accuracy using the ensemble distance method increases as the observation period (t) is extended.

The simulations reveal that the ensemble distance method can improve the percentage of clustering accuracy by more than 11%. In the practical application of monthly rainfall data for the West Java region, it was determined that the optimal number of clusters is three. These clusters exhibit similar rainfall patterns, and cluster models are effective in representing individual models within their respective clusters.

Acknowledgement

The authors would like to express appreciation for the support of the Research Center for Artificial Intelligence and Big Data Unpad and the Directorate for Research and Community Service (DRPM) Ministry of Research, Technology, and Higher Education Indonesia which supports this research under Higher Education Research Grant no. 1549/UN6.3.1/PT.00/2023.

References

Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering - A decade review. *Information Systems*, *53*, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Anderson, D. R., & Burnham, K. P. (2002). Avoiding Pitfalls When Using Information-Theoretic Methods. *The Journal of Wildlife Management*, *66*(3), 912. https://doi.org/10.2307/3803155

Brewer, M. J., Butler, A., & Cooksley, S. L. (2016). The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, *7*(6), 679–692. https://doi.org/10.1111/2041-210X.12541

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Caiado, J., Crato, N., & Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics and Data Analysis*, *50*(10), 2668–2684. https://doi.org/10.1016/j.csda.2005.04.012

Claeskens, G. (2016). Statistical Model Choice. *Annual Review of Statistics and Its Application*, *3*(1), 233–256. https://doi.org/10.1146/annurev-statistics-041715-033413

Corduas, M., & Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. *Computational Statistics and Data Analysis*, *52*(4), 1860–1872. https://doi.org/10.1016/j.csda.2007.06.001

Cryer, J. D., & Chan, K. (2008). *Time Series Analysis with Application in R* (2nd ed.). Springer.

Biabiany, E., Bernard, D.C., Page, V., Paugam-Moisy, H. (2020). Design of an expert distance metric for climate clustering: The case of rainfall in the Lesser Antilles. *Computers & Geosciences*, *145*. 104612. https://doi.org/10.1016/j.cageo.2020.104612

Ergüner Özkoç, E. (2021). Clustering of Time-Series Data. In *Data Mining - Methods, Applications and Systems*. IntechOpen. https://doi.org/10.5772/intechopen.84490

Eszergár-Kiss, D., & Caesar, B. (2017). Definition of user groups applying Ward's method. *Transportation Research*

Procedia, *22*, 25–34. https://doi.org/10.1016/j.trpro.2017.03.004
Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.
Gan, G., Ma, C., & Wu, J. (2007). Data Clustering: Theory, Algorithms, and Applications. In *Data Clustering: Theory, Algorithms, and Applications*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898718348
Gullo, F., Ponti, G., Tagarelli, A., Tradigo, G., & Veltri, P. (2012). A time series approach for clustering mass spectrometry data. *Journal of Computational Science*, *3*(5), 344–355. https://doi.org/10.1016/j.jocs.2011.06.008
Hendrawati, T., Wigena, A. H., Sumertajaya, I. M., & Sartono, B. (2020). A new approach to clustering time series data using the arima model uncertainty. *Communications in Mathematical Biology and Neuroscience*, *2020*, 1–14. https://doi.org/10.28919/cmbn/4778
Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall.
Javed, A., Lee, B. S., & Rizzo, D. M. (2020). A benchmark study on time series clustering. In *arXiv*, 1, p. 100001. https://doi.org/10.1016/j.mlwa.2020.100001
Kalpakis, K., Gada, D., & Puttagunta, V. (2001). Distance measures for effective clustering of ARIMA time-series. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 273–280. https://doi.org/10.1109/icdm.2001.989529
Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data* (L. Kaufman & P. J. Rousseeuw (eds.)). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316801
Keogh, E., & Kasetty, S. (2003). On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. *Data Mining and Knowledge Discovery*, *7*(4), 349–371. https://doi.org/10.1023/A:1024988512476
Kumar, M., & Patel, N. R. (2008). Clustering Data with Measurement Errors. In *Statistical Methods in e-Commerce Research*, 243–267. John Wiley and Sons Inc. https://doi.org/10.1002/9780470315262.ch11
Liao, T. W. (2005). Clustering of time series data - A survey. *Pattern Recognition*, *38*(11), 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025
Maharaj, E. A. (2000). Clusters of time series. *Journal of Classification*, *17*(2), 297–314. https://doi.org/10.1007/s003570000023
Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. John Wiley & Sons.
Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, *31*(3), 274–295. https://doi.org/10.1007/s00357-014-9161-z
Piccolo, D. (1990). a Distance Measure for Classifying Arima Models. *Journal of Time Series Analysis*, *11*(2), 153–164. https://doi.org/10.1111/j.1467-9892.1990.tb00048.x
Piccolo, D. (2010). The Autoregressive metric for comparing time series models. *Statistica*, *70*(4), 459–480. https://doi.org/10.6092/issn.1973-2201/3598
Rani, S., & Sikka, G. (2012). Recent Techniques of Clustering of Time Series Data: A Survey. *International Journal of Computer Applications*, *52*(15), 1–9. https://doi.org/10.5120/8282-1278
Rahkmawati, Y., Sumertajaya, I.M,, Nur Aidi, M. (2019). Evaluation of Accuracy in Identification of ARIMA Models Based on Model Selection Criteria for Inflation Forecasting with the TSClust Approach. *Int J Sci Res Publ*. 9(9):p9355. doi:10.29322/ijsrp.9.09.2019.p9355.
Triacca, U. (2016). Measuring the distance between sets of ARMA models. *Econometrics*, *4*(3). https://doi.org/10.3390/econometrics4030032
Wei, WW. (2006). *Time Series Analysis: univariate and multivariate methods*. second. Boston: Pearson Addison Wesley.