Performance Evaluation of AIC and BIC in Time Series Clustering with Piccolo Method

Triyani Hendrawati[1,2], Aji Hamim Wigena[2], I Made Sumertajaya[2], Bagus Sartono[2]

{triyani.hendrawati@gmail.com,ajiwigena@ymail.com, imsjaya@gmail.com}

Department of Statistics Padjadjaran University, Bandung, Indonesia[1], Department of Statistics IPB University (Bogor Agricultural University), Bogor, Indonesia[2]

Abstract. Piccolo method use parameters of Autoregressive model tocluster time series data. One set of time series data can produce several model, but only one model is used for clustering. Akaike's Information Criterion (AIC) or Bayesian information Criterion (BIC) can be used to selection model. But if it is used different criterion to selection model, can be produced different model, so it can cause different cluster. The aim of this research is to evaluate performance of AIC and BIC in time series clustering with Piccolo method. The simulation comparing performance of AIC with BIC in time series clustering using the Piccolo method was carried out. Results shows that Bayesian information Criterion (BIC) is better than Akaike's information Criterion (AIC).

Keywords: AIC, AR metric, BIC, Piccolo distance, time series clustering.

1 Introduction

Time series data is dynamic data which naturally changes as a function of time, where the value of each point is an observation that occurs sequentially [1]. Methods for clustering of time series data are developed by modifying the algorithm for conventional cluster, so it can handle time series data. Other method is change time series data into ordinary data form (not time series), so it can use the algorithm for ordinary data cluster [8]. Aghabozorgi *et al.* [1], Liao [8], Rani and Sikka [12] have reviewed studies on clustering of time series data.

Sometimes time series data set have unequal length of observations period, so Euclidian distance can't be used for measure distance, as alternate you can use the model-based approach. One method of model-based clustering is Autoregressive (AR (*p*)) model. Several studies on Autoregressive (AR (*p*)) metrics are Piccolo [10], [11], Maharaj [9], Kalpakis *et al.* [6], Caiado*et al.* [2], Corduas and Piccolo [3], Iannario and Piccolo [5], and Triacca [13].

Piccolo [10], Corduas and Piccolo [3], and Piccolo [11] use Autoregressive (AR (*p*)) model for clustering time series data. From a set of time series data can be obtained several Autoregressive (AR (*p*)) models, but only one model is selected based on certain

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290340

criteria. There are several criteria for models selection namely Akaike's information Criterion (AIC) and Bayesian information Criterion (BIC). The model obtained may be different, it is depends on the criteria is used. For example, the first person makes the best model is model 1 based on the AIC criteria, the second person makes the best model is model 2 based on BIC criteria, and so on, which true model is never known. In this paper, we evaluate performance of AIC and BIC in time series clustering with Piccolo method. We use simulation with generating data set.

2 Materials

2.1 Analysis of Time Series Data Clusters

In ordinary data clustering, the distance between objects is actually calculated precisely, but in time series data clustering, the distance is calculated by approximate [1]. One of the easiest ways to calculate distances between two time series is to calculate distances at all time points. However, not all time series data have the same length of observation period, so this method cannot be done.

There are three kinds of approaches for time series data clusters namely raw-data-based, feature-based, and model-based [8]. The usual approach is to directly use data rows. This approach is also called the raw data based approach because it usually works directly with time series raw data. Raw data-based algorithms usually use the usual clustering method, and the distance used is Euclid's distance. The time series raw data is converted into a lower dimension feature trait, this is done in a feature based approach. After that the conventional clustering algorithm can be applied to the extracted feature vector.If we use a model-based approach, the time series data is converted into parameters model. Then parameters model are used to measure distance between two models.

2.2 Autoregressive Model (AR (*p*))

The Autoregressiveprocessis a regression on themselves. The current value of the series Y_t is a linear combination of the p most recent past values plus random components. The Autoregressive model (AR (p)) is:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \quad \ldots (1)$$

where:
$\phi_p \neq 0$ and$\sigma_e^2 > 0$; e_t is random components that are mutually independent, with assumptions $e_t \sim N(0,\sigma_e^2)$, p is the Autoregressive order .

The method to estimate model parameters use maximum likelihood, Bayes, or least square method. The criteria for goodness of the ARIMA model can be measured by AIC, with formula:

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

While the formula for criteria BIC is:

$$BIC = -2 \log(\hat{\sigma}_e^2) + k \log(n) \quad \ldots (3)$$

where:n is the number of observations, k is the number of parameters, $\hat{\sigma}_e^2$ is the maximum likelihood estimator for the variance of the error [4].
The first term formula of BIC is the same as AIC, it only differs in the second term. In the second term, it can be seen that the BIC value is influenced by the number of sample sizes and parameters in the model.

2.3 Autoregressive (AR) Distance

Autoregressive (AR) metric is measure the structural similarity between two invertible ARIMA process [10], [3], [11]. Piccolo [10] was introduce the measurement of structural similarities between two ARIMA process. $X_t$and$Y_t$can be expressed as coefficients $\hat{\pi}$ on AR (∞) namely$\hat{\boldsymbol{\pi}}_x = (\hat{\pi}_{1,x}, \hat{\pi}_{2,x}, \ldots, \hat{\pi}_{j,x}, \ldots)$; $\hat{\boldsymbol{\pi}}_y = (\hat{\pi}_{1,y}, \hat{\pi}_{2,y}, \ldots, \hat{\pi}_{j,y}, \ldots)$. The distance between two ARIMA process X_t and Y_t is expressed as a Euclidean distance between coefficients $\hat{\pi}$ on AR (∞) with the formula:

$$d(x,y) = \sqrt{\sum_{j=1}^{\infty} \left(\hat{\pi}_{j,x} - \hat{\pi}_{j,y}\right)^2} \quad \ldots (4)$$

In practice AR (∞) can be approached with AR (p) [16].

3 Simulation

In this section we use generating data for simulation. We generating data for 3 cluster namely cluster 1(model a), cluster 2 (model b), and cluster 3 (model c).
Model a: AR (2) with parameters (ϕ_1, ϕ_2) = (0.2, 0.1)
Model b: AR (2) with parameters (ϕ_1, ϕ_2) = (0.4, 0.5)
Model c: AR (2) with parameters (ϕ_1, ϕ_2) = (0.6, 0.2)
Each cluster is consist 10 generating data series, giving a total of thirty time series [7].

We have two scenario for generating data, that is scenario 1: model a, b, and c was generating with periods of observation (n) are 50, 300, and 1000 time points. White noise was chosen to be from Gaussian distribution N (0, 0.1). Scenario 2: model a, b, and c was generating with periods of observation (n) are 50, 300, and 1000 time points. White noise was chosen to be from Gaussian distribution N (0, 1).

Time series data from generating data is stated in several Autoregressive (AR (∞)), in this study we use model approach (AR (p)), with maximum p is 5. Then selection model with AIC and BIC, choose best model with minimum value. Then calculate the distance between Autoregressive (AR (p)) models. The coefficients parameters of model AR (p) are used for calculate distance between two models with Euclidean distance. After matric of distance is formed, then we clustering with Ward method.

We simulated 1000 times for periods of observation (n) are 50, 300, and 1000 time points. Average accuracy of clustering with selection model AIC and BIC was calculated. The results of clustering was evaluated. The results of clustering is true if generating data series which come from one cluster is groped to the same cluster. The results of simulation can be seen at Table 1.

Table 1. The average accuracy of clustering with selection model AIC and BIC.

Scenario	n	Selection model with AIC criterion	Selection model with BIC criterion
1	50	77.61%	82.76%
	300	99.02%	99.17%
	1000	99.99%	100%
2	50	76.97%	82.21%
	300	98.95%	99.15%
	1000	99.99%	99.99%

Base on Table 1, scenario 1 with period of observation (n) is 50 observation, the average accuracy of clustering with selection model AIC and BIC is very different, that is 77.61% and 82.76%. It also happened in scenario 2, when period of observation (n) is 50 observation, the average accuracy of clustering with selection model AIC is 76.97% and BIC is 82.21%.For both scenario, if period of observation (n) is longer, average accuracy of clustering with selection model AIC and BIC is almost similar. Average accuracy of clustering is rise when period of observation (n) is longer. However the percentage of accuracy of cluster with BIC is better than AIC for all scenario.

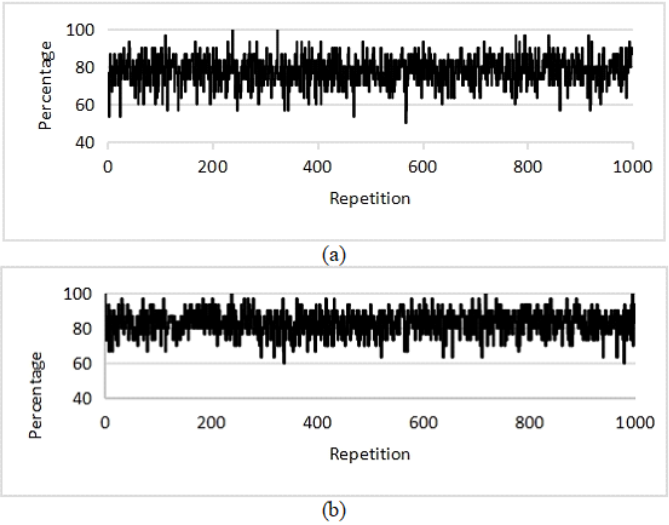

Fig. 1. Percentage of accuracy of clustering results with the observation period (n) 50, (a) choosing model using the AIC criteria, (b) choosing model using the BIC criteria.

In Figure 1, it appears that inscenario 1 when the length of the observation period (n) is 50 time points, obtained the percentage of accuracy in selection model using BIC criteria has a narrower interval than using AIC criteria. The accuracy of the results using AIC criteria is ranged from 53.3% - 100%, while BIC criteria ranges from 60% - 100%. This shows that BIC criteria are more stable compared to AIC.

4 Application

As a real data application, a monthly rainfall intensity data clustering is conducted. Data was obtained from BadanMeteorologiKlimatologidanGeofisika(BMKG). The monthly rainfall intensity data is taken from 49 rainfall stations in the West Java region. The period of rainfall data used is from 1981 to 2014. Each rainfall station has different observation periode, it is caused data availability. The data used is complete data without missing value.

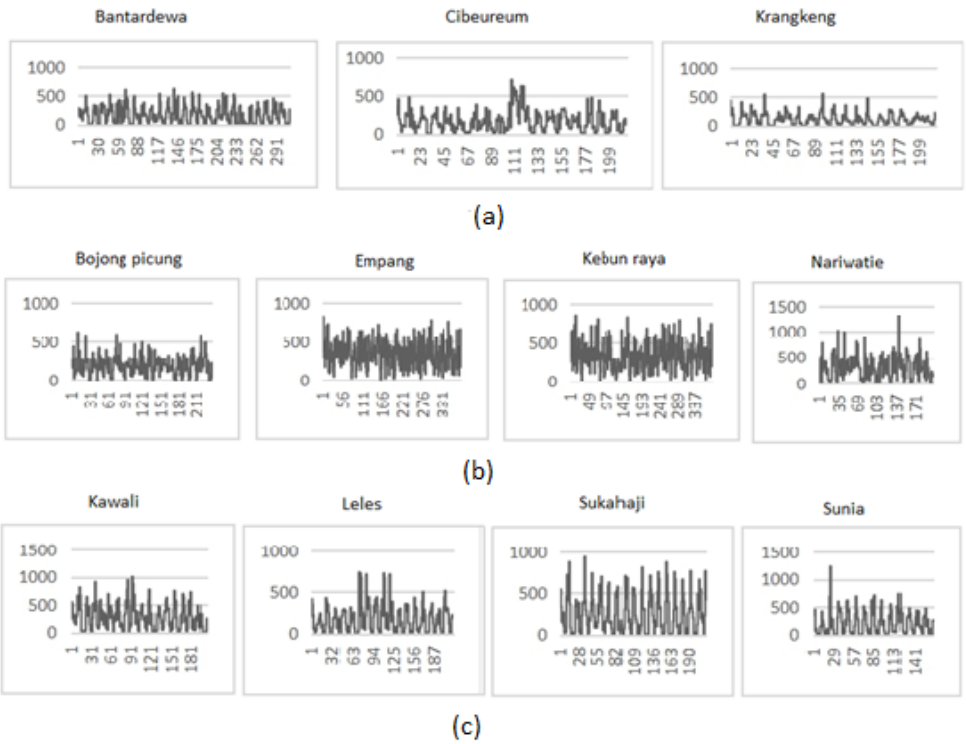

Fig. 2. Plot of rainfall data for each cluster, (a) cluster 2, (b) cluster 4, (c) cluster 6.

The results of the clustering of monthly rainfall intensity data from 49 rainfall stations in the West Java region based on the smallest BIC value were obtained optimal results for six cluster. The clusters are as follows:

- Cluster1:Banjaran, Bekasi, Bondan, Ciawi, Cibuni, Cimalaka, Depok, Gunung Mas, Indramayu, Jatiseeng, Losarang, Lwgede, Pacet, Pegaden, Purwakarta, Salam darma, Stageofcemara, Sukadana, and Wanayasa.
- Cluster2:Bantardewa, Cibeureum, and Krangkeng.
- Cluster3: Barugbug, Cibukamana, Dangdeur, Darmaga, Jatiwangi, Setupatokselatan, Subang, and Tukmudal.
- Cluster4:Bojongpicung, Empang, Kebunraya, and Nariwatie,
- Cluster5: Cinangling, Cisalak, Cisondari, Dempet, Emp Agra, Gegesik, Juntinyuat, Karangkendel, Montaya, Rajamandala, and Singaparna,
- Cluster6:Kawali, Leles, Sukahaji, and Sunia.

In Figure 2, it can be seen rainfall data plots based on clusters for cluster 2, cluster 4, and cluster 6. Due to space limitations, only a portion of the cluster results are presented. In one cluster there is a similarity in rainfall patterns, whereas different groups have different rainfall patterns.

5 Conclusion

Based on the simulation conducted, there are differences in the accuracy of the results of the clustering using the Akaike's information Criterion (AIC) and Bayesian information Criterion (BIC) value criteria. BIC criteria are more stable compared to AIC [17]. BIC give better results than the AIC criteria for all scenario. The results of the clustering of rainfall stations in the West Java region based on the smallest BIC value were obtained by six clusters. Each group shows a similar pattern of rainfall.

References

[1]Aghabozorgi S, Shirkhorshidi AS, Wah TY. 2015. Time-series clustering – A decade review. *Information Systems* 53. 16–38. Elsevier Ltd.

[2] Caiado J, Crato N, Pena D. 2006. A Periodogram-based metric for time series classification. *Computational Statistic & Data Analysis* 50. 2668-2684.

[3] Corduas M, Piccolo D. 2008. Time series clustering and classification by the autoregressive metric. *Comput. Stat. Data Anal*. 52 (4) 1860–1872.

[4] Cryer JD, Chan K. 2008. *Time series analysis with Application in R second edition*, Springer, USA.

[5] Iannario M, Piccolo D. 2013. Spectral decomposition of the AR metric, *Advances in Theoretical and Applied Statistic*, p. 109-118, Springer-Verlag, Berlin, Heidelberg.

[6] Kalpakis K, Gada D, Puttagunta V. 2001. Distance Measures for Effective Clustering of ARIMA Time-Series. *In Proceedings of the IEEE International Conference on Data Mining*, p 273-280, San Jose, CA, USA.

[7] Kumar M, Patel NR. 2005. Clustering Data with Measurement Errors, *Rutcor Research Report*. Rutgers University, New Jersey.

[8] Liao TW. 2005. Clustering of time series data-a survey. *Pattern Recognition* 38. 1857 – 1874. Elsevier Ltd.

[9] Maharaj EA. 2000. Clusters of time Series. *Journal of Classification* 17: 297 - 314. DOI: 10.1007/s003570000023.
[10] Piccolo D. 1990. A distance measure for classifying ARIMA models. *Journal of time series analysis*. vol.11 No.2.
[11] Piccolo D. 2010. The Autoregressive metric for computing time series models. *Statistica*. LXX.n.4.
[12] Rani, Sikka. 2012. Recent Techniques of Clustering of Time Series Data: A Survey. *International Journal of Computer Applications* (0975 – 8887). Vol. 52 No.15.
[13] Triacca U. 2016. Measuring the distance between set of ARMA models. *Econometrics*. 4. 32.
[14] Wei W. 2006. *Time Series Analysis: univariate and multivariate methodssecond edition*. Pearson education inc., USA.
[15] Somov, A. 2011. Wildfire safety with wireless sensor networks. EAI Endorsed Transactions on Ambient Systems. pp. 1-11.
[16] Montero P and Vilar J.A., 2014. " TSclust : An R package for time series clustering", in Journal of Statistical Software, 62(1).
[17] Ding J, Tarokh V, Yang Y. 2018. Bridging AIC and BIC: A new criterion for autoregression, IEEE Transactions on Information Theory, vol. 64, no. 6.