Research Article

Received: date:12.04.2023
Accepted: date:15.06.2023
Published: date:30.06.2023

Examination of Pre-Founded Time-Series Models for Classic Canada Lynx Data by K-Means Cluster Method and BDS Test

Reşat Kasap [1*]

[1] Gazi University, Faculty of Science, Department of Statistics, Ankara, Türkiye; rkasap@gazi.edu.tr
Orcid: 0000-0002-9306-3101[1]
*Correspondence: rkasap@gazi.edu.tr

Abstract: Canadian lynx data are widely used and modeled in the literature. Although many different models have been made so far, no model-based classification studies have been carried out in terms of residuals to investigate the similarities or differences between these models. This study reviewed previously obtained models for the Canadian lynx time series. The starting point of the current study is residuals, and some statistical data analysis tools are used for this. The residual series of the models are clustered with the K-means cluster method. Besides, a new model is proposed for this time series, and the model is included in the data analysis together with other models in the literature. In addition, chaos analysis was performed for all residual time series of the models.

Keywords: Time series, the K-means cluster method, BDS Test Statistic, Canadian lynx data

1. Introduction

There are some classical data patterns in the literature. A new methodology is first applied to these datasets. Time series analysis also includes such data, one of them which is Canada lynx time series. The Canadian lynx data investigated in time series analysis was collected annually for the period 1821–1934. It gives the number of the Canadian lynx "trapped" in the North-West Canada. The review and classification of previous models for the Canadian lynx time series is the research topic of our study. Known time series modeling methods such as the Box-Jenkins or Autoregressive Integrated Moving Average (ARIMA) method can be used to study such time series. However, the stationarity of the ARIMA model time series is limited by the normality of residuals and the requirement for independence. The residuals are the errors between the observed time series and the model constructed with ARIMA, which should be uncorrelated and normally distributed [1].

The Canadian lynx time series includes a ten-year cycle [2-5]. The data have been examined in many studies in the literature. Some studies are just linear, bilinear, exponential, etc. It includes classical modeling approaches such as Linear model forms that have been proposed by [4], [6], and [7]. [8] examined the exponential model structure. Self-excited Threshold Autoregressive (SETAR) models were given by [9], while bilinear models were proposed by [10] and [7]. These studies aim to examine the Canadian data as classic data in the literature with modeling techniques. Thus, the different appearance of this classical data in different models can be seen.

Lai investigated some of the linear and nonlinear autoregressive models on the classic Canadian lynx data [11]. Kaboudan used this data as a real-world data application to evolve best-fit regression models [12]. Teruia and Van Dijk applied Canadian Lynx data to the time-varying method, which allows for a locally (non)linear modelling [13]. Khashei and Bijari used Canadian Lynx data to apply hybrid models for time series forecasting [14]. The usefulness of the Threshold Quantile Autoregression (TQAR) model is illustrated in an application to the dynamics of the Candian lynx population by [15]. Zainuddin et al.

Citation: R.Kasap , " Examination of Pre-Founded Time-Series Models for Classic Canada Lynx Data by K-Means Cluster Method and BDS Test", Journal of Statistics and Applied Sciences, no. 6, pp. 18-25, Dec. 2022, doi:10.52693/jsas.1195048

forecasted Canadian lynx data by using a novel hybridization of bootstrap and double bootstrap artificial neural networks [16]. A forecasting approach for Canadian Lynx data based on machine learning techniques was applied by [17]. Chen et al. proposed a novel method based on error decomposition and a nonlinear combination of forecasters and used this method for forecasting Canadian Lynx data as a real-world data application [18].

In the concept of the current study, it is aimed to examine studies that are based on non-hybrid modeling techniques. Also, the selected time series models which were previously modeled with non-hybrid modeling techniques by 12 different methodologies or researchers are examined by taking into account the modeling in [19]. The examination in question is made on the remnants of each model. The examination in question is made on the residuals of each model. The main purpose of the study is to cluster the models using the K-means clustering method and using the BDS test statistic to investigate chaos. The current study is structured as follows: Section 2 gives previously obtained time series models by using Canadian lynx data, Section 3 includes the methods and results for data mining analysis of the models mentioned in Section 2, and the final comments of the study are given in Section 4.

2. Materials and Methods

2.1. Obtained Time Series Models by Using Canadian Lynx Data

Canadian lynx series, whose models are the subject of investigation, was determined annually between 1820-1934. This data is one of the most popular data sets for theoretical or applied studies in time series analysis. The sequence graph of the said time series is given in Fig. 1.

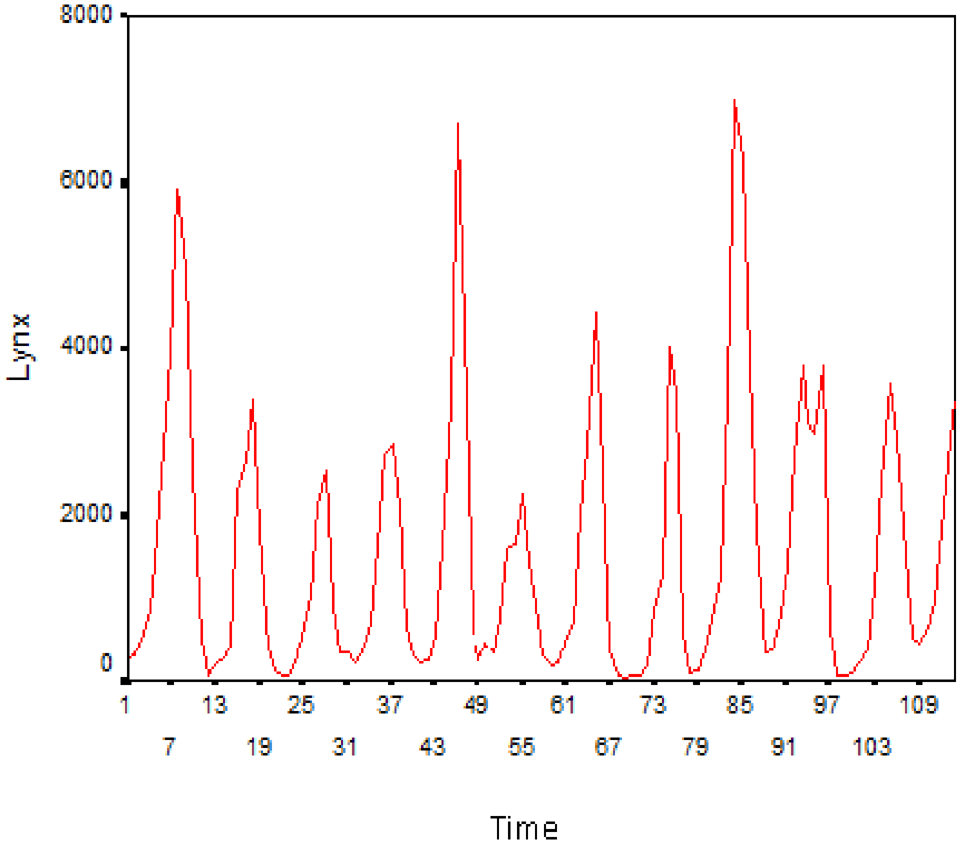

Figure 1. Canadian Lynx series (1820-1934)

The models related to Canadian lynx data are usually obtained after performing logarithmic transformation for a 100-unit slice. The models below assume a white noise process with a mean ε_t of 0 and a variance of σ^2. The models obtained at various times of the data mentioned above, sometimes using different methodologies, are given below:

Model I – Linear [4]

$$X_t = 2,9 + 1,41X_{t-1} – 0,77X_{t-2} + \varepsilon_t, \tag{1}$$

Model II - Exponential [8]

$$X_T= \{1,167 + (0,316 + 0,982X_{T-1})K_T\}X_{T-1} – \{0,437 + (659 +1,26X_{t-1}) K_t\}X_{t-2} + E_t\,, \quad K_t = \exp(-3,89X_{t-1}^2). \tag{2}$$

Model III - SETAR (2,2,2) [9]

$$X_t = \left\{ \begin{array}{ll} 0,62 + 1,25X_{t-1} – 0,43X_{t-2} + 0,195\varepsilon_t, & X_{t-2} \leq 3,25 \\ 2,25 + 1,52X_{t-2} – 1,24X_{t-2} + 0,25\varepsilon_t, & X_{t-2} > 3,25 \end{array} \right. \tag{3}$$

Model IV - Bilinear (9) [10]

$$X_t – 0,8845X_{t-1} + 0,1699X_{t-2} + 0,1271X_{t-4} – 0,5514X_{t-10} + 0,5280X_{t-11} = 1,117 – 0,1653X_{t-8}\, \varepsilon_{t-10} – 0,097X_{t-5}\, \varepsilon_{t-8} + 0,0922X_{t-1}\, \varepsilon_{t-1} + \varepsilon_t\,, \tag{4}$$

Model V - AR (11) [6]

$$X_t = 1,13X_{t-1} – 0,51X_{t-2} + 0,23X_{t-3} – 0,29X_{t-4} + 0,14X_{t-5} – 0,14X_{t-6} + 0,08X_{t-7} - 0,04X_{t-8} + 0,13X_{t-9} + 0,19X_{t-10} – 0,13X_{t-11} + \varepsilon_t\,, \tag{5}$$

Model VI [6]- AR (11)

$$X_t = 1,0938X_{t-1} - 0,3571X_{t-2} - 0,1265X_{t-4} + 0,3244X_{t-10} – 0,3622X_{t-11} + \varepsilon_t\,, \tag{6}$$

Model VII [7]- Bilinear (13)

$$X_t - 0,77227X_{t-1} + 0,091572X_{t-2} – 0,083073X_{t-3} + 0,261493X_{t-4} - 0,225585X_{t-9} + 0,245841X_{t-12} = 1,486292 – 0,7893X_{t-3}\, \varepsilon_{t-9} + 0,4798X_{t-9}\, \varepsilon_{t-9} + 0,3902X_{t-6}\, \varepsilon_{t-2} + 0,1326X_{t-1}\, \varepsilon_{t-1} + 0,07944X_{t-2}\, \varepsilon_{t-7} - 0,3212X_{t-4}\, \varepsilon_{t-2} + \varepsilon_t\,, \tag{7}$$

Model VIII [7]- Linear

$$X_t - 1,0541X_{t-1} + 0,4538X_{t-2} – 0,32597X_{t-3} + 0,37912X_{t-4} – 0,23452X_{t-5} + 0,1757X_{t-6} – 0,09598X_{t-7} + 0,12843X_{t-8} – 0,27435X_{t-9} – 0,11427X_{t-10} + 0,18534X_{t-11}+0,17218X_{t-12}= \varepsilon_t\,, \tag{8}$$

Model IX [7]- Linear

$$X_t - 1,01705X_{t-1} + 0,39997X_{t-2} – 0,25851X_{t-3} + 0,22037X_{t-4} – 0,21099X_{t-9} + 0,25343X_{t-12} = \varepsilon_t\,, \tag{9}$$

Model X [9](Tong, 1983)- SETAR (2,5,2)

$$X_t = \left\{ \begin{array}{ll} 0,768 + 1,067X_{t-1} - 0,2X_{t-2} + 0,164X_{t-3} – 0,428X_{t-4} + 0,181X_{t-5} & \\ \quad + 0,174\varepsilon_t, & X_{t-2} \leq 3,05 \\ 2,254 + 1,474X_{t-1} – 1,202X_{t-2} + 0,238\varepsilon_t\,, & X_{t-2} > 3,05 \end{array} \right. \tag{10}$$

Model XI [9]- SETAR (2,7,2)

$$X_t = \begin{cases} 0{,}546 + 1{,}032X_{t-1} - 0{,}173X_{t-2} + 0{,}171X_{t-3} - 0{,}431X_{t-4} + 0{,}332X_{t-5} & \\ - 0{,}284X_{t-6} + 0{,}210X_{t-7} + 0{,}161\varepsilon_t, & X_{t-2} \leq 3{,}116 \\ 2{,}632 + 1{,}492X_{t-1} - 1{,}324X_{t-2} + 0{,}225\varepsilon_t, & X_{t-2} > 3{,}116 \end{cases} \quad (11)$$

Model XII [8]- Exponential

$$X_t = 0{,}481X_{t-1} - 0{,}247X_{t-2} + 0{,}318X_{t-3} + 0{,}23X_{t-4} + 0{,}352X_{t-5} \quad (12)$$

$$+ 0{,}096X_{t-6} - 0{,}085X_{t-7} - 0{,}289X_{t-8} - 0{,}181X_{t-9} + Y_t,$$

$$Y_t = \{1{,}514 + (0{,}480 - 3{,}332Y_{t-1} - 0{,}610Y_{t-1}^2 + 8{,}906Y_{t-1}^3)$$

$$\exp(-\gamma Y_{t-1}^2)\}Y_{t-1} + \{- 0{,}902 + (- 0{,}228 + 0{,}923Y_{t-1} + 0{,}193Y_{t-1}^2$$

$$- 4{,}216Y_{t-1}^3)\exp(-\gamma Y_{t-1}^2)\}Y_{t-2} + \varepsilon_t, \ \gamma = 3{,}89.$$

In addition to the above models, it can be written by remodeling the original data in question. When the model obtained by remodeling is modeled by taking the logarithm of the original data, the proposed model is named Model XIII-linear and expressed as follows, $X_t = \log Z_t$

$$X_t = 6.645 + 1.347X_{t-1} + 0.657X_{t-2} + \varepsilon_t + 0.294\varepsilon_{t-10} \quad (13)$$

is obtained.

The data mining analysis based on the statistical methodology used to investigate the time series models obtained by modeling the Canadian lynx data can be summarized as follows:

i- Clustering the models' residual sequences with the help of the K-means cluster method.
ii- A criteria for identification of non-linear (or chaos), the BDS (Brock, Dechert, Scheinkman) test statistic.

Cluster analysis is a framework that develops tools and methods that group large individuals (or cases, units, items, objects, etc.) with the help of a specific data matrix containing multiple variables. Items are grouped or clustered using metrics based on “similarity” [20]. The similarity measure is calculated using a distance function based on an approximation such as Euclidean, Mahalanobis, or Manhattan. Items with small distance function values are classified under the same cluster [21]. For this data, clustering with hierarchical and principal components has been applied before in other studies [29]. In this study, one of the most popular is the cluster analysis method to determine the similarities of the models in the literature: K-means cluster method.

2.2. K-Means Cluster Method

This method can be defined as a non-hierarchical clustering method, which is proposed by [22]. The K-means cluster method uses an algorithm to assign each unit to the group or cluster having the nearest centroid [21]. This method tries to construct homogeneous groups or clusters of items based on selected variables, using an algorithm that can handle a large number of cases [20]. In the method, it is necessary to determine the number of clusters at the beginning. This information can specify initial cluster centres [23-24]. The algorithm of K-means can be explained as follows: Firstly, the optimal K value (the number of groups or clusters) has to be determined. Then the K number of items is randomly chosen, and they are assigned for each cluster separately. Afterwards, the remaining items are assigned to the relevant clusters by considering the minimum distance function value, and the new cluster centers are calculated in each iteration. It is needed to repeat the steps above until no more reassignments for items take place [21].

2.3. The BDS Test Statistic

The BDS has been used as a test statistic, as a criterion to identify the chaos (or non-linear), which tests the null hypothesis that the variable of interest is Independently and Identically Distributed (IID). Now, let us briefly consider the test statistic-BDS itself. It is based on the so-called correlation integral introduced by [25].

The time series to be analyzed (X_t:1,2,..., T) is used to form the N-histories

$$X_t^N = (X_t, X_{t+1}, \ldots, X_{t+N-1}).$$

Each N-history can be considered to be a point in an N-dimensional space, where N is called the embedding dimension. These N-histories can be used to define a correlation integral

$$C_N(e) = \frac{2}{T_N\,(T_N - 1)} \sum\nolimits_{t<s} \sum I_e(X_t^N X_s^N),$$

where T_N = T-N+1, and I_e is the indicator function of the event

$| X_{t+i} - X_{s+i} | < e$, i=0,1,...,N-1.

The correlation integral, $C_N(e)$, can be interpreted as an estimate of the probability that X_t^N and X_s^N are within a distance e. Given this interpretation, we can see that under the independence hypothesis

$$C_N(e) \to C_1(e)^N, \text{ as } T\to\infty$$

holds. That is, $P(| x_{t+i} - x_{s+i} | < e)$, (i=0,1,..., N-1) is, due to independence, equal to $\prod_{i=1}^{N-1} P(\,| X_{t+i} - X_{s+i} |< e)$, which is estimated by $C_1(e)^N$ as the variables are identically distributed [26-27]. Thus, the BDS statistic reduces to

$$W_N(e) = [\sqrt{T}\ (C_N(e) - C_1(e)^N)] \,/\, \hat{\sigma}_N(e)\,,$$

where $\hat{\sigma}_N(e)$ is an estimate of the standard deviation under the null hypothesis. WN(e) distribution converges to a standard normal with expectation zero and variance unity as T approaches infinity. Thus, one can now calculate the statistic that has a standard normal asymptotic distribution under the independence hypothesis. If the absolute values of the test statistic are large, the null hypothesis of IID (randomness) is to be rejected [28].

3. Results of Analysis

Some of the results obtained from the analyzes using the methods described above are given in the text. As already mentioned, the study of models of the Canadian lynx series is based on residuals. All other results regarding the techniques are given in Appendices A and B. Accordingly, some descriptive statistics of the model residuals for 13 items (models) are given in Table 1.

Table 1. Some Values for Model Residuals

Model name-type	Average	Standard deviation	Goodness of fit-p
Model I-Linear	-0.01022	0.238354	0.520603
Model II- Exp.	-0.00525	0.212661	0.105848
Model III-SETAR	-0.10415	0.260800	0.228786
Model IV-Bilinear	0.00188	0.202352	0.620368
Model V-Linear	-0.00703	0.197697	0.400336
Model VI-Linear	-0.01299	0.202263	0.681150
Model VII-Bilinear	-0.00644	0.152882	0.478163
Model VIII-Linear	0.00379	0.183809	0.535887
Model IX- Linear	-0.00410	0.192566	0.967262
Model X-SETAR	0.00605	0.204812	0.672208
Model XI-SETAR	0.00400	0.199168	0.724205
Model XII- Exp.	-0.01952	0.242870	0.054820
Model XIII- Linear	-0.00703	0.539036	0.035021

The p-values of the mean, standard deviation, and normality tests for all model residuals are given in Table 1. Model II, Model XII, and Model XIII, in particular, have the lowest probabilities when the p-values are examined carefully. What they have in common is that the two are exponential.

Using the hierarchical clustering with Dendrogram and principal component analysis method, [29] determined that the models were divided into 4 clusters in both of two methods: {I, II, III}, {IV, V, VI}, {VII, VIII, IX}, and {X, XI, XII}.

According to the K-means cluster method, when considered as two clusters (optimal value according to Silhouette approach), {I, II, III, IV, V, VI, VII, VIII, IX, X, XI, XII} and {XIII} are obtained. Similar to the results of hierarchical clustering, models I, II, and XII stay in the same cluster. In addition, the other models except model XIII are in the same cluster as models I, III, and XII. When the number of clusters value increases, even IV has a different structure than V and VI, and {IV, VI, VII} construct a cluster together. If the number of clusters value is four, the obtained clusters are as follows: {I, III}, {IV, V, VI},{II, VII, VIII, IX, X, XI, XII}, and {XIII}. Similar to hierarchical clustering results, the models with different mathematical structures can stay in the same clusters. The proposed Model M also forms a cluster by itself in K-means. All the results for the K-means cluster method can be seen in Appendix A.

Lai compared the Canadian lynx data of AR and some other five models with BDS statistics [11]. In this study, the residuals of thirteen models obtained for Lynx were examined using BDS statistics. The BDS results for all the models are given in Appendix B below.

Considering the BDS test results of the residuals series given in Appendix B, it can be said that the assumptions are met for Model I, Model II, Model III, Model VII, Model X, and Model XI; that is, the residuals are linear. The other models, the residuals for Model IV and Model XII, were $p< \alpha=0.05$ in all dimensions and failed the BDS test. Also, the performance of BDS Statistic in Model III (dimension 3), Model V (dimensions other than 2), Model VI (dimensions other than 2), Model VIII (dimensions 5 and 6), and Model IX (dimensions 5 and 6) could not pass. That is, it contains a non-linear (or chaos) structure.

The analysis results show that the models obtained for the Canadian lynx series, considered classical data in its field, do not show a complete similarity with each other. In the previous studies, the groups obtained in Hierarchical clustering and principal component analysis formed some different groups from those obtained by K-means. Although all models examine the model in the same time series, contrary to expectations, the models show differences. Another interesting point is that the models with the same mathematical structure tend to fall into different groups or clusters.

4. Conclusions

In the current study, the classical Canadian lynx data, which is one of the widely used time series, has been examined with the BDS test statistics of the residuals of these models by classifying the models obtained in various studies before. The models' similarities and differences were investigated using the K-means cluster method, one of the data mining techniques based on multivariate statistical methods. The most important points in the current working concept can be highlighted as follows: All the models in the current study show some differences from the clusters obtained by the previously used methods. The K-means cluster method yields similar results in some respects, as it is based on a distance matrix calculated from the data. If desired, different time series clustering methods can be used in the future [30]. In addition, according to the results of the BDS test statistics based on the residuals of the models, it can be said that the residuals of some models do not pass the BDS test; that is, they contain a nonlinear (or chaotic) structure.

Author Contributions: All authors contributed equally to this study.

Funding: This research did not receive external funding.

Conflicts of interest: The authors declare no conflict of interest.

Acknowledgements: The authors would like to thank Gazi University Academic Writing Application and Research Center for proofreading the article.

References

[1] Box, G.E.P., Time Series Analysis: Forecasting and Control 1994, Englewood Cliffs, N.J. Prentice Hall.

[2] Cendejas-Zarelli, S., "Annual Canadian Lynx trappings 1821-1934", https://rstudio-pubs-static.s3.amazonaws.com/168257_373a84b37 f48453dad40cbc708f 670ff.html, 2016.

[3] Karnaboopathy, R. and Venkatesan, D., "Data mining in canadian lynx time series", Journal of Reliability and Statistical Studies; ISSN: 0974-8024, (Online):2229-5666, 2012; 5(1): 1-6.

[4] Moran, P.A.P., "The statistical analysis of the Canadian lynx", Australian Journal of Zoology, 1953; 1: 163-173.

[5] Tong, H. and Dabas, P., "Cluster of time series models: an example", Journal of Applied Statistics, 1990; 17(2): 187-198.

[6] Tong, H., "Some comments on the Canadian lynx-with data-with discussion", Journal of Royal Statictical Society, A, 1977; 448-468.

[7] Gabr, M.M. and Subba Rao, T., "On the Identification of Bilinear Systems from Operating Records", IFAC Proceedings 1981; 15(1): 375-380.

[8] Ozaki, T., "The statistical analysis of perturbed limit cycle processes using non-linear time series models", Journal of Time Series Analysis, 1982; 3: 29-41.

[9] Tong, H., Threshold Models in Nonlinear Time-Series Analysis. 1983, Springer-Verlag, New York.

[10] Subba Rao, T., "Contribution to the discussion of Tong and Lim's paper", Journal of the Royal Statistical Society, B, 1980; 278-280.

[11] Lai, D., "Comparison study of AR models of the Canadian lynx data: A close look at BDS statistic", Computational Statistics & Data Analysis, 1996; 22(4): 409-423.

[12] Kaboudan, M.A., "Genetically evolved models and normality of their fitted residuals", Journal of Economic Dynamics and Control, 2001; 25(11): 1719-1749.

[13] Terui, N. and Van Dijk H. K., "Combined forecasts from linear and nonlinear time series models", International Journal of Forecasting, 2001; 18(3): 421-438.

[14] Khashei, M. and Bijari, M., "A new class of hybrid models for time series forecasting", Expert Systems with Applications, 2012; 39(4): 4344-4357.

[15] Chavas, J., "Modeling population dynamics: A quantile approach", Mathematical Biosciences, 2015; 262, 138-146, 2015.

[16] Zainuddin, N.H., Lola, M.S., Djauhari, M.A.,Yusof, F.,Ramlee, M.N.A., Deraman, A., Ibrahim, Y., Abdullah, M.T., "Improvement of time forecasting models using a novel hybridization of bootstrap and double bootstrap artificial neural networks", Applied Soft Computing, 2019: 105676.

[17] Panigrahi, S., Behera H.S., "A study on leading machine learning techniques for high order fuzzy time series forecasting", Engineering Applications of Artificial Intelligence, 2020; 87: 103245.

[18] Chen, W., Xu, H., Chen, Z., Jiang, M., "A novel method for time series prediction based on error decomposition and nonlinear combination of forecasters", Neurocomputing, 2021; 426 (22): 85-103.

[19] Kasap, R. And Kurt, E., "An Analysis on Modeling with Different Approaches for the Same Time Series", National Statistics Symposium 2000; 27-28 April, Ankara.

[20] Hardle, W., Hlavka, Z. "Multivariate Statistics: Exercises and Solutions", Springer 2007.

[21] Johnson, R.A., Wichern D.W. Applied Multivariate Statistical Analysis 2002, 5.th Edition, Prentice Hall, 2002.

[22] MacQueen, J.B., "Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability", Volume 1: Statistics, University of California Press, Berkeley, 1967; 281-297.

[23] Anderson, T.W. An Introduction to Multivariate Statistical Analysis. 1984, Second Edition, John Wiley&Sons, New York.

[24] Mirkin, B., Mathematical Clasification and Clustering. 1996, Kluwer Academic Publisher, London.

[25] Grassberger, P. & Procaccia, I. Characterization of strange attractors. Physical review letters, 1983; 50(5): 346.

[26] Brock, W. A., Brock, W. A., Hsieh, D. A. & LeBaron, B. D. Nonlinear dynamics, chaos, and instability: statistical theory and economic evidence. 1991, MIT press.

[27] Chappell, D., Padmore, J. and Ellis, C. A note on the distribution of BDS statistics for a real exchange rate series. Oxford Bulletin of Economics and Statistics, 1996; 58: 561–565.

[28] Kasap, R. and Kurt, E., An investigation of chaos in RL-diode circuit by using the BDS test Journal of Applied Mathematics Decision Sciences 1998; 2(2): 193-199.

[29] Tong H. Non-linear time series: a dynamical system approach. 1990, Oxford University Press, Oxford.

[30] Modiri, M., Homayounpour, M.M. & Ebadzadeh, M.M. Reservoir weights learning based on adaptive dynamic programming and its application in time series classification. Neural Comput & Applic 2022; 34: 13201–13217. https://doi.org/10.1007/s00521-021-06827-5

Appendix A. The Results of K-Means

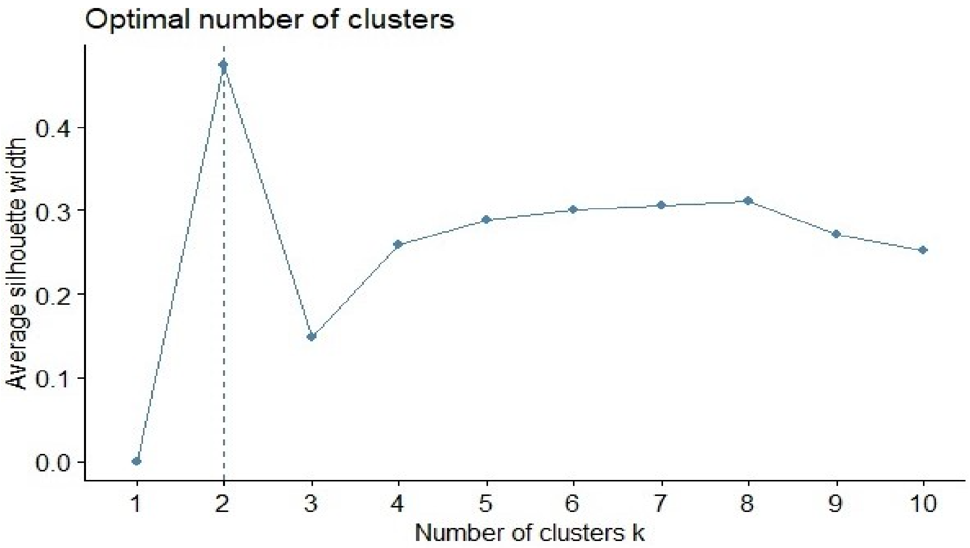

Figure 2. Optimal number of clusters

Table 2. Cluster membership

Model Number	Model Name	2 Clusters	3 Clusters	4 Clusters
1	a	1	1	4
2	b	1	1	1
3	c	1	1	4
4	d	1	3	2
5	e	1	3	2
6	f	1	3	2
7	g	1	1	1
8	h	1	1	1
9	i	1	1	1
10	j	1	1	1
11	k	1	1	1
12	l	1	1	1
13	m	2	2	3