Japanese Journal of Statistics and Data Science (2022) 5:303–319
https://doi.org/10.1007/s42081-022-00148-0

ORIGINAL PAPER

Statistics for COVID-19 Pandemic Data

A *k*-means method for trends of time series

An application to time series of COVID-19 cases in Japan

Norio Watanabe[1]

Received: 27 September 2021 / Revised: 12 January 2022 / Accepted: 10 February 2022 /
Published online: 3 March 2022

Abstract
A *k*-means method style clustering algorithm is proposed for trends of multivariate time series. The usual *k*-means method is based on distances or dissimilarity measures among multivariate data and centroids of clusters. Some similarity or dissimilarity measures are also available for multivariate time series. However, suitability of dissimilarity measures depends on the properties of time series. Moreover, it is not easy to define the centroid for time series. The *k*-medoid clustering method can be applied to time series using one of dissimilarity measures without using centroids. However, the *k*-medoid method becomes restrictive if appropriate medoids do not exist. In this paper, the centroid is defined as a common trend and a dissimilarity measure is also introduced for trends. Based on these centroids and dissimilarity measures, a *k*-means method style algorithm is proposed for a multivariate trend. The proposed method is applied to the time series of COVID-19 cases in each prefecture of Japan.

Keywords Dissimiliarity measure · Common trend · Clustering · *k*-Medoid method

Mathematics Subject Classification 62H30 · 62M10 · 62P10

1 Introduction

Clustering is one of the important issues in multivariate analysis. This is also true for multivariate time series. Clustering methods for time series depend on characteristics of time series. In this paper, we assume that time series contain trends, and then, they are nonstationary, and we propose a non-hierarchical clustering method, which has a *k*-means method style, for a multivariate trend.

✉ Norio Watanabe
watanabe@indsys.chuo-u.ac.jp

[1] Chuo University, Bunkyo-ku 1-13-27, Tokyo 112-8551, Japan

The usual k-means method is based on distances or dissimilarity measures among data and centers of gravity or centroids of clusters.

For time series, some similarity or dissimilarity measures are available. Pearson's correlation coefficient and the cosine similarity are typical similarity measures, and the dynamic time warping method can also be applied (see Bagnal (2017), Berndt and Clifford (1994), Egghe and Leydesdorff (2009), for example). However, suitability of dissimilarity measures depends on the properties of time series. Watanabe (2021) discussed dissimilarity measures for nonstationary time series and proposed other dissimilarity measures explained in the following section. In this paper, we introduce a new dissimilarity measure for applying to clustering. The new one has a similar form to those by Watanabe (2021).

Unlike similarity or dissimilarity measures, it is not easy to define the centroid for time series. Without using centroids, the k-medoid clustering method can be applied to time series using one of dissimilarity measures (cf. Kaufman and Rousseeuw (2009)). However, the k-medoid method is restrictive if appropriate medoids, which represent the centers of clusters, do not exist. Our aim is to analyze a multivariate trend of nonstationary time series. In this case, an idea of common trend can be adopted for definition of the centroid for time series. In this paper, we apply the definition of common trend given by Watanabe (2021). However, Watanabe (2021) does not consider the lags. Therefore, we provide a new definition of common trend by considering lags, since lags are important in time series analysis. The lag is especially crucial for multivariate time series.

The main task of clustering is to find similar patterns in time series. We assume that time series under analysis are ratio scale data and positively correlated each other in some sense. We also assume that standardization is appropriate for comparison with each other. If time series are nonstationary, standardization is also difficult. The dissimilarity measure in this paper is based on weighting as standardization. We propose a k-means method style algorithm for a multivariate trend using a new dissimilarity measure and the common trend.

The proposed method is applied to the time series of COVID-19 cases in each prefecture of Japan. We also discuss clustering of original nonstationary time series themselves not trends briefly using COVID-19 series.

2 Dissimilarity measures

Target data in this paper are P-variate time series and we consider problems on their trends based on dissimilarity among time series. First, we discuss dissimilarity between two time series in this section.

Let (x_n, y_n) $(n = 1, 2, \ldots, N)$ be an observed bivariate time series. We assume that

$$x_n = T_n + v_n \tag{1}$$

$$y_n = U_n + w_n, \tag{2}$$

where $\{(v_n, w_n)\}$ is a bivariate zero mean stationary process, and $\{T_n\}$ and $\{U_n\}$ are trends or mean value functions. We assume that trends are (conditionally) deterministic

and that they are estimated appropriately. The moving average method is a simple way for trend estimation. Other methods are found in Kim (2009) and Watanabe and Watanabe (2015), for example. Moreover we assume that the seasonality or periodic movements are absent in estimated trends; so-called cyclic components can be included in trends. Note that the period of cyclic component is relatively large comparing to the one of seasonality.

It is not easy to capture dissimilarity between nonstationary time series, unlike stationary time series. Watanabe (2021) introduced the dissimilarity functions between $\{T_n\}$ and $\{U_n\}$ as follows:

Definition 1 (Watanabe 2021) The simple dissimilarity measure function $\delta_S(\ell)$ is given by

$$\delta_S(\ell) = \sqrt{\frac{\sum_{n=1}^{N-\ell}(T_n - U_{n+\ell})^2}{\sum_{n=1}^{N-\ell} T_n^2 + \sum_{n=\ell+1}^{N} U_n^2}} \tag{3}$$

for $\ell = 0, \pm 1, \pm 2, \ldots$.

Definition 2 (Watanabe 2021) The weighted dissimilarity measure function $\delta_W(\ell)$ is given by

$$\delta_W(\ell) = \inf_{(r_T, r_U) \in R} \sqrt{\frac{\sum_{n=1}^{N-\ell}(r_T T_n - r_U U_{n+\ell})^2}{r_T^2 \sum_{n=1}^{N-\ell} T_n^2 + r_U^2 \sum_{n=\ell+1}^{N} U_n^2}} \tag{4}$$

for $\ell = 0, \pm 1, \pm 2, \ldots$, where $R = \{(r_T, r_U) |\, r_T^2 + r_U^2 = 1, r_T \geq 0, r_U \geq 0\}$.

The region of minimization can be replaced by $R_1 = \{(r_T, r_U) |\, r_T^2 + r_U^2 = 1, r_T \geq 0\}$ or $R_2 = \{(r_T, r_U) |\, r_T^2 + r_U^2 = 1, r_U \geq 0\}$. The choice depends on the property of time series.

Definition 3 (Watanabe 2021) The normalized dissimilarity measure function $\delta_N(\ell)$ is given by

$$\delta_N(\ell) = \inf_{(r_T, r_U) \in R, (c_T, c_U) \in C} \sqrt{S_\ell}, \tag{5}$$

where

$$S_\ell = \frac{\sum_{n=1}^{N-\ell}(r_T(T_n - c_T) - r_U(U_{n+\ell} - c_U))^2}{r_T^2 \sum_{n=1}^{N-\ell}(T_n - c_T)^2 + r_U^2 \sum_{n=\ell+1}^{N}(U_n - c_U)^2} \tag{6}$$

for $\ell = 0, \pm 1, \pm 2, \ldots$ and C is a bounded subset of 2-dimensional Euclidean space.

The independent variable of these functions is the lag ℓ similarly to the cross correlation function for the stationary time series.

Watanabe (2021) showed that the k-medoid clustering method (Kaufman and Rousseeuw 2009) can be applied to time series using these dissimilarity measures (an example in Watanabe (2021) are based on $\delta_W(0)$). The k-medoid method can be used when it is difficult to define centroids (cf. Kaufman and Rousseeuw (2009)).

The dissimilarity measures δ_W and δ_N are essentially invariant for the exchange of time series (δ for ℓ is equal to the original δ for $-\ell$). This symmetric property is not required for dissimilarity measure between one trend and a centroid of trends. If a kind of average time series in each cluster is available instead of the medoid, it is not necessarily adequate to use the weights for both time series. The reason is that the weights for the centroid should be invariant. In Watanabe (2021), the common trend for multivariate time series is also defined as shown below. We can use the common trend in each cluster instead of the medoid as the centroid. Then, the weight for the common trend becomes unnecessary. In this paper, we propose another dissimilarity function for clustering based on $\delta_W(\ell)$ but in a slightly different form.

Definition 4 Let $\{C_n\}$ be the given common trend. The dissimilarity measure function $\delta_C(\ell)$ between $\{T_n\}$ and the common trend $\{C_n\}$ is given by

$$\delta_C(\ell) = \inf_{r>0} \sqrt{\frac{\sum_{n=1}^{N-\ell}(C_n - rT_{n+\ell})^2}{r^2 \sum_{n=\ell+1}^{N} T_n^2}} \tag{7}$$

for $\ell = 0, \pm 1, \pm 2, \ldots$.

The restriction $r > 0$ means that time series under consideration are assumed to be positively correlated each other in some sense.

It is easily shown that

(1) $0 \le \delta_C(\ell) \le 1$,
(2) $\delta_C(\ell) = 0 \iff aT_{n+\ell} = C_n (\forall n)$ where a is a positive constant,
(3) if $|aT_{n+\ell} - C_n| \le \epsilon$ and $|aT_n| \ge d > 0$ $(\forall n)$ for $\exists a$, then $\delta_C(\ell) \le \epsilon/d$.

Moreover, we have the following theorem.

Theorem 1 *Suppose that*

$$r_0 = \frac{\sum_{n=1}^{N-\ell} C_n^2}{\sum_{n=1}^{N-\ell} C_n T_{n+\ell}} > 0. \tag{8}$$

Then, the minimum value of the right-hand side of Eq. (7) is attained at $r = r_0$ and we have

$$\delta_C(\ell) = \inf_{\rho>0} \sqrt{\frac{\sum_{n=1}^{N-\ell}(\rho\, C_n - T_{n+\ell})^2}{\sum_{n=\ell+1}^{N} T_n^2}} \tag{9}$$

$$= \sqrt{1 - \frac{\left(\sum_{n=1}^{N-\ell} C_n T_{n+\ell}\right)^2}{\sum_{n=1}^{N-\ell} C_n^2 \sum_{n=\ell+1}^{N} T_n^2}}. \tag{10}$$

The proof is easy. The cosine similarity is one of the well-known similarity measures (Egghe and Leydesdorff 2009). The cosine similarity between $\{C_n\}$ and $\{T_{n+\ell}\}$ is given by

$$\frac{\sum_{n=1}^{N-\ell} C_n T_{n+\ell}}{\sqrt{\sum_{n=1}^{N-\ell} C_n^2 \sum_{n=\ell+1}^{N} T_n^2}}. \tag{11}$$

We can consider that the above theorem provides some validity of the cosine similarity. (The relation between the cosine similarity and the dissimilarity δ_S in Definition 1 is stated in Watanabe (2021).)

The use of the dissimilarity δ_C with some appropriately defined common trend makes a k-means style clustering possible. In the following section, we introduce the common trend which is well suited to δ_C.

3 Common trend

In this paper, we do not assume any models and define the common trend as the weighted sum of the multiple trends given by the solution of an optimization problem. The formulation is similar to the one by Watanabe Watanabe (2021) except for the existence of lags.

Let $T_n = \left(T_{1n}, \ldots, T_{Pn}\right)$ $(n = 1, 2, \ldots, N)$ be the P-dimensional vector of trends or mean value functions of P-variate time series. First, we introduce the common trend given by Watanabe Watanabe (2021).

Definition 5 (Watanabe 2021) The common trend $\{C_n^{(0)}\}$ is the time series given by the optimization problem

$$\sum_{p=1}^{P} \sqrt{\frac{\sum_{n=1}^{N} (C_n^{(0)} - r_p T_{pn})^2}{r_p^2 \sum_{n=1}^{N} T_{pn}^2}} \quad \longrightarrow \quad \text{minimize} \tag{12}$$

with respect to $r_1, \ldots, r_P$, where $C_n^{(0)} = \sum_{p=1}^{P} r_p T_{pn}$, $r_1 \geq 0, r_2 \geq 0, \ldots, r_P \geq 0$ and $\sum_{p=1}^{P} r_p = 1$.

Each term in the objective function (12) has the same form as Eq. (7) with $\ell = 0$. That is, lags are not considered in this definition. However, it is important to consider lags for multivariate time series. It is especially important for discussing the common trend. For example, studies on business cycle are crucial in the econometric field. The business cycle is related to the common trend of many time series, and these time series consist of leading, coincident, and lagging indicators. This means that plus or minus lags play key roles. In this paper, we generalize the above definition for considering lags.

Let ℓ_p denote the lag for $\{T_{pn}\}$ and assume that $-L_{\max} \leq \ell_p \leq L_{\max}$, where $L_{\max}$ is a given integer.

Definition 6 The common trend $\{C_n\}$ is the time series given by the optimization problem

$$J_C(r_1, \ldots, r_P, \ell_1, \ldots, \ell_P) \longrightarrow \text{minimize} \tag{13}$$

with respect to $r_1, \ldots, r_P$ and $\ell_1, \ldots, \ell_P$, where $r_p \geq 0$, $\sum_{p=1}^{P} r_p = 1$ and $-L_{\max} \leq \ell_p \leq L_{\max}$. The objective function is defined by

$$J_C(r_1, \ldots, r_P, \ell_1, \ldots, \ell_P) = \sum_{p=1}^{P} \sqrt{\frac{\sum_{n=L_{\max}+1}^{N-L_{\max}} (C_n - r_p T_{p,n+\ell_p})^2}{r_p^2 \sum_{n=L_{\max}+1}^{N-L_{\max}} T_{p,n+\ell_p}^2}}, \tag{14}$$

where $C_n = \sum_{p=1}^{P} r_p T_{p,n+\ell_p}$.

The optimization (13) is not easy unless both P and $L_{\max}$ are small. Watanabe Watanabe (2021) proposed the recursive algorithm for $C_n^{(0)}$ in Definition 5. In the following, we propose an extended recursive algorithm for C_n in Definition 6.

Estimation algorithm for common trend (ECT)

Step 1. Initialize C_n by setting $r_p = 1/P$, that is

$$C_n = \frac{1}{P} \sum_{p=1}^{P} T_{pn}. \tag{15}$$

Step 2. For fixed $\{C_n\}$, find r_p and ℓ_p that minimize

$$\frac{\sum_{n=L_{\max}-1}^{N-L_{\max}} (C_n - r_p T_{p,n+\ell_p})^2}{r_p^2 \sum_{n=L_{\max}+1}^{N-L_{\max}} T_{p,n+\ell_p}^2} \qquad (p = 1, 2, \ldots, P). \tag{16}$$

Step 3. Replace r_p by $r_p / \sum_{p=1}^{P} r_p$ and calculate

$$C_n = \sum_{p=1}^{P} r_p T_{p,n+\ell}. \tag{17}$$

Step 4. Calculate J_C in Eq. (14).
Step 5. Go to Step 2 until some termination condition is satisfied.
Step 6. Select $\{C_n\}$ with $\{r_p, \ell_p | p = 1, \ldots, P\}$ that minimizes J_C.

We call this an ECT algorithm. In Step 2, the analytical solution is available. See Theorem 1. This algorithm is applied to practical time series in Sect. 5. A comparison with the Definition 6 is also considered.

4 Clustering

Now, we propose a k-means method based on the dissimilarity function $\delta_C(\ell)$ and the common trend calculated by the recursive algorithm in Sect. 3.

Target data are the P-variate trend $T_n = \left(T_{1n}, \ldots, T_{Pn}\right)$ $(n = 1, 2, \ldots, N)$. Let K be a given number of clusters. We define the function u_{kp} as follows:

$$u_{kp} = \begin{cases} 1 \text{ if } \{\mathrm{T_{pn}|n} = 1, \ldots, \mathrm{N}\} \text{ belongs to k} - \text{th cluster} \\ 0 \text{ otherwise,} \end{cases} \tag{18}$$

where $\sum_{k=1}^{K} u_{kp} = 1(\forall p)$ and $\sum_{p=1}^{P} u_{kp} > 0(\forall k)$.

Let L_{MAX} be the upper bound of lags and L_{KMT} be a number of repetitions. The following is the recursive algorithm for clustering of trends.

K-means method for trend (KMT)

Step 1. (**Initialize**) Set $\{u_{kp}\}$ randomly.

Step 2. (**Centroid**) Estimate the common trend $\{C_{kn}|n = L_{\mathrm{MAX}} + 1, \ldots, N - L_{\mathrm{MAX}}\}$ of k-th cluster by the ECT algorithm $(k = 1, \ldots, K)$.

Step 3. (**Dissimilarity**) Calculate the dissimilarity $\delta_C(\ell)$ between the common trend $\{C_{kn}\}$ and each trend $\{T_{pn}|n = L_{\mathrm{MAX}} + 1, \ldots, N - L_{\mathrm{MAX}}\}$ for $p = 1, \ldots, P$, $k = 1, \ldots, K$ and $-L_{\mathrm{MAX}} \leq \ell \leq L_{\mathrm{MAX}}$. Let δ_{kp} be the minimum value with respect to ℓ.

Step 4. (**Reassignment**) Redefine u_{kp} as follows:

$$u_{kp} = \begin{cases} 1 \text{ if } \delta_{\mathrm{kp}} = \min_{1 \leq \mathrm{j} \leq \mathrm{K}} \delta_{\mathrm{jp}} \\ 0 \text{ otherwise} \end{cases} \tag{19}$$

(the tie is not considered here for simplicity).

Step 5. If some change in $\{u_{kp}\}$ occurs, goto Step 2. If there is no change, calculate the value

$$J = \sum_{k=1}^{K} \sum_{p=1}^{P} u_{kp} \delta_{kp}. \tag{20}$$

Step 6. Go to Step 1 $L_{\mathrm{KMT}} - 1$ times.

Step 7. The classification with the minimum value of J is adopted as the result.

We call this a KMT algorithm. Similarly to the usual k-means method, sensitivity of initial values is large. That is, it is not assured that the solution is the global minimum. Usually, L_{KMT} should be set relatively large. If a tie occurs in Step 4, k can be determined randomly similarly to the usual k-means method. An efficient approach to solve the tie problem is to extend the hard clustering to soft clustering. In this paper, we consider hard clustering only.

An important feature of the ECT and KMT algorithm is that any extra numerical optimization technique is required, though some nonlinear optimization is required for the use of δ_W instead of δ_C.

An application to time series of COVID-19 cases is demonstrated in the next section.

5 Application to time series of COVID-19 cases

5.1 Data set

Time series consist of daily record numbers of COVID-19 cases in prefectures. The number of prefectures of Japan is 47 (= P). Data are provided by NHK (Japan Broadcasting Corporation). The upper plot in Fig. 1 shows 47 time series from January 16, 2020 to August 1, 2021. The length is 564. The first day of each month is indicated by the vertical dotted line.

We estimate the trends of original time series by the moving average method. We adopt the triangular weight function whose support has the length 28 (4 weeks). The length becomes 536 (= N), since any processing for both ends of series is not applied here. The lower plot in Fig. 1 shows the moving averaged series. A purpose here is to find similarity of patterns. For this purpose, we apply the proposed KMT algorithm. First, we examine the ECT algorithm.

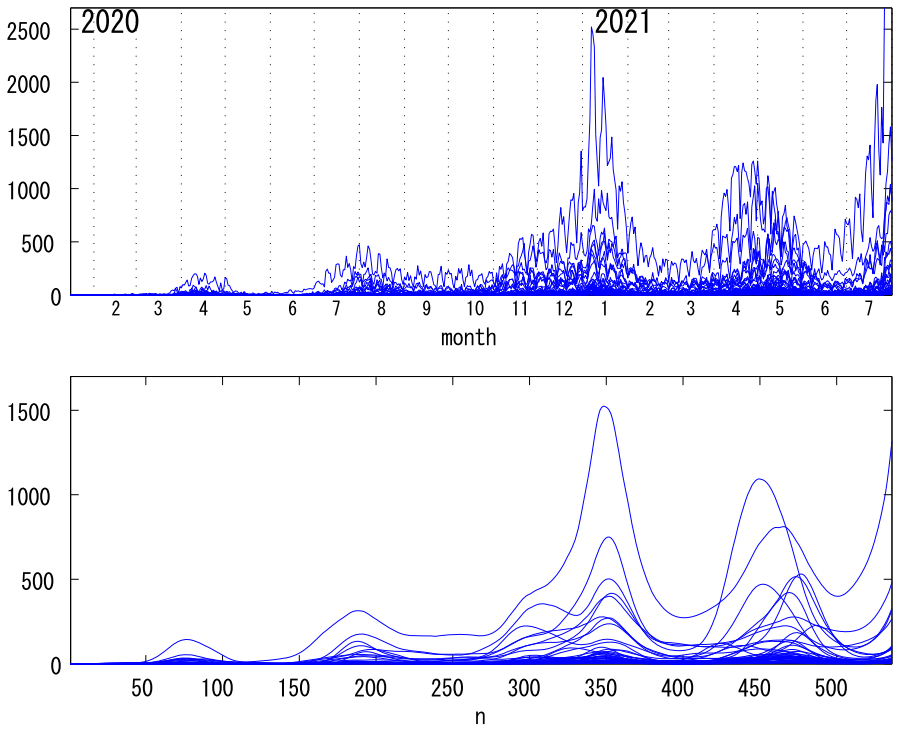


Fig. 1 Original series and moving averaged series

5.2 Estimation of common trend

We consider the estimation of the common trend among three time series of Tokyo, Osaka, and Hokkaido without lags by two methods. The first is to apply numerical optimization technique to minimize the objective function in Definition 5. The second is to apply the ECT algorithm without lags ($L_{\mathrm{MAX}} = 0$). (A similar example is illustrated in Watanabe (2021).) Original time series and moving averaged time series are shown in Fig. 2.

Calculation in this paper is carried out using MATLAB. The MATLAB function 'fmincon' is used for minimization under constraints in Definition 5.

Two estimated common trends are illustrated in Fig. 3. The right is the partly magnified plot. It is found that the difference is quite small.

Values of the objective function J_C obtained by two methods are plotted in Fig. 4. We can say that the recursive ECT algorithm works well for these data. The estimated common trend and three weighted trends are plotted in Fig. 5, where r_0 in Theorem 1 is used as each weight. Tendency of three time series is reflected roughly in the estimated common trend plotted by bold line. However, it seems that there exist lags in three time series, and then, the common trend fluctuates unnaturally.

Figure 6 shows the estimated common trend obtained by the ECT algorithm with lags by setting $L_{\mathrm{MAX}} = 30$.

We can say that the ECT algorithm with lags also works well. However, there exists some differences among patterns even if lags are considered. This suggests the

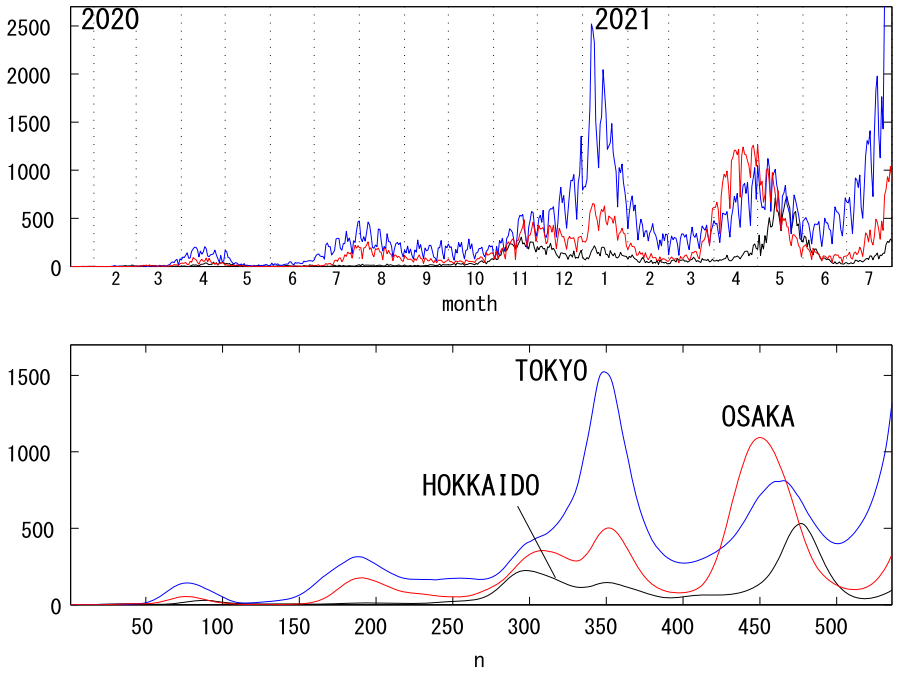


Fig. 2 Three series (Tokyo, Osaka, Hokkaido)

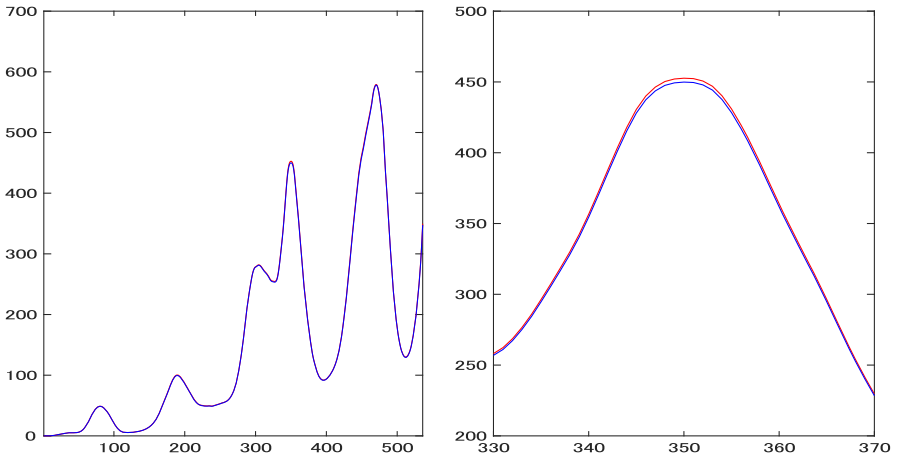


Fig. 3 Comparison between two methods (1) common trends

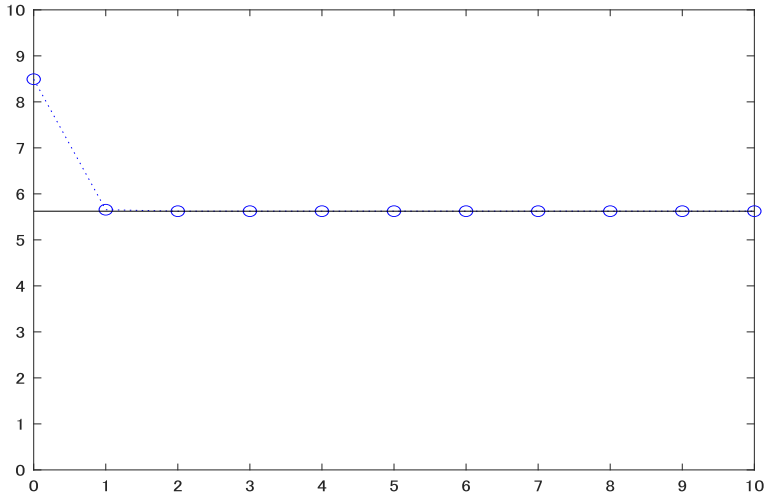


Fig. 4 Comparison between two methods (2) values of J_C

necessity of clustering. In the following, clustering based on the ECT algorithm with lags is applied to all time series.

5.3 *K*-means method for trends

The proposed KMT algorithm is applied to 47 time series by setting $K = 1, \ldots, 8$, $L_{\max} = 30$ and $L_{\mathrm{KMT}} = 1000$. The KMT algorithm with $K = 1$ is not clustering but means the estimation of the common trend of all time series. The results of K-means method for trends are summarized in Table 1 and Fig. 7 ($K = 1, \ldots, 6$). Table 1 includes sizes of clusters, values of J, the maximum of $|\ell_p|$, and the maximum of $\ell_p - \ell_q$ in each cluster. Each trend in Fig. 7 is weighted using r_0 in Theorem 1.

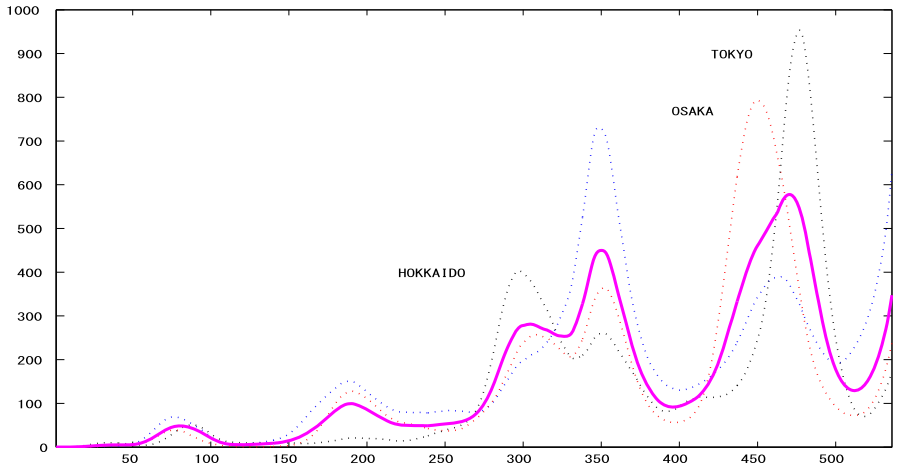


Fig. 5 Common trend of three series

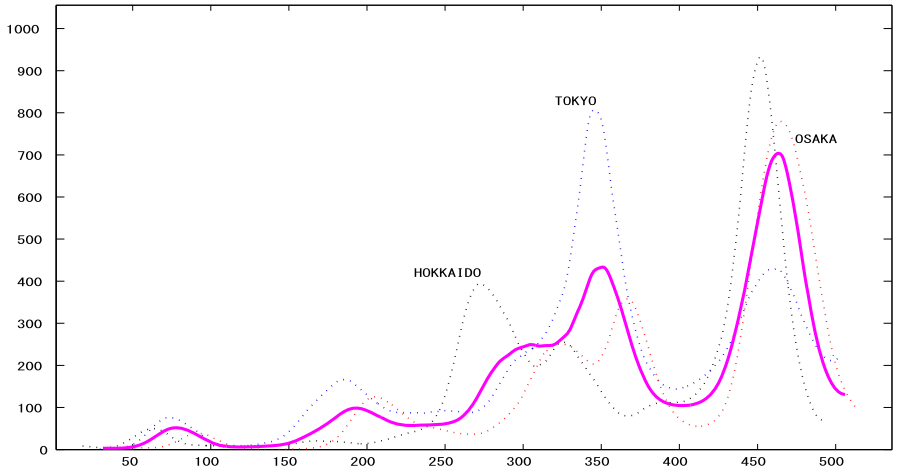
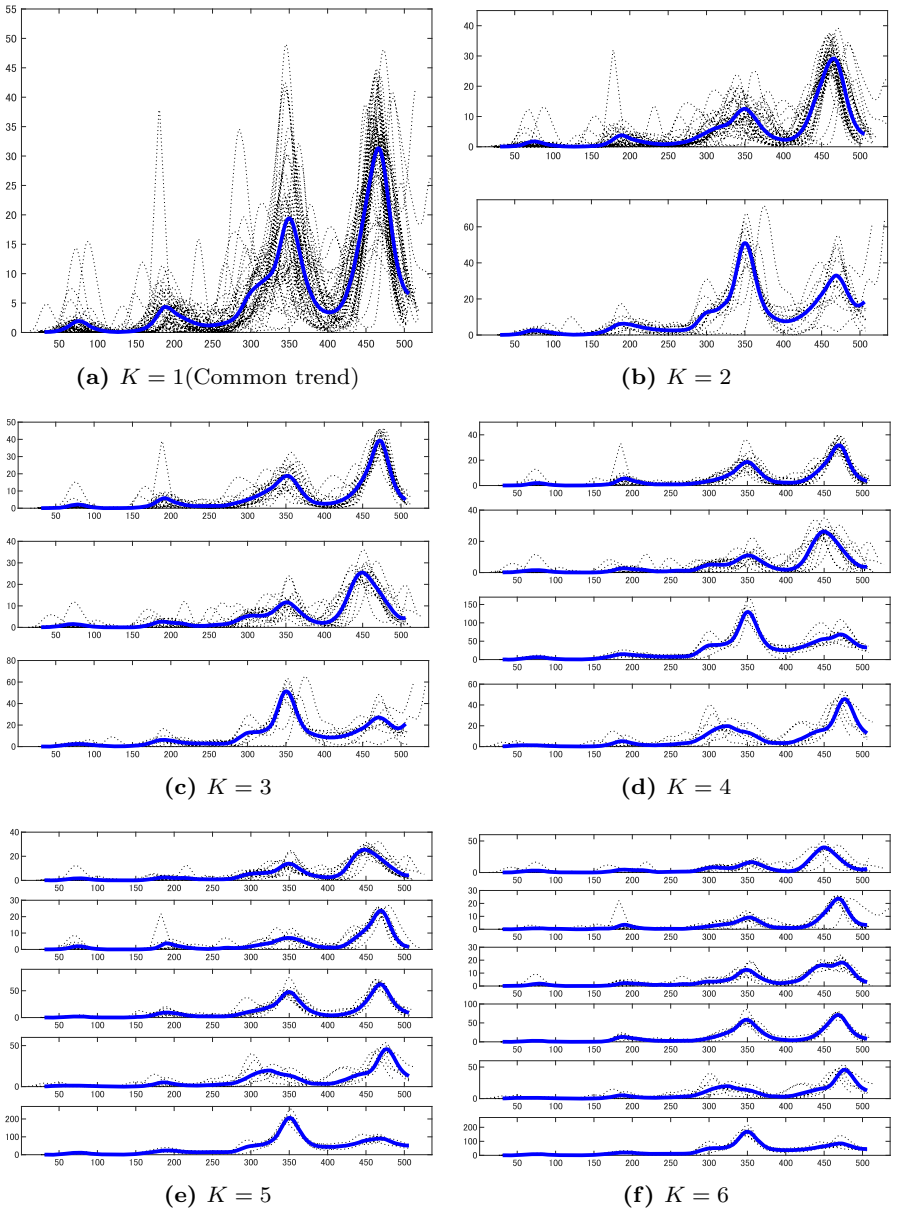


Fig. 6 Common trend of three *lagged* series

The determination of K is important but difficult problem. For usual k-means method, some methods have been proposed; for example, the elbow method, silhouette method, gap statistic, and so on (cf. Yuan and Yang (2019)). However, there is no definitive method. This is true for the proposed K-means method for trends. Therefore, we do not refer the determination of K in detail, since this problem should be discussed separately.

Values of J monotonically decrease and there is no clear "elbow". Selected lags or differences between lags are relatively large when K is 2, 3, 4, or 6, but it is difficult to explain large lags. When K is less than 5, it seems that peaks are not separated well (especially the third peak). One plausible candidate of the number of clusters is 5. The result for $K = 5$ is illustrated again in Fig. 8.



(a) $K = 1$(Common trend) **(b)** $K = 2$

(c) $K = 3$ **(d)** $K = 4$

(e) $K = 5$ **(f)** $K = 6$

Fig. 7 Results for $K = 1, 2, \ldots, 6$

Table 1 Summary of KMT

C.No. K	1	2	3	4	5	6	7	8	J	Max $\lvert\ell\rvert$	Max ℓ−Min ℓ
1	47	–	–	–	–	–	–	–	16.45	30	48
2	33	14	–	–	–	–	–	–	13.85	30	46
3	20	17	10	–	–	–	–	–	12.72	27	40
4	17	15	8	7	–	–	–	–	11.94	23	39
5	14	10	10	7	6	–	–	–	11.27	21	26
6	9	9	8	7	7	7	–	–	10.68	30	33
7	8	8	8	8	7	7	1	–	10.18	22	25
8	9	8	7	7	6	5	4	1	9.82	22	25

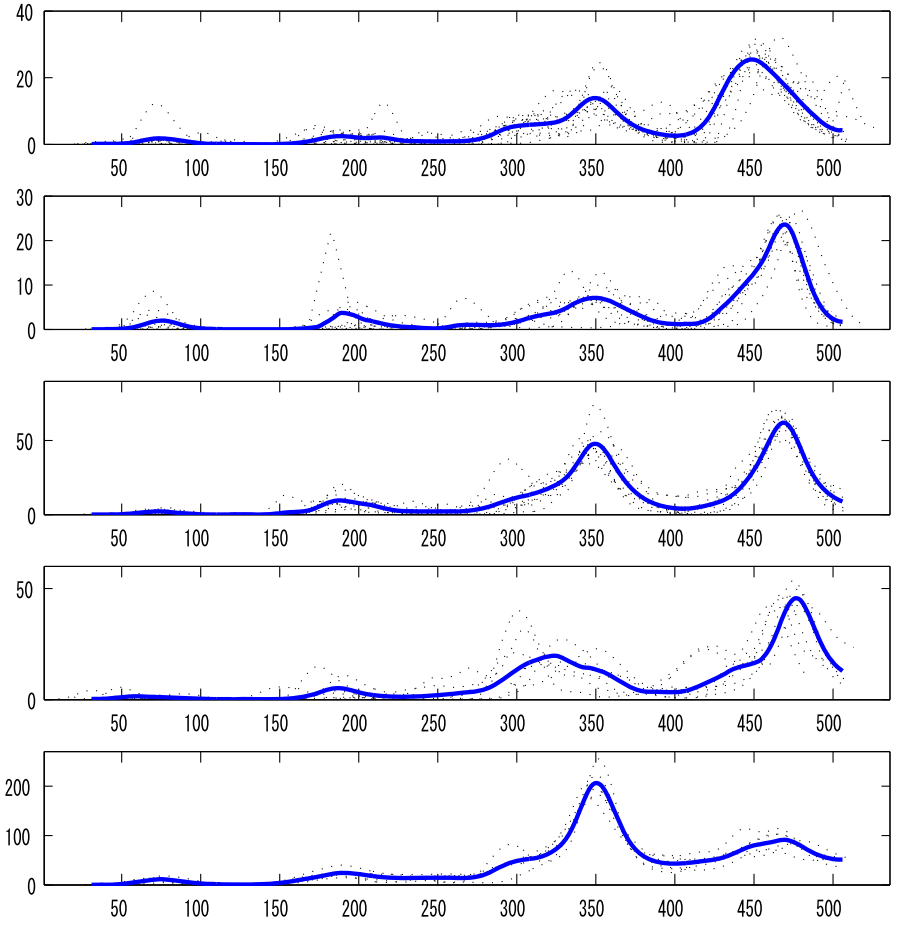


Fig. 8 Result for $K = 5$

Table 2 Prefectures in clusters ($K = 5$)

C.No.	Prefecture	ℓ	C.No.	Prefecture	ℓ	C.No.	Prefecture	ℓ
1	Miyagi	21	2	Aomori	0	3	Gumma	− 2
	Yamagata	14		Akita	− 4		Gifu	2
	Fukushima	5		Toyama	6		Shizuoka	4
	Niigata	3		Ishikawa	3		Kyoto	− 4
	Fukui	− 2		Shimane	5		Yamaguchi	4
	Nagano	-2		Okayama	1		Fukuoka	0
	Mie	8		Tokushima	15		Nagasaki	− 4
	Shiga	10		Kagawa	− 2		Kumamoto	0
	Osaka	0		Saga	− 1		Miyazaki	− 2
	Hyogo	1		Oita	0		Kagoshima	2
	Nara	0						
	Wakayama	− 1						
	Tottori	− 7						
	Ehime	− 6						

C.No.	Prefecture	ℓ	C.No.	Prefecture	ℓ
4	Hokkaido	− 1	5	Ibaraki	3
	Iwate	− 8		Tochigi	− 3
	Yamanashi	20		Saitama	0
	Aichi	− 2		Chiba	3
	Hiroshima	− 1		Tokyo	− 1
	Kochi	10		Kanagawa	0
	Okinawa	9			

Table 2 shows prefectures in each cluster, where ℓ indicates the selected lag to the common trend of each cluster. It is meaningless to compare lags among different clusters.

It is found that Tokyo, Osaka, and Hokkaido belong to different clusters. This means that the estimated common trend in the previous subsection is not so meaningful.

It is said that these COVID-19 series contain four waves. The first peak appears around $n = 70$. Differences among patterns in each cluster appear in the heights of peaks or locations of peaks. In clusters 1, 2, and 4, the fourth peak is remarkably large. On the other hand, the third peak is largest in cluster 5.

Cluster 5 consists of prefectures in Kanto area, whose center is Tokyo, except for Gumma. We can say that tendency of each prefecture in Kanto area is resemble. Cluster 1 mainly consists of most prefectures in Kinki area, whose center is Osaka, and prefectures in south Tohoku area. Cluster 3 mainly consists of prefectures in Kyushu area.

It is expected that such a statistical analysis will provide epidemiologically or medically useful findings.

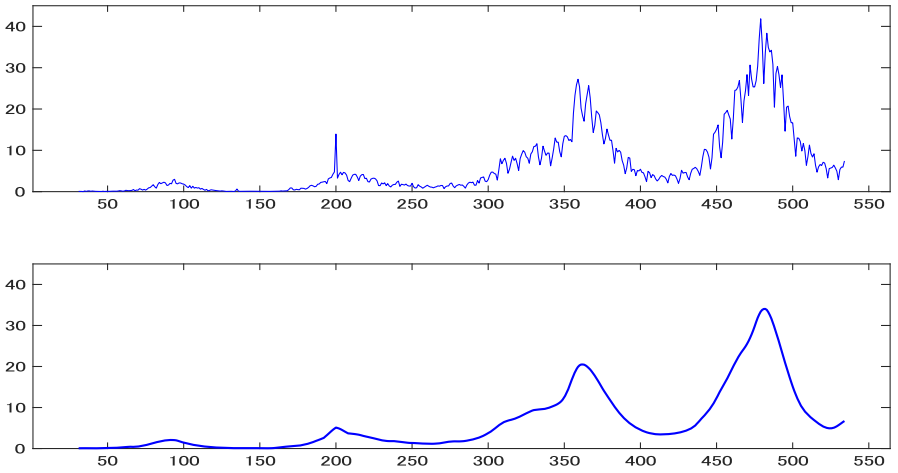


Fig. 9 Common trends

6 Clustering of original series

In this paper, we have focused on the clustering of a multivariate trend. In this final section, we consider the clustering of original series briefly.

The proposed KMT algorithm can be applied directly to original time series. However, the meaning of centroid of cluster becomes vague, since it is defined as the weighted sum of original series in the ECT algorithm.

As an example, the estimated centroid of all original series of COVID-19 cases by the ECT algorithm is shown in the upper graph in Fig. 9.

The estimated series includes seasonality and irregular fluctuation. It is clear that this series cannot be regarded as the common trend. Moreover, the meaning of seasonality of this series becomes vague, since lags are considered. As a result, this series is not appropriate as the centroid. This means that the direct application of the KMT algorithm to original series is not appropriate usually and the modification of algorithm is required for the direct application. Note that the estimated series might be regarded as the common trend approximately, if P is large and seasonality is absent in original time series.

A modification is achieved easily by appending a trend estimation step between Steps 3 and 4 in the ECT algorithm. The moving average method is a simple way for trend estimation. In this example, we adopt the moving average method with the triangular weight function whose support has the length 15 (half a month). The length of the smoothed series does not change here, since the both ends of series are processed in a simple way introduced in Brockwell and Davis (1991) for the sake of brevity. The estimated centroid of all original series of COVID-19 cases by the modified ECT algorithm is shown in the lower graph in Fig. 9. The estimated series can be regarded as the common trend and then as the centroid.

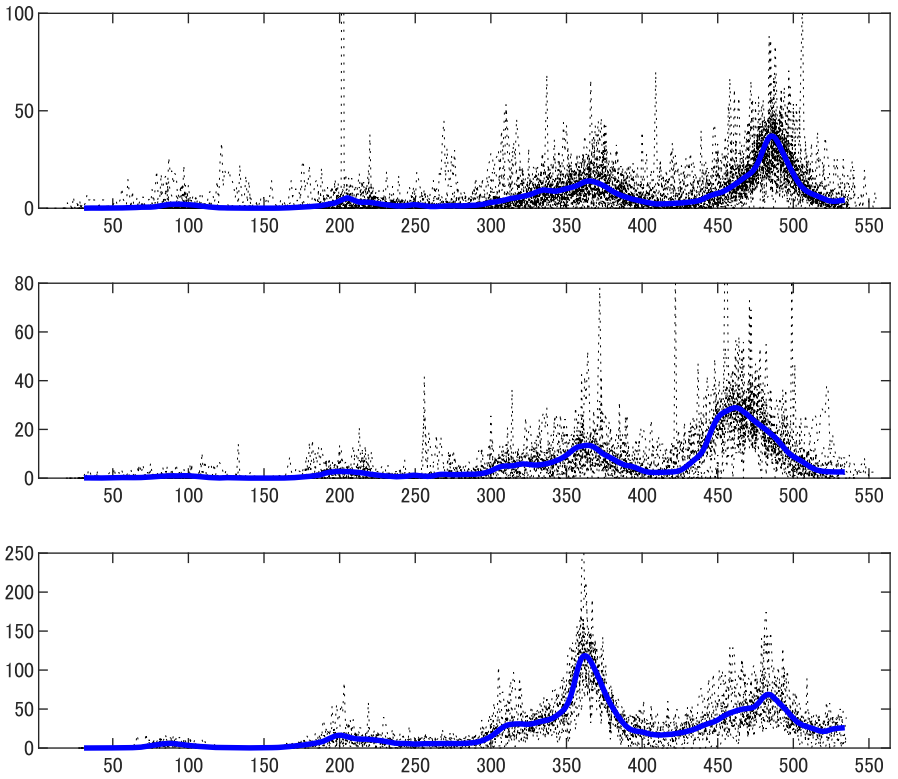


Fig. 10 Result for $K = 3$

Clustering of original series can be achieved by the KMT algorithm combined with the modified ECT algorithm. A clustering result for original series of COVID-19 cases with $K = 3$ is illustrated in Fig. 10.

The result is similar to the one by the multivariate trend of COVID-19 series, but not identical. The differences occur from the seasonality components and large irregular fluctuations in original series.

It should be considered well whether the clustering of original series is appropriate or not, since original series include trends, seasonal components, and irregular components usually. That is, results of clustering depend on various factors.

7 Concluding remarks

For the case when series are stationary processes, the cross correlation functions can be used as the similarity measure and it is meaningless to consider the common trend. In this case, the ECT or KMT algorithm is not efficient. On the other hand, clustering of a multivariate trend is appropriate, if the purpose is to clarify relationship among trends for nonstationary time series. It is expected that the proposed method provides an additional tool for trend analysis.

When some meaningful clusters are found by the proposed method, there is a possibility that results can be applied to prediction. In this case, it is expected to analyze multiple time series in the same cluster from the viewpoint of multivariate time series analysis. However, we should note that the prediction of trends is difficult usually because of nonstationarity.

In Step 3 of the KMT Algorithm, the number of $\delta_C(\ell)$'s to be computed becomes huge, when P is large and $L_{\max}$ should be large. Some simplification will be required in this case. One way is to consider a subset of lags. Another way is to derive lagged time series previously by considering the dissimilarity function between each trend and the common trend of all series, and then to apply the proposed method without lags.

In this paper, we assume that trends are estimated appropriately. For this assumption roles of trend estimation methods are important and essential. In the case of the moving average method, the length of the moving average is crucial. We have to pay sufficient attention for trend estimation.

When all values of time series are positive, we can try the log transformation. For log-transformed time series, the validity of our methods becomes doubtful. However, the log transformation cannot be applied to time series including zero values like COVID-19 series. On the other hand, the proposed method can be applied to time series including not only zero but also negative values.

Similarly to usual k-means method, the determination of the cluster size is important. Further studies are expected for determination of K. An extension to fuzzy clustering, that is, an extension from hard to soft clustering is an issue in future.

Acknowledgements This work was supported by Chuo University Personal Research Grant. We would like to thank reviewers for useful comments.

Declarations

Conflict of interest The author declare no conflict of interest.

References

Bagnal, A., et al. (2017). The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining Knowledge Discovery, 31,* 606–660.

Berndt, D. & Clifford, J. Using dynamic time warping to find patterns in time series. In *AAAI-94 workshop on knowledge discovery in databases* (AAAI Press), 359–370 (1994).

Brockwell, P. J., & Davis, R. A. (1991). *Time series: Theory and methods* (2nd ed.). Springer.

Egghe, L., & Leydesdorff, L. (2009). The relation between Pearson's correlation coefficient and Salton's cosine measure. *Journal of the American Society for Information Science and Technology, 60*(5), 1027–1036.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. Wiley.

Kim, S.-J., et al. (2009). ℓ_1 trend filtering. *SIAM Review, 51*(2), 339–360.

Watanabe, N. Dissimilarity measures for time series and trend analysis: Application to COVID-19 cases series. *Journal of Mathematics and Systems Sciences***(to appear)**.

Watanabe, E., & Watanabe, N., et al. (2015). Weighted multivariate fuzzy trend model for seasonal time series. In L. Filus (Ed.), *Stochastic modeling, data analysis and statistical applications* (pp. 443–450). ISAST.

Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. *Multidisciplinary Scientific Journal, 2*(6), 226–235.