International Journal of Data Warehousing and Mining
Volume 19 • Issue 3

Clustering of COVID-19 Multi-Time Series-Based K-Means and PCA With Forecasting

Sundus Naji Alaziz, Department of Mathematical Sciences, Faculty of Science, Princess Nourah bint Abdulrahman University, Saudi Arabia

Aziza Ali Alshowiman, Department of Mathematical Sciences, Faculty of Science, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

Bakr Albayati, Department of Basic Sciences, Common First Year King Saud University, Riyadh, Saudi Arabia

Abd al-Aziz H. El-Bagoury, Basic Sciences Department, Higher Institute of Engineering and Technology, El-Mahala El-Kobra, Egypt

https://orcid.org/0000-0002-2853-0762

Wasswa Shafik, Faculty of Basic Sciences and Information Technology, Ndejje University, Kampala, Uganda*

https://orcid.org/0000-0002-9320-3186

ABSTRACT

The COVID-19 pandemic is one of the current universal threats to humanity. The entire world is cooperating persistently to find some ways to decrease its effect. The time series is one of the basic criteria that play a fundamental part in developing an accurate prediction model for future estimations regarding the expansion of this virus with its infective nature. The authors discuss in this paper the goals of the study, problems, definitions, and previous studies. Also they deal with the theoretical aspect of multi-time series clusters using both the K-means and the time series cluster. In the end, they apply the topics, and ARIMA is used to introduce a prototype to give specific predictions about the impact of the COVID-19 pandemic from 90 to 140 days. The modeling and prediction process is done using the available data set from the Saudi Ministry of Health for Riyadh, Jeddah, Makkah, and Dammam during the previous four months, and the model is evaluated using the Python program. Based on this proposed method, the authors address the conclusions.

KEYWORDS

Clustering of COVID-19, K-Means, Multi-Time Series Clusters, PCA

1. INTRODUCTION

In the last year, COVID-19 has dominated the entire world with its deadly universal impacts on the health community. In general, there have been various academic research projects that have made priceless attempts to assess epidemic samples and models across a wide range of locations to provide an all-encompassing picture of its development. For example, in Saudi Arabian, the topic of the epidemic becomes a rich soil for a great number of research and academic studies. The focus of the current

DOI: 10.4018/IJDWM.317374 *Corresponding Author

study is concerned mainly with one of those numerous studies that are expected to play a useful role through its non-parametric method of extracting significant information from non-systemized data sets, and low-dimensional structures that underlie our data. One of the common methods that are used to convert MTS into an updated coordinate space to find the main traits is the Principal Component Analysis (PCA). It is a dimensionality reduction method that has another important function which is to measure multi-time series with genuine features such as high dimensionality and similarity measure, which increase the clustering progress complexity more than univariate time series; see Olive (2017), Salem and Hussein (2019), and Chen et al. (2021).

This algorithm which identifies the groupings of data sets with similar properties, is one of the most well-known ones used for clustering time series data. The fundamental theory on which K-means clustering is theoretically based on the idea of partitioning objects into K clusters with the aim that the distance within a cluster is reduced to the minimum. Additionally, it is also the choice of how many initial cluster centers to use that impacts the quality of results; see Li (2019). Nevertheless, one of the expected problems in most areas of applied statistics is the strategy of identifying groups of similar time series in a panel of time series. However, it is not a usual process to extend the distance-based clustering methods to time series data, where the identification of an appropriate distance measure is undeniably a difficult step. Nevertheless, most of such clustering methods, used frequently in current studies and pieces of research, can, unfortunately, work only with fixed-dimensional demonstrations of data patterns. Thus, clustering time series is a serious research problem that is used in numerous applications in variable fields and has recently attracted a large amount of research. In the current paper, the researcher has made the central model of the investigation of the clustering data resulting from autoregressive moving average (ARMA), using k-means and algorithms with the Euclidean distance between estimated model parameters; see Cao et al. (2015).

The main purpose of the current study provides the users with a practical technique that can genuinely guide the investigation to accurate results that match perfectly their needs. Therefore, the researcher has clustered certain Saudi cities whose similar qualities have been specified by the Saudi ministry of health care during the COVID-19 pandemic. The study could be concerned with listing representatives for homogeneous groups (data reduction), recording "clusters" with a detailed description of their unknown properties (data types), numbering useful and suitable groupings(Clusters), or at least documenting unusual data objects (outlier detection). Consequently, we have examined our model with a COVID-19 real dataset to stress the effective role of our model with the existing models. In the following sections, several related works have been discussed and represented, including our proposed methodology.

2. RELATED WORK

To more effectively locate the first cluster centroids, a number of methods have been suggested. In the Encyclopedia of Ecology, Craig Syms, Principal Components Analysis, its main function is to display the relative locations of data points in fewer dimensions while preserving the most significant amounts of information that can explore relationships between dependent variables. Olive (2017) used classical PCA in explaining the reduction and concentration structure with a few linear uncorrelated combinations of the original variables. He implemented PCA in data reduction and interpretation. In relation to the Walk-Random hypothesis, previously rejected in many previous studies, the current study showed a decrease in the time-series correlation at least in part of the Japanese stock market - which means that the models that Dependent on that hypothesis have become more important (Cao et al., 2015).

Baudry et al. (2010) study model-based clustering whose design is constructed on a fixed combination of allotments, in which each mixture element is considered to be significantly related to another set, cluster, or subpopulation. For continuous data, the most frequent component distribution is a multivariate Gaussian (or normal) distribution. The used methodology which is standard for

model-based clustering includes using the EM algorithm to calculate the specific combination models relating to each number of clusters considered and using BIC to choose the number of combination elements, presumed to equalize the number of clusters. Consequently, the clustering is eventually done by assigning each observation to the cluster to which it is most likely to belong a posterior. This is strictly conditioned only on the selected sample and its predicted parameters (Cao et al., 2015).

3. MATERIAL AND METHODS

The basic notion regarding PCA is to specify designs and associations among several characteristics in the dataset. Initially, PCA's previous function is to decrease the parameters of the data without affecting the significant recorded data which can be easily recollected. PCA can be performed by doing simple sequences of stages through which the principal elements are the updated group of variables that are recorded through the first set of variables. The principal elements are calculated in these recently recorded variables which are not only extremely important but individually impartial from each other. Then the process of regularization of the recorded data is all about collecting the required data through a specified methodology in which all the variables and their values are gathered within a homogenous limit. After that, all the variables in the recorded data are categorized objectively through a neutral and recognized scale; see Li (2014).

In the second stage which includes both calculating and drawing the covariance matrix, PCA plays an important role to denote the correlation and dependencies among the characteristics recorded in the data set. This role is ultimately crucial to denote all the distinctively dependent variables because they show sets of information that are not neutral, biased, and insignificant. This subsequently, decreases the general function of the sample. From a mathematical point of view, a covariance matrix is a K × K matrix, where K shows the average of the data set (Li, 2014). Yet, in a case where the researcher has a set of 2-Dimensional data with variables a and b, the covariance matrix is a 2×2 matrix as shown below:

$$\begin{bmatrix} \mathrm{Cov}(c,c) & \mathrm{Cov}(c,d) \\ \mathrm{Cov}(d,c) & \mathrm{Cov}(d,d) \end{bmatrix},$$

Cov (c, c): covariance of a variable with itself, Cov (c, d): Covariance of the variable 'c' with to the variable 'd'. Also, covariance is commutative, Cov (c, d) = Cov (d, c).

The final value of covariance denotes how two variables that are co-dependent, are interrelated to each other. If this value is negative, it states the fact that the variables are indirectly interrelating with each other through an indirect relationship. Yet, if it is a positive one, it will specify the opposite fact that the two variables are directly interrelated with each other. The steps to find the covariance matrix:

Step 1: Computing both the Eigenvectors and Eigenvalues that are regarded as the mathematical values that have to be calculated from the covariance matrix with the object to decide the main elements of the data set. It is generally understood that these two mathematical formulas are constantly estimated as a pair, meaning that with each eigenvector, there will be always an eigenvalue. Thus, with a 2-Dimensional data set, the researcher will measure 2 eigenvectors (of course with their relevant eigenvalues). The concept on which eigenvectors are theorized lies in the fact that the main function of using the Covariance matrix is to grasp and observe the most amount of variance within the data set. This is simply because, with the most amount of variance, there will be always the most valuable information regarding this data set.

Step 2: Calculating the main elements is based on the process of calculation of the eigenvectors and eigenvalues, in which the noteworthy observation and value is the eigenvector with the highest

value of the eigenvalue. Thus, the researcher can easily remove all the main elements beneath that value aiming at decreasing the dimensionality of the dataset.

Step 3: the role of PCA in decreasing the Dimensions of the Dataset is to re-categorize the initial data with the final principal elements, which should represent the greatest and the most important statistics of the data set. However, to substitute the initial data axis with the updated formed principal elements, the researcher needs only to make the mathematical multiplication of both the transpose of the initial data set with the transpose of the recorded feature vector [1, 7].

3.1 Different Models of Multi -Time Series

Time series analysis requires the formulation of a mathematical model representing the given series, which has been developed. Specialists have several mathematical models linking the values of observations, and the values of different components of the time series.

Among the most prominent mathematical models that describe the time series are the additive, multiplicative and mixed models, given by

- Collective model: $Y_t = T_t + S_t + C_t + I_t$,
- Tax model: $Y_t = T_t \times S_t \times C_t \times I_t$,
- Mixed model: $Y_t = T_t + S_t \times C_t \times I_t$,

where, T_t is the general drift, S_t is occasional changes, C_t is cyclical changes, and I_t is random changes. And time series are analyzed to isolate the regular and irregular influences, and to know the extent of the effect. Each of them is based on the value of the observed phenomenon. Thus, the goal of the investigation is to return the overall value of the phenomenon to its constituent elements; see Dozie and Ijomah (2020).

The existence of time-series compounds can be detected by analyzing the information graphically, the general trend is represented in that vehicle that pushes the curve of the series evolved over time up or down, while the periodic component is reflected in the graph in the form of peaks or troughs on a regular basis, allowing us to determine the period of this occurrence the phenomenon, as for the random variable, is the fluctuation occurring at the level of the chain. As for the seasonal variable, it is evident through the regularity in recording a value on the last quarter of each year or a decrease at the start of a new year.

3.2 Autocorrelation Function (ACF):

It is a measure that measures the internal correlation strength of the time series, where the static time series can be, distinguished from the non-static through the values of the autocorrelation coefficients. The ACF is an important means of knowing the stability of the time series, as it is it tends to slow down rapidly towards zero with increasing displacement periods (k) or breaks off after a number of periods of displacement (q = k) i.e.:

$$\rho_k = 0; \forall k > q,$$

where, ρ_k is the autocorrelation function for ACF. Since the sample autocorrelation function is only estimate of the self-correlations, its values are from it is likely to be small and not zero, meaning that:

$$r_k \neq 0; \forall k > q,$$

r_k is estimates of the autocorrelations. But if the time series is unstable due to an upward or downward trend in the average, the ACF function of the sample does not interrupt and does not slowly descend towards zero, because the observations tend to be in the same direction as the arithmetic mean of the time series for several time periods, and as a result, we obtain large correlations at long displacement periods; Jebb et al. (2015).

3.3 Self-Regression Model (AR):

The autoregressive model represents the relationship between the current values and the preceding estimates that the time series has achieved and is used in various fields, including the description of a specific phenomenon, whether that is natural or economic. When the current value of the series is a function of its value in the previous period in addition to some errors, the models formed from this process are called autoregressive models.

If y_t represents the current value of the time series $x_{(t-1)}, x_{(t-2)}, \ldots x_{(t-n)}$ the values of the same series in previous periods, and it was found that y_t depends on or is affected by its previous value, so we can express this relationship with a self-regression model of rank p, as follows:

$$x_t = \mu + \varnothing_1 x_{t-1} + \varnothing_2 x_{t-2} + \cdots + \varnothing_p x_{t-p} + \varepsilon_t ,$$

Or in the form of deviations:

$$z_t = \mu + \varnothing_1 z_{t-1} + \varnothing_2 z_{t-2} + \cdots + \varnothing_p z_{t-p} + \varepsilon_t ,$$

whereas: $\varnothing_1, \varnothing_2, \ldots, \varnothing_p$ are the parameters of the autoregressive model and ε_t is called the error term which is independent of previous z_t values, and follow a moderate-mean distribution with mean 0 and constant variance δ_a^2 .

3.4 Moving Averages Model (MA)

The time series whose value in time ***t*** can be obtained from random errors in the current period ε_t and previous periods $(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$, and the resulting model from this process is called the moving averages model. The general formula for this model of the rank q, symbolized by the symbol MA(q) is:

$$x_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} ,$$

whereas: x_t is a dependent variable, expressing the estimating value of *y* at the time *t*, $\theta_1, \theta_2, \ldots, \theta_q$ is the mean of the dependent variable *y* (constant term) estimated model, μ is a parameter, $(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q})$ is an error in preceding time periods *t*, ε_t is the random error at time *t* which is not explained by the model variables. In another form of deviation, we have

$$z_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q}.$$

The rank of the model q is determined by its autocorrelation function, as it discontinues after a period of q, while the partial autocorrelation function gradually decreases in a descending curve.

3.5 Self-Regression and Integrative Moving Averages (ARIMA) Model

ARIMA models are models generated from autoregressive models and averages models moving, after taking appropriate differences to make the time series stable, and these models are symbolized by (p, d, q) ARIMA, where

p: indicates the degree of self-regression.
d: Refers to the number of differences needed to make the time series stable.
q: denotes the score of the moving averages.

For more detail, we refer to Niennattrakul and Ratanamahatana (2007) and Ansari and Ahmed (2001). The ARIMA model is one of the most popular models for predicting the values of economic variables such as the prices of some commodities, inflation, and products, among others. It is also used in the analysis of non-economic variables such as the evolution of some diseases over time. When the time series is not stable, it must first be converted to a stable time series before constructing the mathematical model, by taking the differences d or using one of the transformations and the number of differences required to convert the series into a stable series called the "degree of integration" as it transforms from ARMA (p, q) to ARIMA model (p, d, q) is as follows:

$$\varnothing_p\left(B\right)y_t = \mu + \theta_q\left(B\right)\varepsilon_t,$$

or in terms of deviation, we have

$$\varnothing_p\left(B\right)w_t = \theta_q\left(B\right)\varepsilon_t,$$

where

$$y_t = \left(1-B\right)^d z_t,$$
$$\varnothing_p\left(B\right) = 1 - \varnothing_1 B - \varnothing_2 B^2 - \cdots - \varnothing_p B^p,$$
$$\theta_p\left(B\right) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p,$$

and

$$w_t = \begin{cases} \nabla^d Y_t = \left(1-B\right)^d y_t, & d>0 \\ Y_t, & d=0 \end{cases}.$$

For further detail, please see Ratanamahatana (2005) and Al-Aziz and Alotaibi (2019). Therefore, ARIMA models can be considered stable ARMA models with different ranks.

3.7 Cluster Analysis

A collection of methods known as a cluster analysis is used to create groups (clusters) from objects with multiple variations of data. The objective is to create smaller groups with consistent characteristics that stand out from bigger groups with more variety. In the same way that the differences and variances between them as independent groups are maximized, the established groups or clusters should be regularized as much as feasible.

3.8 K-Means Clustering Time Series

There are overall techniques and strategies of time series clustering that can be categorized as being Non-Hierarchical (K-means clustering). So, any given data of any Time-series should be with a specific significance and interest due to their ubiquity. This can be applied to all the various fields of knowledge including natural science like physics, chemistry, and biology, mathematical sciences including algebra and engineering, the commercial business like financing, banking, accounting, and economics. They can be also applied to the field of health care and sanitation and governmental programs. In all these respects, each time series gives various worthy pieces of information and observation. Furthermore, the k-means method is not only a Non-hierarchical clustering method as with its simple tracts and steps, but it also involves a cautious mixture of both k and the original centers to escape any sort of limited minimization of its variation that will cause biased categorization of its elements. This is clearly illustrated in complex objects and phenomena that can help the researcher to figure out peculiar patterns in time-series datasets.

3.9 The Stages of Applying ARIMA Models in Forecasting

The first stage to build an ARIMA model for time series is data initialization. So if the data is stable by drawing data, partial and self-correlation, then prepare the data for recognition, but if the series is not stable in the mean and variance when instability in the mean is addressed by taking the first difference (d= 1), if it does not stabilize, take the second difference (d = 2) and it often stabilizes between the first or second difference. As for instability in the variance. It is addressed through an appropriate conversion procedure for the data and the logarithmic transformation and the square root transformation are among the most used transformations for contrast [13].

After achieving the stability of the time series. The process of model identification begins, and using data about how the time series is generated, then getting the value of p, d, and q that is needed in the ARIMA general linear model then obtaining preliminary estimates of the model parameters. The optimal forecasting is when found upcoming estimates of the time series by using the appropriate sample obtained under the previous stages and obtaining an estimate in which the resulting error is very small and the variance is minimal.

To ensure the correctness of the model rank (q, d, p) ARIMA, a set of statistical criteria has been adopted that help in the comparison between the candidate models, where the best model is chosen that has the lowest value for these criteria from these criteria:

Akaike Information Criterion (AIC): This method was proposed by scientist Akaike (1979), defined:

$$AIC\left(M\right) = -2\left(Conditonal\ \ Maximum\ \ Likelihood\right) + 2$$

If the model has M parameters according to the data, then the formula for the AIC standard in terms of the amount of error variance is as follows:

$$\mathrm{AIC}\left(\mathrm{M}\right) = \mathrm{n}\left(\mathrm{Ln}\left(\tilde{\mathrm{A}}\hat{}_{\mathrm{a}}^{2}\right)\right) + 2\mathrm{M},$$

where M represents the whole value of model parameters, *n* is the value of views, and $\sigma\hat{}_{a}^{2}$ is the amount of error variance.

Bayesian Information Criterion (BIC): The formula for computing BIC is:

$$\mathrm{BIC}\left(\mathrm{M}\right)=\mathrm{n}\left(\ln\left(\tilde{\mathrm{A}}\hat{}_{\mathrm{a}}^{2}\right)-\left(\mathrm{n}-\mathrm{M}\right)\right)\ln(1-\frac{\mathrm{M}}{\mathrm{n}})+\mathrm{M}\left(\ln\left(\mathrm{n}\right)\right)+\mathrm{M}\left(\ln\left[\left(\frac{\tilde{\mathrm{A}}_{\mathrm{x}}^{2}}{\tilde{\mathrm{A}}_{\mathrm{a}}^{2}}-1\right)\Big/\mathrm{M}\right]\right),$$

where $\hat{\sigma}_{x}^{2}$ is the amount of variation of the series, and $\sigma\hat{}_{a}^{2}$ is the amount of error variance and calculated as follows: $\sigma\hat{}_{a}^{2}=\sum_{t=1}^{n}\frac{\left(x_{t}-\hat{x}_{t}\right)^{2}}{\left(n-p\right)}$. After neglecting some of the terms, the final form can be shown as follows:

$$\mathrm{BIC}\left(\mathrm{M}\right)=\mathrm{n}\left(\ln\left(\tilde{\mathrm{A}}_{\mathrm{a}}^{2}\right)\right)+\mathrm{M}\left(\ln\left(\mathrm{N}\right)\right),$$

where N in the previous formula is the series views value and at the model parameters final value; see Wang et al. (2022).

If the researcher is given specific data that will be referred to as $\mathrm{S}_{\mathrm{n}}=\left\{\mathrm{X}_{1},\ldots,\mathrm{X}_{\mathrm{n}}\right\}$, the function of the K-Means algorithm is to calculate K cluster centers $¼_{1},\ldots,¼_{\mathrm{k}}$ (K is given) and, for each point X_{i}, specify the task to one of the K clusters.

Cluster value is denoted by means of vectors $\left(\mathrm{r}_{1},\ldots,\mathrm{r}_{\mathrm{n}}\right)$ where each r_{i} is a vector of size K

$r_{i}=\left(0,0,\ldots,1,\ldots,0,0\right)T$, for $i=1,\ldots,n$,

$r_{i}\left(k\right)=1,$ if and only if X_{i} belongs to cluster k, for $k=1,\ldots,k.$ (Each point can belong to only one cluster.) we might have $r_{1}=\left(1,0,0\right),\ r_{2}=\left(0,0,1\right),\ r_{3}=\left(1,0,0\right),\ r_{4}=\left(0,1,0\right)$, in which case X_{1}, X_{3} are assigned to cluster 1, X_{2} is assigned to cluster 3, while X_{4} is assigned to cluster 2.

The K-means algorithm requests the vectors $\left\{\mu_{i}\right\}_{i=1}^{n}$ and $\left\{r_{i}\right\}_{k=1}^{k}$ that reduce a score to the minimum supported by the regularization of the total value of the specific distances between every point and its corresponding centers

$$J=\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{k}r_{i}\left(k\right)x_{i}-\mu_{j}{}^{2}.$$

There will be two optimizations at each step in the final solution that can be simplified as follows:

1. Keeping the present values $\left\{\mu_k\right\}_{k=1}^{k}$ unchanged, find the values $\left\{r_i\right\}_{i=1}^{n}$ that reduce the score to the minimum J ("E-Step").
2. Keeping the final values $\left\{r_i\right\}_{i=1}^{n}$ unchanged, find the values $\left\{\mu_k\right\}_{k=1}^{k}$ that reduce the score to the minimum J ("E-Step").
   J ("M-Step").

In the "E-step" with the current values $\left\{\mu_k\right\}_{k=1}^{k}$ fixed, the values $\left\{r_i\right\}_{i=1}^{n}$ that minimize J can found by inspection:

$$r_i\left(k\right)=\begin{cases}1, if\ \ k=\arg\min_{i=1,..,k}\ x_i-\mu_j^{\ 2}\\ 0, \qquad\qquad\qquad\qquad ,otherwise\end{cases}.$$

For $i=1,...,n$. In other words, we simply assign each point to the closest cluster mean. In the "M-step", the values $\left\{\mu_k\right\}_{k=1}^{k}$ that minimize J can be found by simple differentiation

$$\frac{\partial j}{\partial \mu_k}=2\sum_{i=1}^{n}r_i\left(k\right)X_i-\mu_k=0,$$

which gives

$$\mu_k=\frac{\sum_{i=1}^{n}r_i\left(k\right)x_i}{\sum_{i=1}^{n}r_i\left(k\right)}.$$

3.9.1 Euclidean Distances and L_p Norms

The Euclidean distance measure is considered one of the most straightforward comparable estimations for time series. The researcher can initially presuppose that both time series are equal in length, referred to as n. thus, both can be viewed as specific points in n-dimensional Euclidean space. So, they can be easily defined all the possible variations between sequences C and Q and $D\left(C,Q\right)=Lp\left(C,Q\right)$, i.e., the gap between the two points calculated by Lp norm (when p = 2, it reduces to the familiar Euclidean distance). This denotes a clear illustration of the fundamental theory behind the Euclidean distance metric.

More formally, let *μ(C)* and σ(C) be the mean and standard deviation of $C=\left\{c_1,...,c_n\right\}$ the sequence C is replaced by the normalized sequences *C*, where:

$$C_i^1=\frac{c_i-\mu\left(C\right)}{\tilde{\mathrm{A}}\left(C\right)}.$$

Even after regularization, the Euclidean as a tool of distance calculation is in most cases not functional for some time series domains. This is due to the fact that the method does not have the flexibility to permit any sort of quickening and slowing along the time axis.

3.9.2 Dynamic Time Warping (DTW)

It is a path-searching algorithm, it finds the minimum cost path between the complete matrixes of pairwise spaces one time series and another one. We will label X and Y. Define $X \coloneqq \left(x_1, x_2, \ldots, x_n\right)$ and $Y \coloneqq \left(y_1, y_2, \ldots, y_n\right)$. This matrix of pairwise distances is referred to as the cost matrix C. Define the function $c\left(x, y\right)$ as the cost or local distance function between two points *x* and *y*. If $c\left(x, y\right) = 0$, the two points are identical. Therefore, low cost implies similarity, high cost implies dissimilarity. DTW finds a path through the cost matrix of the minimum total cost. Each valid path through the cost matrix is called a "warping" path. The set of all warping paths is labeled W. This recursive function gives the minimum cost path:

$$\gamma\left(i, j\right) = d\left(q_i, c_i\right) + \min\{\gamma\left(i-1, j-1\right), \gamma\left(i-1, j\right), \gamma\left(i, j-1\right).$$

To describe the background theoretical framework of the DTW algorithm, it is known that the method has gained most of its solid grounds through its peculiar capability in categorizing the time-series homogenous estimations, which can reduce the minimum level of instability and biased categorization in time through permitting some sort of "elastic" conversion of time series to spot comparable figures with different segments. Consequently, when having two-time series exemplified by the series values (or curves represented by the sequences of vertices) DTW gives the ideal answer in the O (MN) time which could be developed in advanced steps through applying new procedures such as multi-scaling. The sole limit that can affect the data sequences is the requirement of the data set to be tested at intermediate points in time (however, this problem is not without a practical solution, as it can be handled through re-sampling). If the values of all the sequences are specified by some feature space Φ than to compare two different sequences $X, Y \in \Phi$ one needs to use the local distance measure which is defined to be a function:

$$d : \Phi \times \Phi \rightarrow R \geq 0.$$

Intuitively d has a small value when sequences are a similar and large value if they are different. Since the Dynamic Programming algorithm lies in the core of DTW it is common to call this distance function the "cost function" and the task of optimal alignment of the sequences becomes the task. By arranging, all sequence points by minimizing the cost function (or distance), the algorithm starts by building the matrix distance C Î R, N×M representing, all pairwise distances between X and Y; see Safarineja et al. (2009), Ambigavathi and Sridharan (2020), and Wan et al. (2021).

4. THE FRAMEWORK

The practical application of Covid-19 data in Saudi Arabia over the past four months for four cities in the Kingdom using Python to analyze the data.

4.1. Exploring the Dataset

We have the Covid-19 data series in Saudi Arabia from the Ministry of Health account for four-time series, these are Riyadh, Jeddah, Makkah, and Dammam for the months of July, August, September, and October.

Table 1 shows the data frame for case count in cities, note that the number of rows is 124 and the columns are 9, where the case count for 4 cities is: cases1 for July, cases2 for August, cases3 for September, and cases4 for October.

Table 1. Dataset for PCA

	Case 4	Date 4	Case 3	Date 3	Case 2	Date 2	Case 1	Date 1	
Jeddah	30	1/10/2022	72	1/9/2022	31	1/8/2022	172	1/7/2022	0
Jeddah	40	2/10/2022	43	2/9/2022	72	2/8/2022	164	2/7/2022	1
Jeddah	9	3/10/2022	31	3/9/2022	50	3/8/2022	169	3/7/2022	2
Jeddah	11	4/10/2022	48	4/9/2022	40	4/8/2022	169	4/7/2022	3
Jeddah	11	5/10/2022	65	5/9/2022	49	5/8/2022	149	5/7/2022	4
…	…	…	…	…	…	…		…	...
Dammam	15	27/10/2022	21	27/9/2022	19	27/8/2022	79	27/7/2022	119
Dammam	15	28/10/2022	24	28/9/2022	24	28/8/2022	67	28/7/2022	120
Dammam	11	29/10/2022	18	29/9/2022	16	29/8/2022	70	29/7/2022	121
Dammam	2	30/10/2022	21	30/9/2022	20	30/8/2022	0	30/7/2022	122
Dammam	5	31/10/2022	0	31/9/2022	21	31/8/2022	0	31/7/2022	123

In Table 2, statistical data is presented for different cases. Where the number is 124, and the highest value of the median and standard deviation of cases1. The lowest value is case 4, where (25%, 50%, and 75%), 25% of them are the highest case1. Moreover, the least cases4, the minimum is case 4 for the month of October, and the maximum number is case1 for the month of July.

4.2. Create a Covariance Matrix Data

The data in Table 1 contain 4 variables: "Cases of Riyadh", "Cases of Jeddah", "Cases of Makkah", and "Cases of Dammam". After implementing the covariance matrix on the data, we noticed that the values are large. Next, we obtained a visual representation of the covariance matrix.

We noticed that the matrix values after derivation became larger than the previous values. Next, we got a visual representation of the covariance matrix after derivation: The values in Figure 1 are quite large after differentiation, and they are all positive and we got a covariance matrix graph.

In Figure 2, the horizontal axis shows the days number. While the vertical axis shows the Covid-19 cases number, as the number of infections is high in the city of Dammam in the first 20 days, as it reached more than 450 cases of the spread of Covid-19, represented by the red color for the period (20-40) days, the city of Riyadh recorded the highest cases of Covid-19 as the number of infections

Table 2. The data of statistical

	Case 1	Case 2	Case 3	Case 4
count	124.000000	124.000000	124.000000	124.000000
mean	167.822581	66.83871	43.306452	23.080645
std	89.771168	47.34296	17.466892	14.049041
min	0.0000000	0.000000	0.0000000	2.000000
25%	102.00000	43.00000	32.000000	10.00000
50%	148.00000	56.00000	43.000000	23.000000
75%	224.50000	75.25000	53.000000	34.250000
max	431.00000	281.0000	88.90000	59.000000

Figure 1. Represent the variance of the covariance matrix

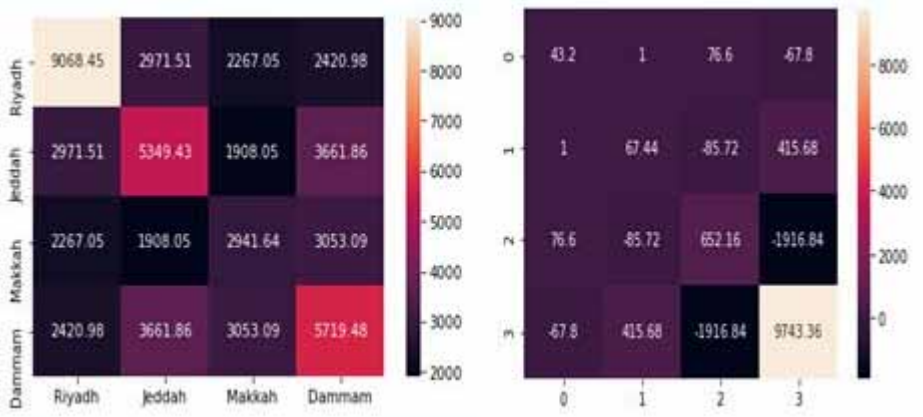
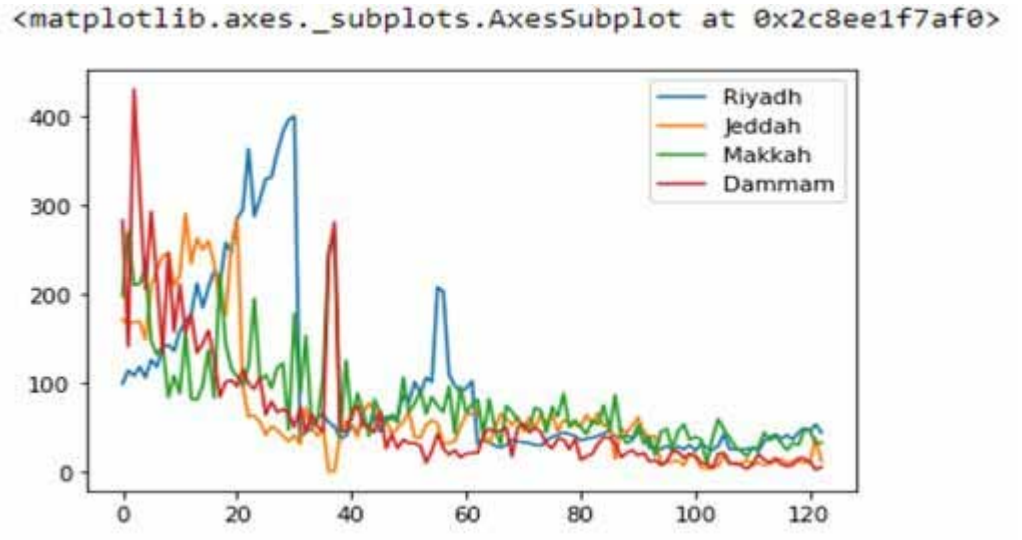


Figure 2. The graph of the covariance matrix



reached (400) infections and also in 60 days the number of Riyadh cases decreased to 200 cases, represented by the blue color, and between the period (60-120), the number of cases for all cities ranged from 100 to the least and were generally stable.

Figure 3 shows the enlarge in the number of cases of Covid-19 in the month of July as the number of infected cases reached 450 cases, and the lowest in the month of October as the cases of infection were reduced to 4 injuries.

4.4. Principal Component Analysis Model

We need the data center and measurement to scale our data, and we measure based on the data. To present our data by PCA for two main elements. Taking into consideration that the primary data consisted of 5 columns, 4 features, and 1 target column after we finished the PCA process, we only have 2 columns for features. The target data set is not changed. Thereupon, the researcher has attributed the target column to the updated set of key elements.

Figure 3. Visualization of the data

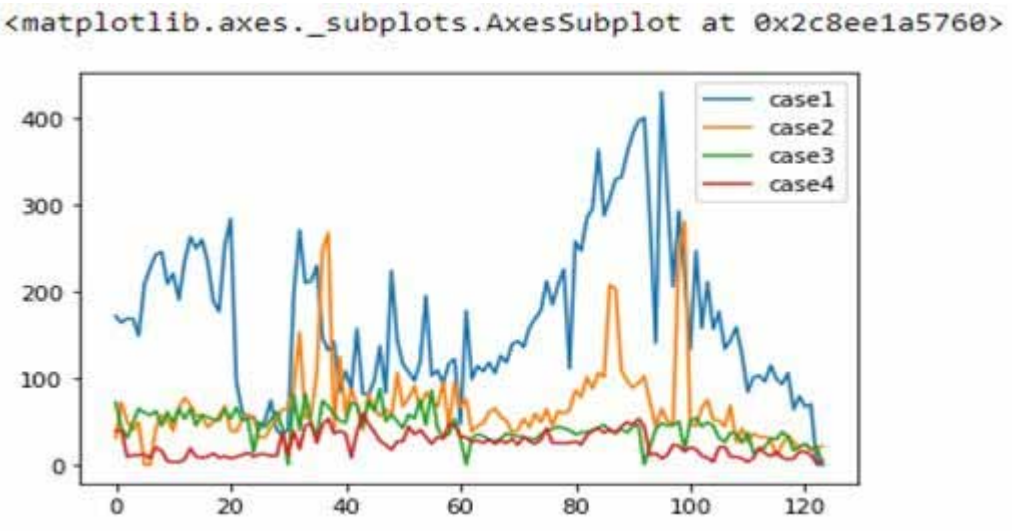


Table 3 provides a new dataset after performing PCA for all cities. In the following step, the researcher will explain, the peculiar way through which PCA plays a significant role in clarifying the data through simplified procedures and forms.

Figure 4 denotes that the original dimensional data suffers no risks when it will be lessened to 2 dimensions using PCA. This is due to the fact that the dataset can be illustrated through only two elements.

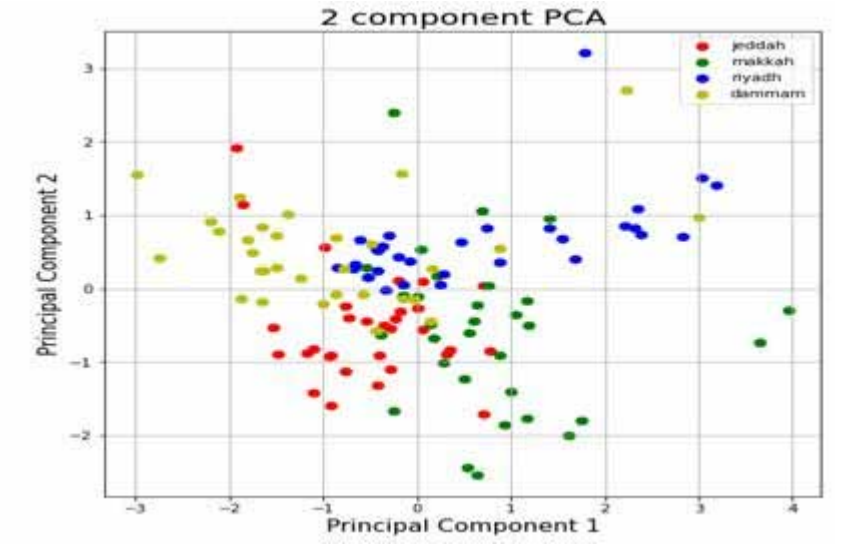
4.5. Multi- Time Series with ARIMA Time Series

We used Table 1 for the dataset of case counts for visualization in a multi-time series of our data as 4-line plots on a lone graph ("cases of Riyadh", "cases of Jeddah", "cases of Mecca", "cases of Dammam") we see that in the x-axis is the months.

Table 3. New dataset after performing PCA

	Principal Component 1	Principal Component 2	City
0	0.702306	-1.706907	Jeddah
1	0.704602	0.034754	Jeddah
2	-0.980006	0.568428	Jeddah
3	-0.726840	-0.391734	Jeddah
4	-0.420540	-1.312249	Jeddah
…	…	…	…
119	-1.811037	0.662015	Dammam
120	-1.760154	0.491794	Dammam
121	-2.109273	0.777924	Dammam
122	-2.744051	0.423253	Dammam
123	-2.986117	1.549685	Dammam
124 rows * 3 columns			

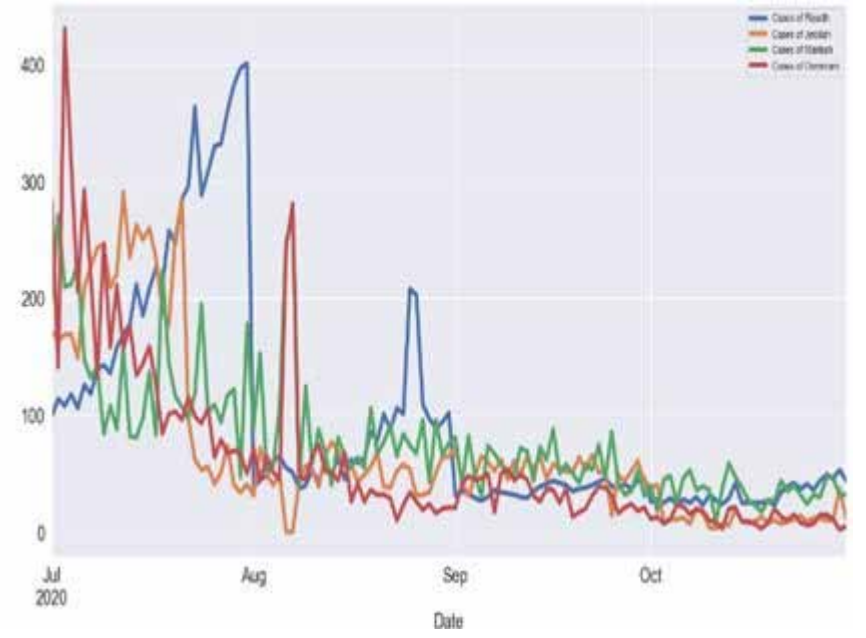
Figure 4. Dimensional data



Note that, this data is relative visualization in Multi-Time Series is same of in PCA Figure 5 the numbers represent search interest relative to the highest point on the chart for the given case count. A value of 400 is the highest cases count. A value of 200 is half the cases count. Likewise, a score of 0 is less than 1% as cases count.

Numerous methods for determining drifts in time series have been proposed by scholars to characterize patterns in time series. Using a rolling average as an example will show you the most

Figure 5. Relative visualization for Multi-Time Series



well-known case. This process emphasizes that the researcher should compute the median of all the points on each side of the chosen point for each time point. With the use of this average, we will be able to handle any such drifts without improper classification or documented instability. By identifying the drifts of "Cases of Riyadh," "Cases of Jeddah," "Cases of Makkah," and "Cases of Dammam" on a single figure, we may determine this.

In Figure 6, we have the trends that we are looking for, we have most of the cases compared to the previous plot. First-order differencing from Table 4, we compute the correlation coefficients of all of these time series.

Table 5 shows that the data sets are positively correlated. However, from looking at the times series, it looks as though their seasonal components would be positively correlated and their trends positively correlated also.

The researcher notes the first-order variation of these time series and then calculates the correlation of those because that will be the correlation of the case elements, roughly. The researcher should take into consideration that eliminating the drift may cause a noticeable correlation in the cases. So, the first-order differences.

There is a slight negative correlation when we observe the drift and the case elements; see Figure 7.

Autocorrelation is taken into the correlation of variables and correlation of time series, to spot the autocorrelation of the "Cases of Riyadh", "Cases of Jeddah", "Cases of Makkah", and "Cases of Dammam" series: on the x-axis, we have the lag, and on the y-axis, we have how correlated the time

Figure 6. The trends in the time series

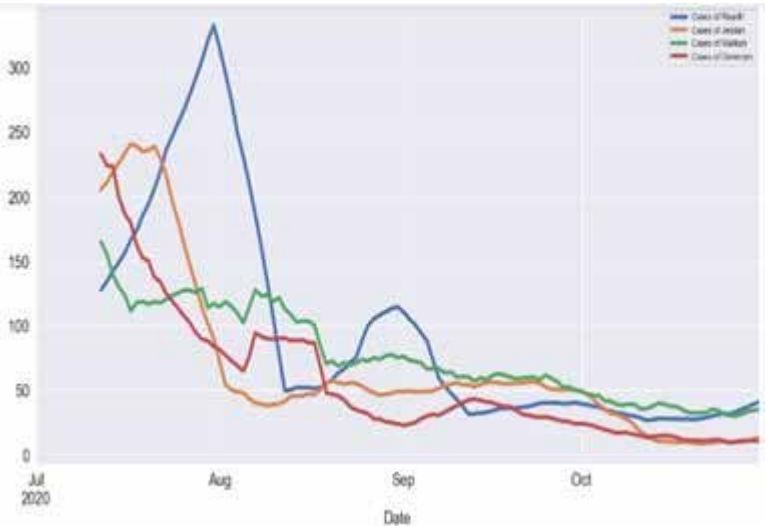


Table 4. Data of the correlation coefficients

Areas	Cases of Riyadh	Cases of Jeddah	Cases of Makkah	Cases of Dammam
Cases of Riyadh	1.000000	0.432905	0.438935	0.334882
Cases of Jeddah	0.432905	1.000000	0.480814	0.661881
Cases of Makkah	0.438935	0.480814	1.000000	0.744190
Cases of Dammam	0.334882	0.661881	0.744190	1.000000

Figure 7. The first-order differences

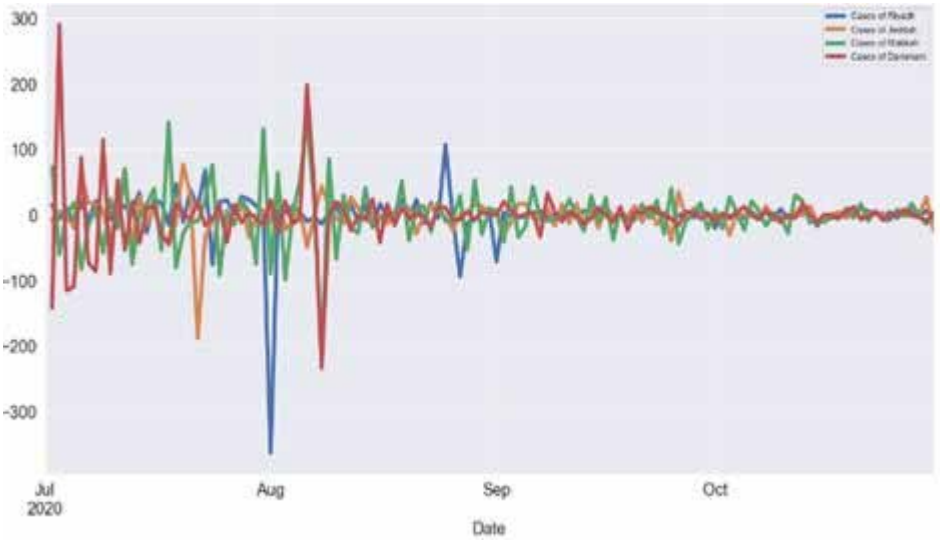


Table 5. The correlation constants to the first-order differences

Areas	Cases of Riyadh	Cases of Jeddah	Cases of Makkah	Cases of Dammam
Cases of Riyadh	1.000000	0.025817	0.082705	-0.044630
Cases of Jeddah	0.025817	1.000000	-0.126251	-0.119325
Cases of Makkah	0.82705	-0.126251	1.000000	0.053077
Cases of Dammam	-0.044630	-0.119325	0.053077	1.000000

series is with itself at that lag; see Table 5. So, this means that if the initial time series recaps every two days, the researcher would guess to observe a spike in the autocorrelation function at 2 days.

We looked at the graph and what we should expect to see here is a spike in the autocorrelation function at 4 months. It reveals that the time series correlation with itself has changed by 4 months.

If the researcher added more lags in the axes, he/she would see that it is 4-months at which we have this noticeable climax in correlation. There are other cases in an 8-month period, where it is also correlated with itself. There are other cases at 12. However, the more we move further away, the less, we can notice any sort of correlation. We have a correlation of itself with itself at a lag of 0. The dotted lines in the above figure explain the vital role of statistics in giving detailed information regarding that correlation. In this case, we can say that the "Cases of Riyadh" series is genuinely autocorrelated with a lag of 4 months.

From Figure 8, we have concluded the ARIMA Time Series, if we want to divide the data set into two sections, one for training and the other for testing, we can easily store the recorded data in the last month in the testing section.

Creating the ARIMA model is the simplest and most straightforward step in the entire process since the researcher has achieved all the required steps from the multi-time series. By doing such a step, the researcher can confidently calculate the ARIMA model, supply it in the data set, declaring the order of the required ARIMA model; see Table 6. The researcher will be able to see the summary of the model in the final output as well.

Figure 8. Autocorrelation and correlation of variables and correlation of time series for: (A) Cases of Riyadh, (B) Cases of Jeddah, (C) Cases of Makkah, and (D) Cases of Dammam

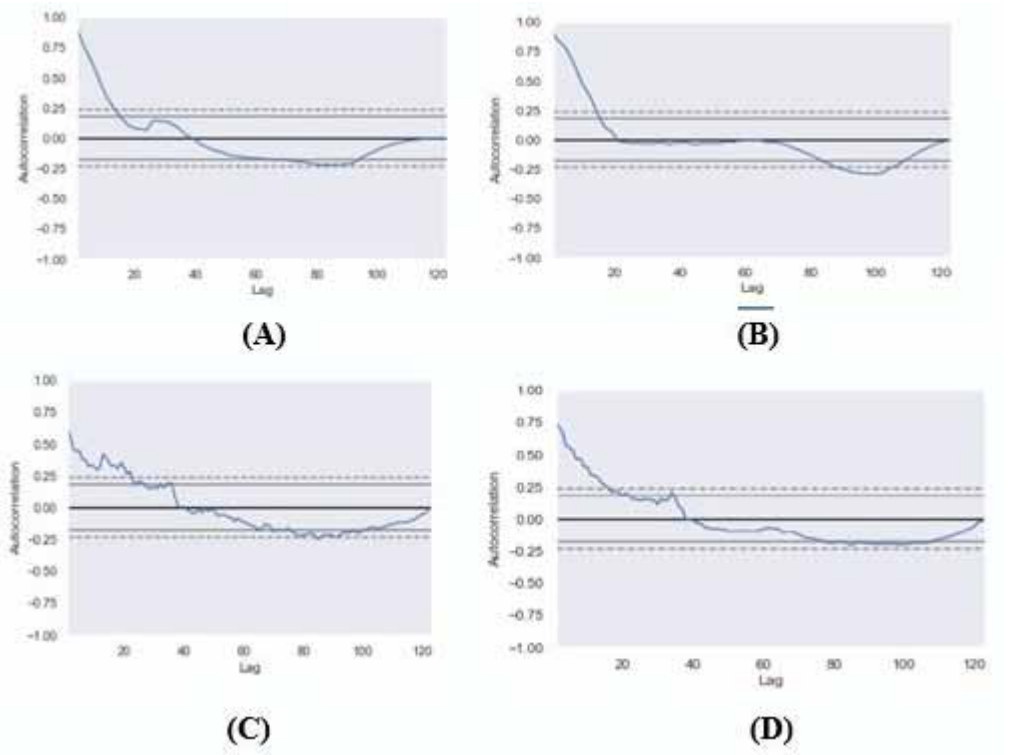


Table 6. The summary of the ARIMA models

AIC	BIC	Model:	Cases of cities
970.796	985.927	ARIMA (1.1.3)	Cases of Riyadh:
890.342	905.472	ARIMA (1.1.3)	Cases of Jeddah:
955.003	970.134	ARIMA (1.1.3)	Cases of Makkah:
975.547	990.678	ARIMA (1.1.3)	Cases of Dammam:

The ARIMA model (1.1.3) is the best model as it has smaller values of BIC and AIC. We have added the time series to extract all the forecasts in Figure 9.

Note that, all the time series extracted from them the forecasting that the line represents in purple is the forecast of the number of future cases below 10 cases. As we see the forecastings do a pretty good job of matching the actual trend despite having a certainly acceptable lag, as we have represented all the ARIMA characters in Figure 10.

Note that, ARIMA forecasting are represented by color (blue, green, purple, and pink), all the line purple and that the forecasting "Cases of Riyadh" is represented in orange color. The forecasting "Cases of Jeddah" is represented in red, the forecast "Cases of Makkah" is represented in brown and are the highest and close to 50 cases, and the forecast "Cases of Dammam" represented in gray color is the lowest and represents 3 cases.

Figure 9. All the forecast

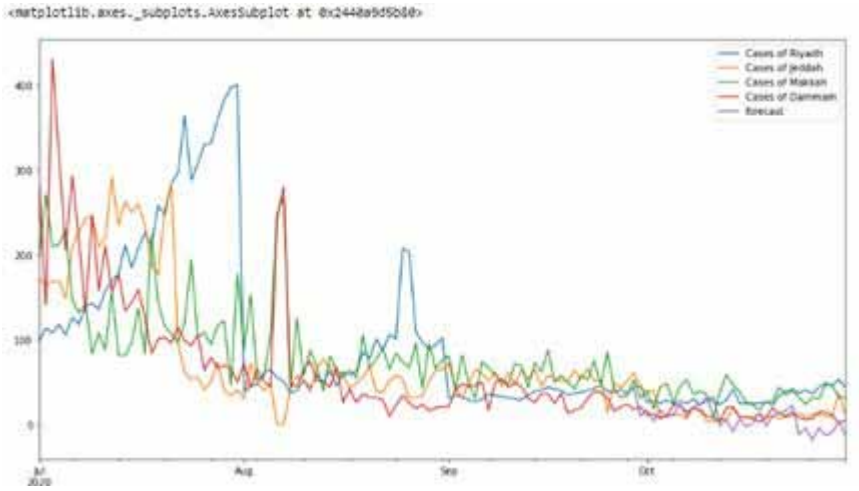
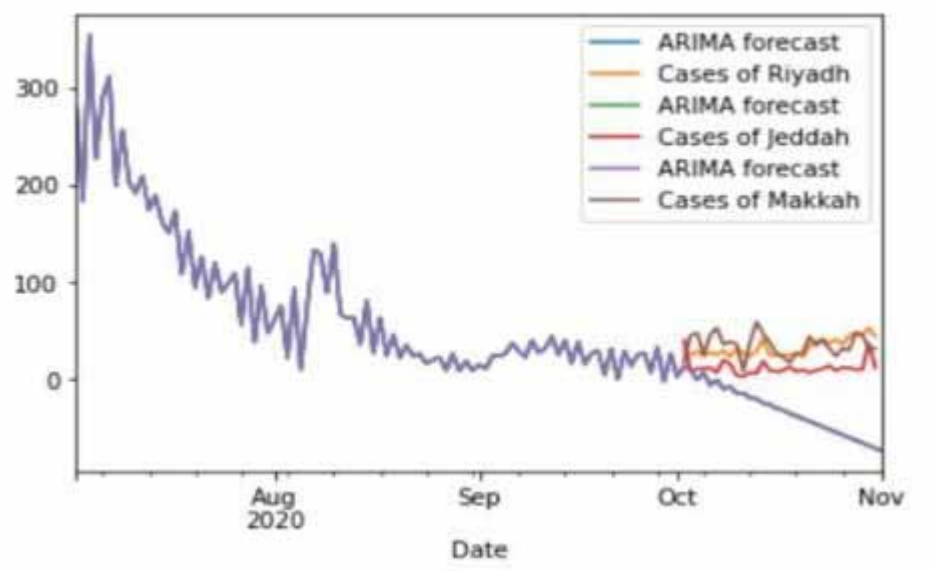


Figure 10. All the future forecasting (modified this figure)



Clustering time series: for implementing K-Means clustering we will work on the case count dataset of Table 1. We divided the data into two groups, the first group Riyadh and Jeddah, and the second group Makkah and Dammam, we took only two variables from the data.

Figure 11 of K-Means was about choosing the number of clusters (k) and selecting random centroids for each cluster.

We picked 3 clusters, and then select random observations from the data as the centroids; see Figure 12.

The black dots represent the three centroids of each group, we have chosen these points at random, thus these points are selected randomly in each update we may get different points and we defined some conditions for applying the K-Means Clustering algorithm for “Riyadh cases”, “Jeddah cases”

Figure 11. Visualize the data points for: (A) Cases of Riyadh, and Cases of Jeddah, (B) Cases of Makkah, and Cases of Dammam

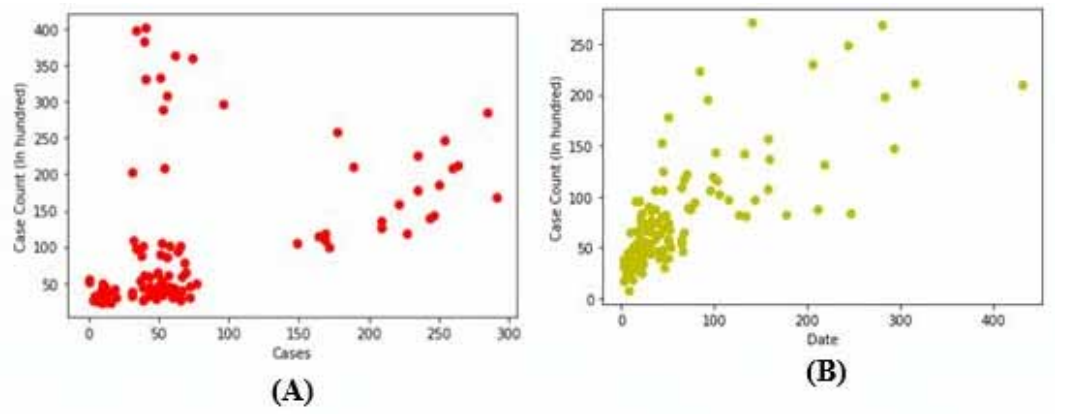
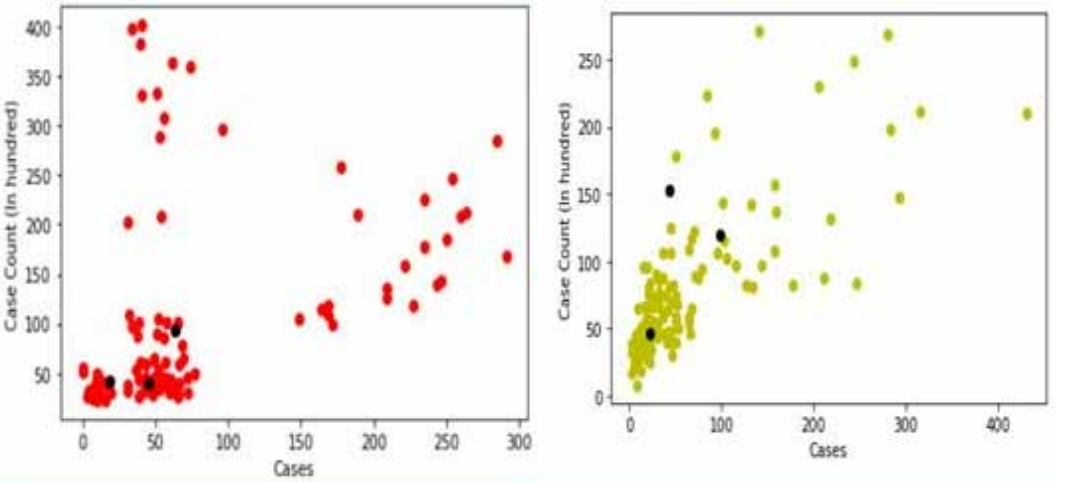


Figure 12. Centroids for visualize the data points



and for "Makkah cases" "Dammam cases" to find the Values of the difference between the centroids. We showed the visualization of three groups; see Figure 13.

4.6 Clustering from PCA

From Figure 14, we have a new dataset after PCA, which is a procedure we will use for clustering k-means on it, to create a K-means cluster with two clusters.

Here, 0, 1, 2, and 3 are used only to represent block identifiers and have no mathematical significance. We saw what the algorithm centroid values generated for the final groups; the output will be a 2D array of shape 2×4.

First row in figure15 shows estimations for the coordinates of the first centroid i.e.(0.59672778, -1.23419116), the second row shows the values for the coordinates of the next (2.43773215, 0.99210534), the third row shows values for the coordinates of the next (-1.68360581, 0.70132128), and the fourth-row details the values for the coordinates of the other centroid (-0.26267505, -0.0271317). In short, the algorithm is perfect and shows accurate results. Then plot the data points back on the diagram illustrating the clustering of data, for detail please see Figure 16.

Figure 13. Visualize three clusters

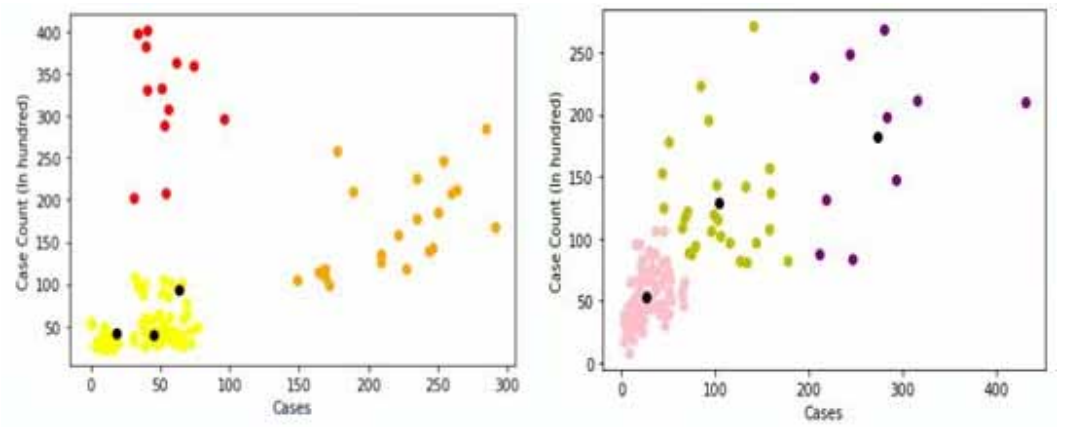


Figure 14. The labels for the data points

```
array([0, 3, 2, 3, 0, 3, 3, 0, 3, 3, 3, 0, 3, 0, 3, 3, 3, 3, 3, 0, 3, 0,
       3, 3, 2, 3, 2, 3, 3, 3, 2, 0, 1, 0, 0, 3, 1, 1, 0, 0, 3, 0, 0, 0,
       0, 0, 0, 3, 0, 3, 3, 0, 0, 0, 3, 0, 3, 3, 3, 0, 3, 2, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 1, 1, 3, 3, 3, 3, 3, 3, 2, 3, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
```

Figure 15. Values for the coordinates

```
array([[ 0.59672778, -1.23419116],
       [ 2.43773215,  0.99210534],
       [-1.68360581,  0.70132128],
       [-0.26267505, -0.0271317 ]])
```

As expected, in the fourth cluster, in the colors (blue, red, yellow, and green), we saw that the clusters that are close to each other have been clustered together. We have illustrated the visualization of four groups, we plot the data points in colors (blue, red, yellow, green) while the centroids are in black; see Figure 17.

We have four clusters that appear in colors (blue, red, yellow, and green), (4) centers appear with black dots, the first point appears in blue, the second in red, the third in green, and the last in yellow. Dynamic Time Warping: We implemented the DTW algorithm, and we divided the data into two groups, the first is "Cases of Riyadh and Jeddah", and the second is "Cases of Makkah and Dammam".

The x-axis represents the number of days, and the y-axis represents the number of cases in Figure 18. From 0 to 20 days, Jeddah cases were the highest, and nearly 300 cases arrived, and from 20 to 40 days, the cases of Riyadh were the highest and they arrived at about 400 cases, after 40 to

Figure 16. Visualize the data clustering

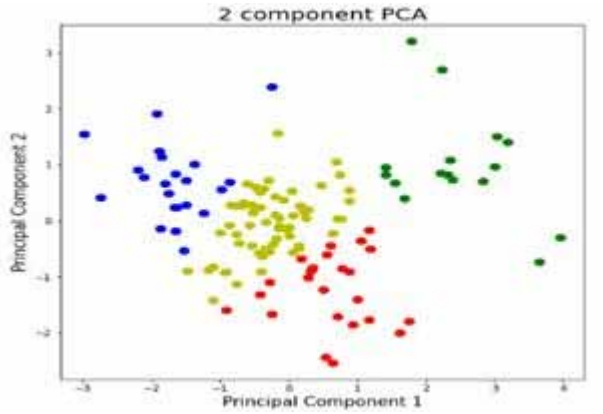


Figure 17. Visualize of the centroids for the data clustering

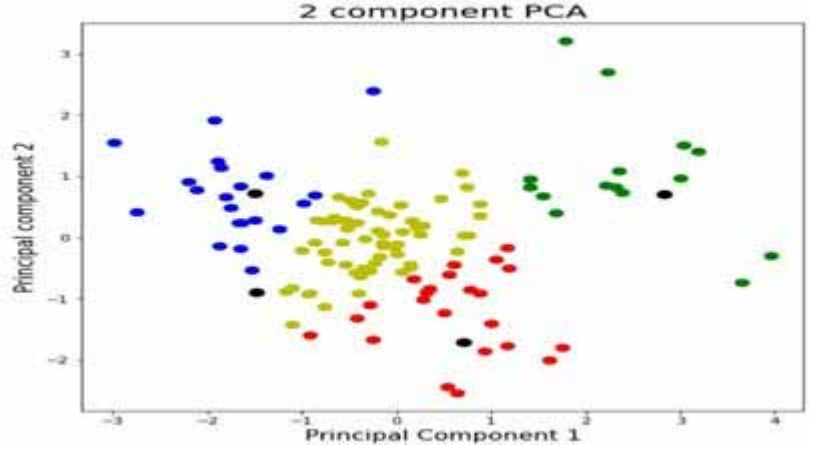
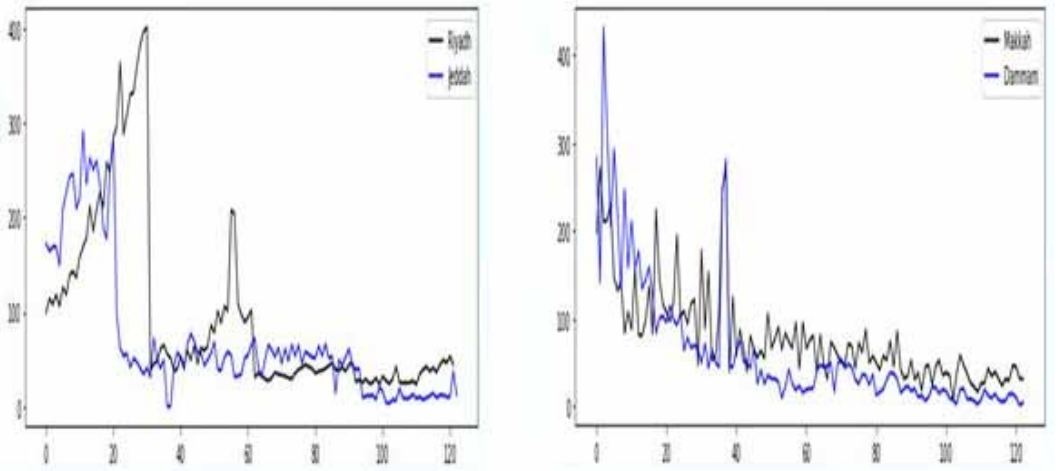


Figure 18. Visualize of the centroids for the data clustering



60 days, the decline in Riyadh cases reached almost 200 cases, but from 60 to 120 days, the cases varied between 10 to 100 cases.

From 0 to 20 days, Dammam cases were the highest, and nearly 400 cases arrived, and from 20 to 40 days, the highest cases were roughly equal between Makkah and Dammam, and about 250 cases arrived, and from 40 to 120 days the cases decreased and varied between 10 to 100 cases.

We computed the cost matrix A using Euclidean distance as a measure of local cost for Riyadh and Jeddah, and for Makkah and Dammam (which arrives at the absolute value in the one-dimensional case). Using dynamic programming, we computed the accumulated cost matrix B for Riyadh and Jeddah, and for Makkah and Dammam, which yields the DTW distance DTW (X, Y). We derived the optimum torsion path P $*$ using regression, for Riyadh, Jeddah, Mecca, and Dammam. As a sanity check, we computed the total cost of the optimal warping path, which agrees with DTW (X, Y). The total cost of the optimal warping path is 3178.0 for Riyadh and Jeddah, and for Makkah and Dammam is 2816.0.

Finally, we visualized the cost matrix A and the accumulated cost matrix B along with the optimal warping path in blue; see Figure 19 and Figure 20.

Figure 19. Visualization with the optimal warping path by the blue for Riyadh and Jeddah

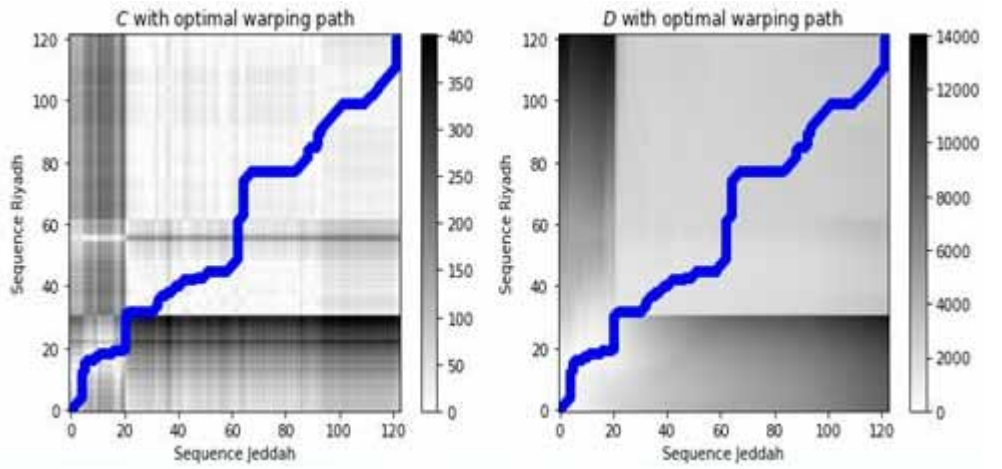
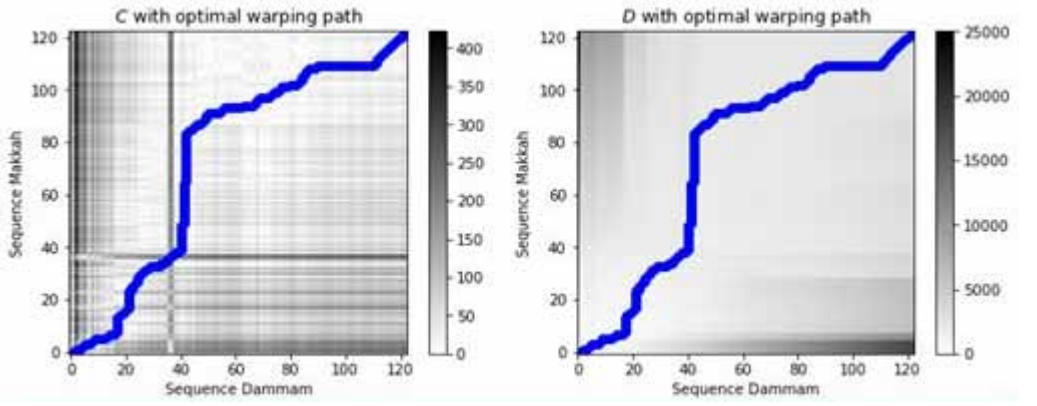


Figure 20. Visualization with the optimal warping path by the blue for Makkah and Dammam



5. CONCLUSION

The current circumstances of the worldwide Covid-19 epidemic show alarming signs and threats. Nevertheless, all the various governments of the world nations cooperate together with their governmental facilities and other organizations to reach accurate statistics and figures regarding the current situation, the accuracy of the statistics is a necessary step to figure out the most appropriate methodology to handle that epidemic. In this paper, we have presented an efficient clustering method that selects the optimal initial centroids of multi-Time Series Based K-Means and PCA algorithm with forecasting. With help of the proposed method, we have efficiently created clusters of different cities of KSA according to similar health care quality during COVID-19. Clustering multi-time series is one of the important tools to help the decision-makers to clarify the ambiguity of the current critical situation and establish a coherent plan.

The ARIMA statistical model was used in Covid-19 data to extract statistical data for the Covid-19 data for four months for Riyadh, Jeddah, Makkah, and Dammam extracted the data visualization where we observed a decrease in cases between the beginning of July and the end of October. Then use PCA of the original data and we get to have two columns where the original 4 dimensions data were reduced by only two components, we deduced identifying trends by taking the average rolling mean to take the average points, which tends to reduce noise and diversity seasonally. Then find the autocorrelation between the variables and the time-series correlation. Finally, we found the forecasting from the time series and ARIMA model, Where the Cases in Jeddah are the best and most appropriate because they have the lowest values of the BIC and AIC evaluation criteria. We have presented an efficient clustering method that selects the optimal initial centroids of the K-means clustering algorithm. With the help of the proposed method, we have efficiently created clusters of different cities according to similar health care quality during COVID-19., we were able to visualize the three groups where the black points represent the middle point of each group, and dynamic time warping (DTW) was used to get a better idea of the optimal twisting path.

In this paper, while experimenting with COVID-19 datasets, our model outperforms in terms of a reduced number of constant iterations, which consequently reduces the execution number of times series by using clustering K-means.

CONFLICT OF INTEREST

The authors declare that no competing exist regarding the publication of this paper.

REFERENCES

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, *66*(2), 237–242.

Al-Aziz, S. N., & Alotaibi, R. (2019). Model-Based Discriminant Analysis and Two-Step Clustering for Breast Cancer patients. *World Applied Sciences Journal*, *37*(6), 500–511.

Ambigavathi, M., & Sridharan, D. (2020, April). Analysis of Clustering Algorithms in Machine Learning for Healthcare Data. In *International Conference on Advances in Computing and Data Sciences* (pp. 117-128). Springer.

Ansari, M. I., & Ahmed, S. M. (2001). Time series analysis of tea prices: An application of ARIMA modelling and cointegration analysis. *The Indian Economic Journal*, *48*(3), 49.

Baudry, J. P., Raftery, A. E., Celeux, G., Lo, K., & Gottardo, R. (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, *19*(2), 332–353.

Cao, D., Tian, Y., & Bai, D. (2015, July). Time series clustering method based on Principal Component Analysis. In *5th International conference on information engineering for mechanics and materials* (pp. 888-895). Atlantis Press.

Chen, Y., Wang, S., Xiao, X., Liu, Y., Hua, Z., & Zhou, Y. (2021). Self-paced enhanced low-rank tensor kernelized multi-view subspace clustering. *IEEE Transactions on Multimedia*, *24*, 4054–4066. doi:10.1109/TMM.2021.3112230

Dozie, K. C. N., & Ijomah, M. A. (2020). A comparative study on additive and mixed models in descriptive time series. *American Journal of Mathematical and Computer Modelling*, *5*(1), 12–17.

Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: Examining and forecasting change. *Frontiers in Psychology*, *6*, 727.

Li, H. (2014). Asynchronism-based principal component analysis for time series data mining. *Expert Systems with Applications*, *41*(6), 2842–2850.

Li, H. (2019). Multivariate time series clustering based on common principal component analysis. *Neurocomputing*, *349*, 239–247. doi:10.1016/j.neucom.2019.03.060

Niennattrakul, V., & Ratanamahatana, C. A. (2007, April). On clustering multimedia time series data using k-means and dynamic time warping. In *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)* (pp. 733-738). IEEE.

Olive, D. J. (2017). Principal component analysis. In *Robust multivariate analysis* (pp. 189–217). Springer. doi:10.1007/978-3-319-68253-2_6

Ralanamahatana, C. A., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M., & Das, G. (2005). Mining time series data. In *Data mining and knowledge discovery handbook* (pp. 1069–1103). Springer.

Safarineja, B., Menhaj, M. B., & Karrari, M. (2009). Distributed data clustering using expectation maximization algorithm. *Journal of Applied Sciences (Faisalabad)*, *9*(5), 854–864.

Salem, N., & Hussein, S. (2019). Data dimensional reduction and principal components analysis. *Procedia Computer Science*, *163*, 292–299. doi:10.1016/j.procs.2019.12.111

Wan, X., Li, H., Zhang, L., & Wu, Y. J. (2021). Multivariate Time Series Data Clustering Method Based on Dynamic Time Warping and Affinity Propagation. *Wireless Communications and Mobile Computing*, *2021*, 1–8.

Wang, M., Pan, J., Li, X., Li, M., Liu, Z., Zhao, Q., & Wang, Y. (2022). ARIMA and ARIMA-ERNN models for prediction of pertussis incidence in mainland China from 2004 to 2021. *BMC Public Health*, *22*(1), 1–11.

Sundus Naji Alaziz is a professor at Department of Mathematical Sciences, Faculty of Science, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Bakr Albayati is a professor at Department of Basic Sciences, Common First Year King Saud University, Riyadh, Saudi Arabia.

Abd AL-Aziz Hosni El-Bagoury obtained his BSc, MSc and Ph.D (Mathematical Statistics) from Mathematics Department, Tanta University, Egypt. His research activities mainly focused on: Search theory, probability theory, optimization, operations research, stochastic process, applied probability and mathematical modeling. He has published papers in leading and well reputed international Journals of Mathematical Sciences either ISI or Scopus journals. He is referee in some international journals such as Mathematical Methods in the Applied Sciences, Information Science letters, Journal of Statistics Applications& Probability and Applied Mathematics & Information Sciences.