Original Research Article

FUNCTIONAL DATA ANALYSIS ON GLOBAL COVID-19 DATA

ABSTRACT

No one can deny that COVID-19 has spread around the world and is still emerging as a new variation in some areas. Since the time of its breakout, three coronavirus waves have emerged. Daily cases of COVID-19 from 79 countries around the world were selected for the study. The main objective of this research was to model and analyze the behavior of the disease's first wave in the world and on the Asian continent using functional data analysis methods. Functional models were fitted to the data using B-spline basis functions at different orders, and the best-fit curves were further analyzed with respect to their functional behavior and the rate of change. These curves were visualized during the preliminary analysis and later clustered within each continent using functional cluster analysis. The results indicated that all continents, apart from Asia, had two clusters based on their functional behavior, whereas the world data had three clusters. All the continents except the Asian continent had different functional forms and numbers of peaks, but they all had the same number of clusters. The world's 18 countries were divided into two categories, with the remaining 61 countries clustered into a single group. The identified cluster indices were further modeled using multinomial logistic regression models with six popular health index variables. In the world, people over the age of 70 made a significant contribution to selecting cluster 2 over cluster 1 and cluster 3 over cluster 2. On the Asian continent, the female smoker variable preferred cluster 2 over cluster 1, while cluster 3 over cluster 1 could be determined by the median age variable. The findings of the overall study would be helpful for the researchers to understand the spread of the disease and the impact of the health indices on its functional behavior.

Keywords: COVID-19, functional data analysis, functional cluster analysis, multinomial logistic regression

1. INTRODUCTION

Although media attention is only the top of the iceberg, no one can dispute that the coronavirus (COVID-19) has spread at an alarming rate around the world. According to sources, COVID-19 was first found on December 31, 2019, in Wuhan, China. After the initial epidemic, COVID-19 moved quickly to other Asian and non-Asian nations while continuing to grow throughout China. Although the disease shares many similarities with the severe acute respiratory syndrome virus that swept over Asia in 2003, it has been shown to spread much more swiftly, and there was no vaccine available at the outset of COVID-19.

There was no specific cure, vaccination, or medicine for the sickness caused by COVID-19 as of January 2022. The COVID-19 outbreak is forcing many governments and areas around the world to implement strict policies and upgrade medical facilities to safeguard people. Patients are treated based on their health and clinical circumstances and common signs or symptoms such as fever, cough, shortness of breath, and breathing difficulties. Furthermore, extremely supportive treatment for persons who have been infected is very helpful. The World Health Organization (WHO) proclaimed the COVID-19 outbreak a global health emergency and urged individuals to continue strengthening and upgrading their readiness for COVID-19-related health problems.

Most of the studies were conducted using descriptive analysis techniques and popular predictive models. Limited work has been done to study the behaviour of its functional forms and the rates of

change of the disease spread. This functional behaviour will probably lead to the study of the common functional characteristics and further could be used to identify functional clusters of countries in the world. In addition, such as different health indices could have made a significant contribution to the functional behaviour of the spread of the disease and its prevention. Therefore, the study of the country wise factors, the study mainly considers Functional data analysis (FDA) as the tool to model the counts of COVID-19 cases which are updated daily in the form of discrete observations with this, the functional models convert discrete time intervals into continuous functional forms. Additionally, the FDA was used for the analysis because earlier research had been conducted on the discretely observed COVID-19. Moreover, cluster analysis is applied to find categories and, using the respective clustered index, Population density, Median age, aged 65 older, aged 70 older, Female smokers and Hospital bed per thousand, construct multinomial logistic models to identify the contribution to each cluster. The likelihood ratio test which is a comparison of the likelihood ratio (-2LL) for the researcher's model with predictors (called model chi square) to the likelihood ratio for a baseline model with simply a constant was used. This ratio's significance is determined using the Chi square.

- H_0: There is no difference between null model and final model.
- H_1: There is difference between null model and final model.

The county-level variables linked with the COVID-19 case-fatality rate (CFR) were identified using publically accessible datasets and a negative binomial generalized linear model [12]. A higher number of hospitals per 10,000 people, a ban on religious gatherings, a higher percentage of individuals living in mobile homes, and a higher percentage of uninsured persons were all linked to lower CFR. A higher percentage of the population over 65, a higher percentage of black or African Americans, a higher asthma prevalence, and a bigger number of hospitals in a county were all linked to increased CFR.

The pattern of COVID-19 victims, who suffered from some underlying disorders, was investigated using a single population, Hungary [9]. The grouping is structured by age, gender, and underlying medical conditions. For age-based and age-independent analysis, K-Means and two-step clustering algorithms were used. Their clustering result can predict comparable cases allocated to any cluster, which will be a severe concern for the population.

From May to August 2021, 422 cancer patients were studied in an institution-based cross-sectional study [4]. The primary data was collected using a standardized interviewer-administered questionnaire. To evaluate the independent association of covariates with the outcome variable, descriptive statistics and binary logistic regression were used. His study discovered that many cancer patients had little understanding of the vaccine and that the first and second rounds of COVID-19 vaccination had a low success rate.

Alotaibi (2021) conducted a statistical and deterministic investigation into the spread of COVID-19 in Saudi Arabia [5]. A statistical analysis of the acquired data was carried out, which included data fitting using some statistical distributions, as well as a seven-day moving average for deaths, recovered, and infected. Additionally, a simple Susceptible-Infected-Recovered (SIR) mathematical model based on the concept of piecewise modeling, where the model is defined within intervals, captures the crossover exhibited by the spread throughout the Saudi Arabian kingdom.

Kumar (2020) used cluster analysis, a data mining technique, to categorize real groups of infectious disease "new coronavirus disease (COVID-19)" data from various Indian states and union territories (UTs) based on their high similarity [13]. The findings enable them to identify clusters of infected Indian states and UTs. The main goal of clustering in this study was to improve monitoring techniques in infected states and UTs in India, which will be very useful to the government, doctors, police, and others involved in understanding the seriousness of the spread of COVID-19 to improve government policies, decisions, medical facilities, treatment, and to reduce the number of infected and deceased people.

Various curve estimate algorithms are heavily used by the FDA. Curve fitting commonly includes interpolation nowadays, when advanced monitoring equipment is routinely employed in many sciences and is so accurate that measurement errors can be ignored. Smoothing may be used in modeling if the observed data has errors that need to be removed. The discretely observed COVID-19 data has been chosen. This is why FDA was used to analyze this data set in this study.

2. METHODOLOGY

2.1 Data Collection

The data was collected from the World Health Organization coronavirus (COVID-19) Dashboard. At first, the dataset, which contains data from 1 March 2020 to 30 May 2020, is divided into subsets according to continents. Although there are 7 continents, only 6 of them are selected since most people live there. Next, 15 countries are selected from each continent except South America and Oceania by considering tourist attractions. For the first part of the data analysis, new cases per million for each country were used. Fig.1 shows a physical map specifying the location of the selected countries that were considered in the analysis.

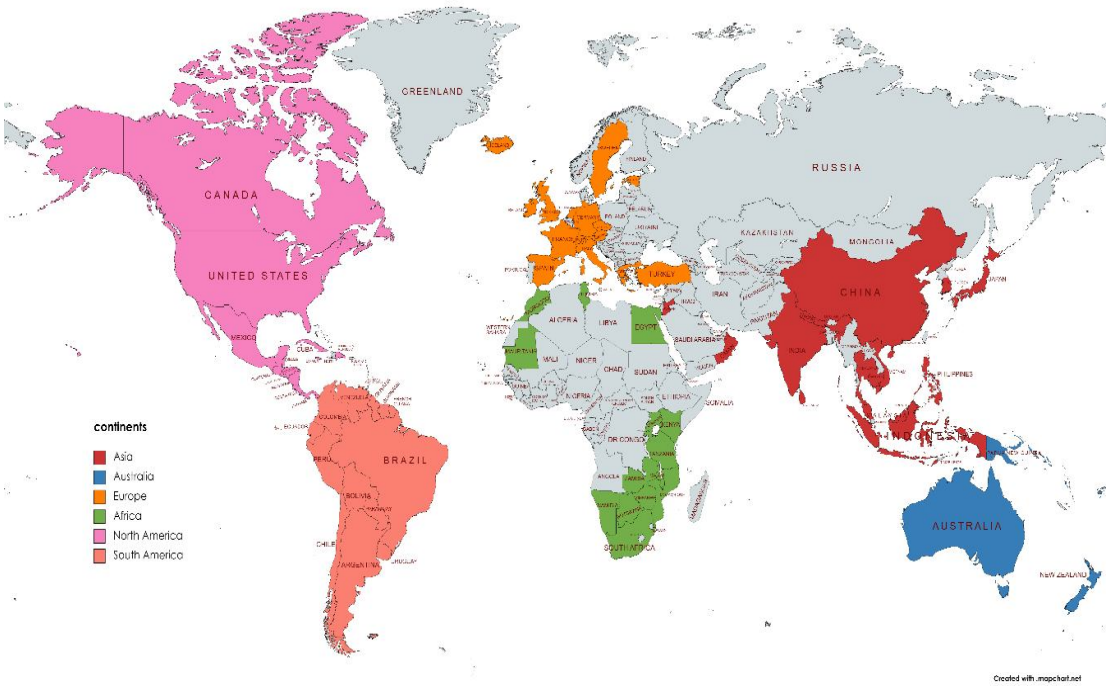

Fig. 1. Locations of the selected countries within each continent
(https://www.mapchart.net/world.html)

2.2 Functional Data analysis

Functional data analysis is a type of statistical analysis that involves data collected from a group of locations on a continuum. The data collection sites are frequently densely spaced over the continuum to ensure that the curvature shape is effectively captured. The term "functional data fitting" refers to a model of the functional observations. The model representation for time-varying curves in 2-dimensional space could be a linear combination of basis functions.

Functional curve fitting in this study primarily follows the three methods outlined above, with the goal of determining an adequate representation of the functional curves:

1. Observe the data curves.
2. Depict some local features of the curves
3. To improve data fitting with a fixed number of basis functions.

2.2.1 Functional Data Representation

A random sampling of independent real-valued functions, $X_1(t), X_2(t), \ldots, X_n(t)$ on a compact interval $I = [0, T]$on the real line is typical of first-generation functional data. Curve data is another name for this type of information. These real-valued functions can be thought of as realizations of a one-dimensional stochastic process, which is frequently considered to be in Hilbert space, such as $L^2(I)$. A stochastic process $X(t)$ is said to be a L^2process if and only if it satisfies $E\left(\int_I X^2(t)dt\right) < \infty$.

2.2.2 Creating a Smooth Curve

Even though it is seen discretely, the data object to be analyzed should be considered a function rather than an individual data point in FDA. These discrete data may contain observational error. Converting raw data into functional objects is the initial stage in a FDA. The discrete observations are fitted with a curve to represent the continuous underlying process. The discrete points are then set aside, while the functional objects are kept for further analysis. These smooth functions are denoted by the following linear combination of basis functions:

$$X(t) = \sum_{k=1}^{K} C_k \, \varphi_k(t)$$

Where C_k are basis coefficients and $\varphi_k$is the known basis function while the K is the size of the maximum basis required.

2.2.2.1 The B-Spline Basis

De Boor invented the popular B-spline basis function system. These are a subset of spline functions that, as ordinary splines, have the property of being piecewise polynomials connected by knots on the time axis.

To get an estimate for $X(t)$, one only needs to figure out the basis functions and then estimate the coefficient values C_k using the least squares approach, weighted least squares technique, roughness penalty approach, and so on. In the B-spline basis system, the φ_k's piecewise spline functions of order that are coupled smoothly at the time points where the knots are located and specified over the full region from t_0 to t_{M-1}. In the interval $[t_i, t_{i+1})$, write $B_{i,k}$ for the k[th] order piecewise spline. After that, the B-spline basis functions are built as follows:

$$B_{i,0}(t) = \begin{cases} 1, & \text{if } t_i \le t < t_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

and

$$B_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(t)$$

Where k indicates the order of B-spline . Each basis function evaluated at time $t, B_{i,k}(t)$, might be produced given knowledge of the values of its "neighbors," $B_{i,k-1}(t)$ and$B_{i+1,k-1}(t)$as shown by the recursive relationship above. With this construction method, each φ_k is only allowed to be positive on the smallest number of subintervals divided by the knots. To be more exact, B-spline basis functions have a property known as compact support, which states that any $\varphi_k$can be positive only over *k* subintervals. Due to the $M \times K$band-structured model matrix $\varphi(t)$with entries being the B-spline functions evaluated at different time points $t = (t_0, t_1, \ldots, t_{M-1})'$,this property assures the fitting algorithm's computational speed. regardless of how many knots are in the interval (t_0, t_{M-1}).

The order of the spline functions can be determined by the number of derivatives of the basic curves that must be smooth. The m^{th}derivative $(m < K)$ of the k^{th} order B-spline can be computed recursively by

$$D^m B_{i,k}(t) = (k-1)\left\{\frac{D^{m-1}B_{i,k-1}(t)}{\tau_{i+k-1} - \tau_i} - \frac{D^{m-1}B_{i+1,k-1}(t)}{\tau_{i+k} - \tau_{i+1}}\right\}$$

Where the i^{th} knot is τ_i.

2.3 Cluster Analysis

Traditional cluster analysis methods and functional clustering approaches are extremely similar. To create a distance matrix, one must first calculate distances between all pairs of curves to be clustered. The Euclidean distance, often known as the L_2norm, is one of the most widely used distance measures:

$$d_{L_2}[y_i(t), y_j(t)] = \sqrt{\int [y_i(t) - y_j(t)]^2 dt} = d_{ij}$$

The distance matrix obtained from the initial set of curves can be simply applied to popular functional clustering algorithms. Popular clustering methods currently include, but are not limited to, hierarchical based methods with various types of links, partitioning methods such as the K-means and K-medoids methods, and Ward's method, a hybrid method.

2.3.1 K-mean Clustering

K-means type clustering algorithms have been widely applied to functional data and are more popular than hierarchical clustering algorithms. It is natural to view cluster mean functions as the cluster centers in functional clustering.

For a set of functional data $\{X_i(t); i = 1, \ldots, n\}$, the k-means functional clustering aims to find a set of cluster centers$\{\mu^c; c = 1, \rightleftarrows \ldots, L\}$, assuming there are *L* clusters, by minimizing the sum of the squared distances between X_i and the cluster centers that are associated with their cluster labels $\{C_i; i = 1, \ldots, n\}$, for a suitable functional distance d. In other words, the $n$observations $\{X_i\}$ are divided into L groups in such a way that

$$D = \frac{1}{n} \sum_{i=1}^{n} d^2(X_i, \mu_n^c)$$

is minimized throughout all sets of functions $\{\mu_n^c; c = 1, \ldots, L\}$, where

$$\mu_n^c(t) = \sum_{i=1}^{n} X_i(t)\, 1_{\{C_i = c\}} / N_c$$

and

$$N_c = \sum_{i=1}^{n} 1_{\{C_i = c\}}$$

The L_2 norm is frequently selected as the distance d. Because functional data is often contaminated by measurement errors and can be sparsely or irregularly sampled, a common approach to minimize D is to project infinite-dimensional functional data onto a low-dimensional space of a set of basis functions, similar to functional correlation and regression.

2.3.2 Determine the number of clusters

One of the most difficult aspects of k-mean is determining the number of clusters *k*. However, there are methods to help making decisions. One method is to start with a single cluster and subsequently split it until no reasonable clusters can be discovered, if at all. Making box plots of all similarity metrics between the aligned curves and their cluster centroid for different values of *k* is another option. A large number of outliers could indicate that a larger number of clusters should be explored. Furthermore, if the corresponding means of the similarity measures are plotted against the various *k: s*, we may look for a point where the graph begins to level out and choose the *k* where this point emerges.

2.4 Logistic Regression Analysis

When the dependent variable is categorical (or nominal), logistic regression is utilized. The number of dependent variables in binary logistic regression (BLR) is two, whereas the number of dependent variables in multinomial logistic regression(MLR) is more than two. As a method for studying categorical-response variables, logistic regression competes with discriminant analysis. Many statisticians believe that logistic regression is more adaptable and better suited to modeling the majority of scenarios than discriminant analysis. This is because, unlike discriminant analysis, logistic regression does not require that the independent variables are regularly distributed.

2.4.1 Multinomial logistic Regression

To predict membership in more than two groups, use MLR. It functions in a similar way to binary logistic regression. The analysis divides the result variable into a series of two-category comparisons.

If we have n independent observations with p-explanatory variables and a qualitative response variable with *k* categories, one of the categories must be considered the base level, and all logits must be generated relative to it in the multinomial case. Any category can be used as the starting point, therefore we'll start with category *k*. Because there is no ordering, any category can be called *k*. Let P_j denote the multinomial probability of an observation falling in the j^{th} category ,to find the relationship between this probability and the p explanatory variables, $x_1, x_2, \ldots, x_p$, the multiple logistic regression model, then is

$$\log\left[\frac{P_j(x_i)}{P_k(x_i)}\right] = \alpha_{0i} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \ldots + \beta_{pj}x_{pi},$$

where $j = 1,2,\ldots,(k-1)$ and $i = 1,2,\ldots,n$. Since all the P's add to unity, this reduces to

$$P_j(x_i) = \frac{exp(\alpha_{0i}+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\ldots+\beta_{pj}x_{pi})}{1+\sum_{j=1}^{k-1} exp(\alpha_{0i}+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\ldots+\beta_{pj}x_{pi})}$$

and

$$P_k(x_i) = \frac{1}{1+\sum_{j=1}^{k-1} exp(\alpha_{0i}+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\ldots+\beta_{pj}x_{pi})}$$

for $j = 1,2,\ldots,(k-1)$.

Note that $\sum_{j=1}^{k} P_j(x_i) = 1$and that there are $(k-1) \times (p+1)$ coefficients. The model parameters are estimated by the method of maximum likelihood [8]. The multinomial logistic model has an interesting interpretation in terms of logistic regression.

3. RESULTS AND DISCUSSION

3.1 Preliminary Analysis

Here basic central propensity statistical measures such as minimum, maximum, quartiles, median, mean and the standard deviation of the daily new cases per millions for the six continents over four months' period were considered.

Table. 1. Summary Statistics of new cases over the four-months period

Continent	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.	Standard deviation
Asia	0.000	0.028	0.526	7.854	2.415	349.509	29.418
Europe	0.000	3.549	12.235	30.183	44.072	752.648	43.256
Oceania	0.000	0.000	0.000	1.385	0.543	24.776	3.618
Africa	0.000	0.000	0.243	2.780	1.486	120.083	10.129
North America	0.000	0.149	3.113	17.211	22.823	310.162	30.696
South America	0.000	0.244	3.296	30.471	25.223	1700.680	80.456

According to Table.1, the most significant contribution to the mean of the daily new cases per million was made by the South American continent, while European countries showed the second highest mean value, which was 30.183. Furthermore, the North American continent recorded 17.211 cases per million as its mean value, and this value was almost half of the mean of the Asian continent. Looking at it more specifically, the highest standard deviation was made up by South America, and the reverse is true for Oceania. There was a small difference between the standard deviation of the Asian and North American continents. As well, the standard deviation of Africa was three times larger than that of Oceania.

3.2 Functional Data Analysis

The three steps described under Section 2.2 can be used to construct the curves below.

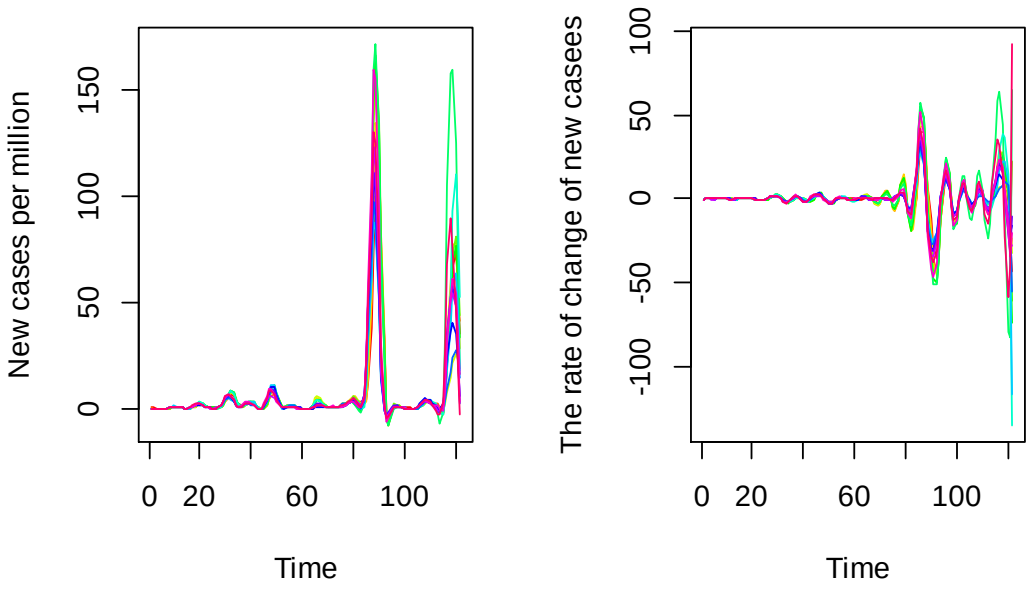

Fig.2. Smoothing curves and the first derivative curve of Asian countries

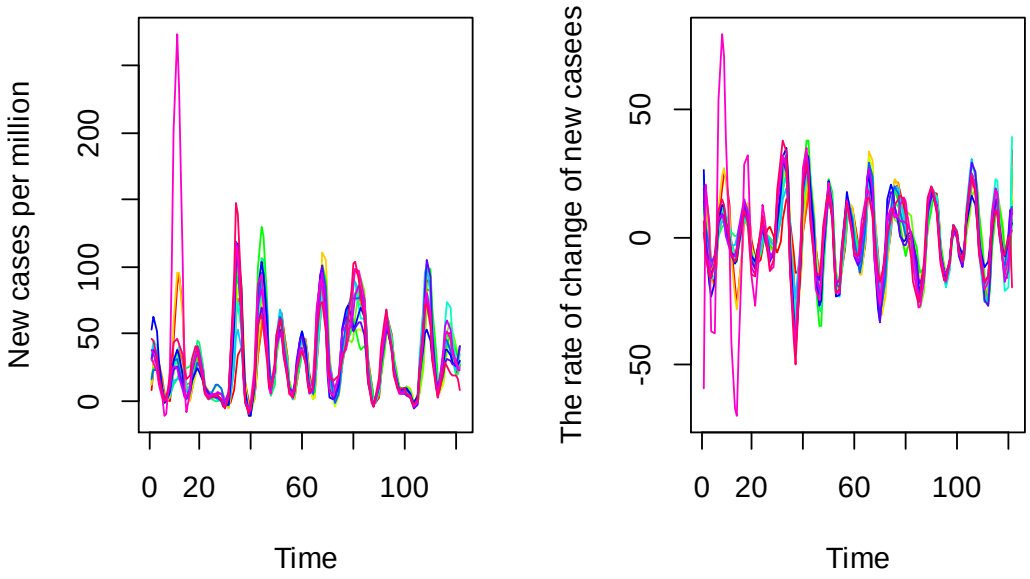

Fig.3. Smoothing curves and the first derivative curve of Europe countries

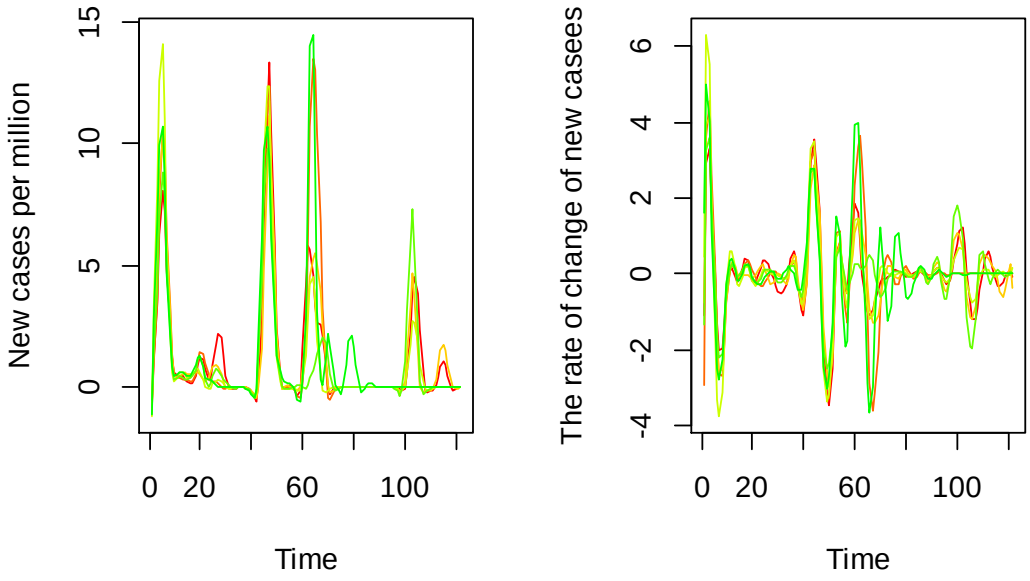

Fig.4. Smoothing curves and the first derivative curve of Oceania countries

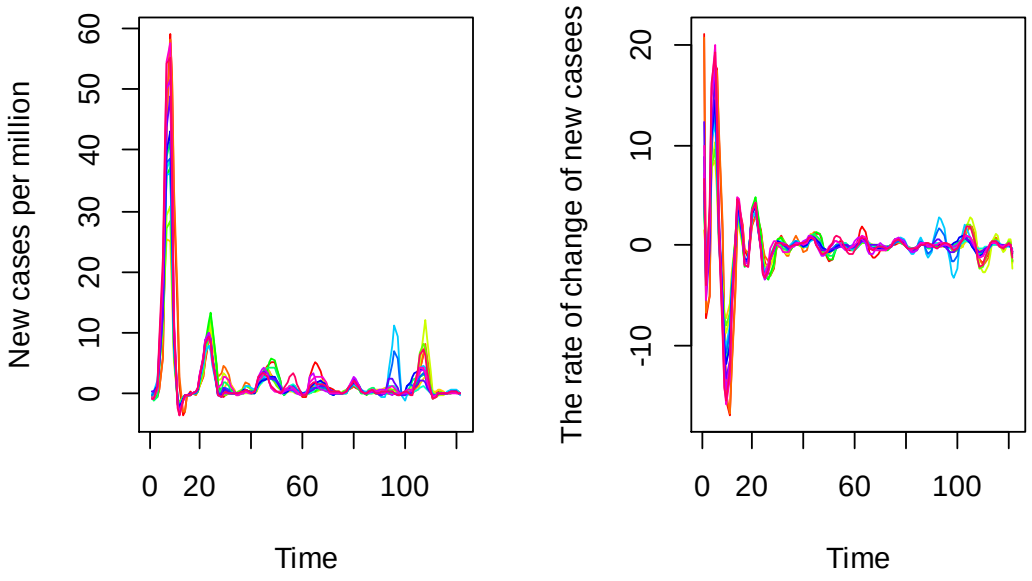

Fig.5. Smoothing curves and the first derivative curve of African countries

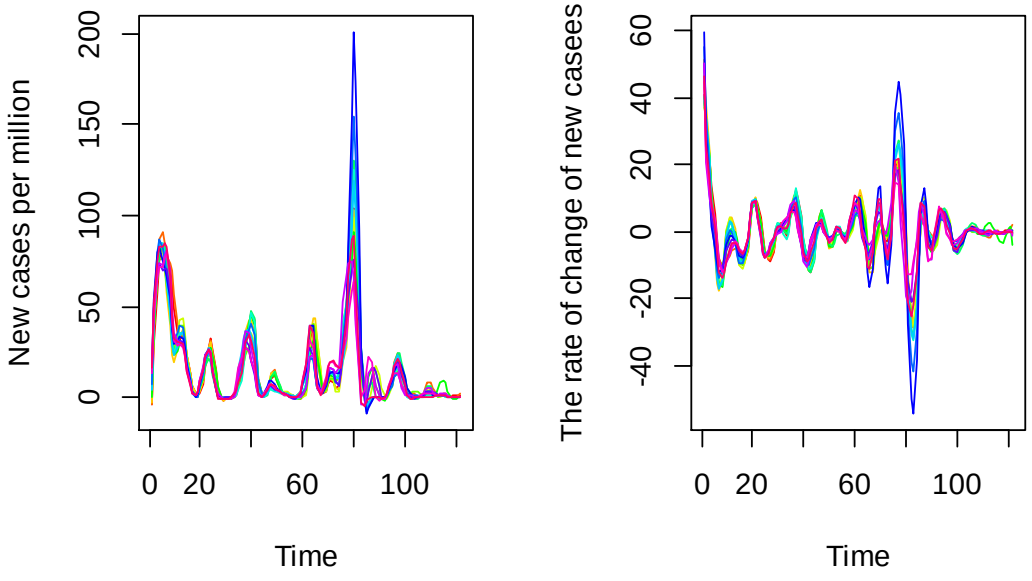

Fig.6. Smoothing curves and the first derivative curve of North American countries

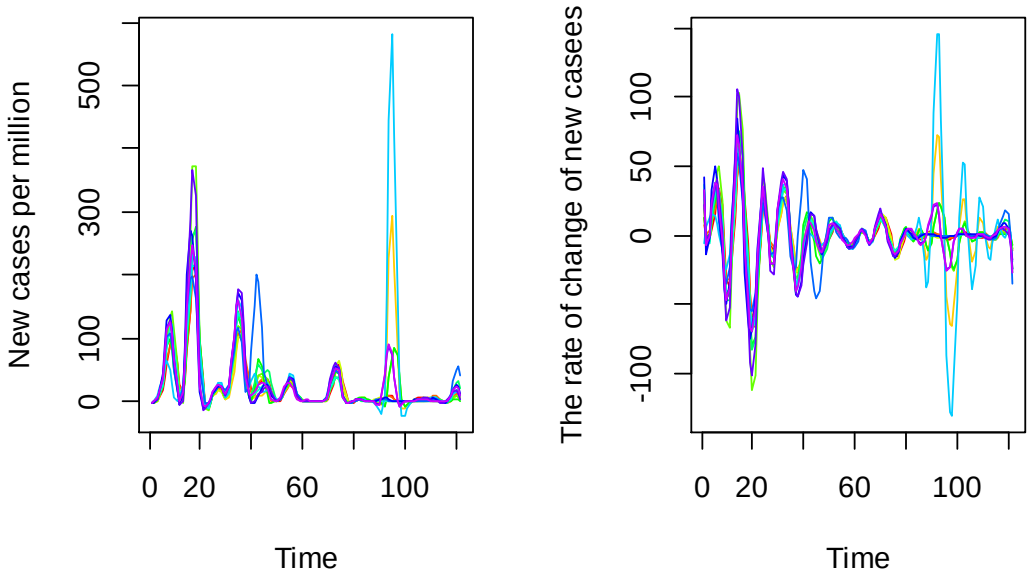

Fig.7. Smoothing curves and the first derivative curve of South American countries

The data sets indicate non-periodic curves; therefore, a B-spline basis has to be used. The optimal number of basis for these data is 61, and multiple knots are not used to obtain smooth curves. The functional curves represent daily new cases, while the rate of change curves show how speedily the COVID-19 epidemic is spreading. All the areas' first derivative graphs, with the exception of the Asian and African continents, exhibit fluctuation through time, with the Asian continent indicating a large change near the end of the period and the converse is true for African countries.

3.3 Cluster Analysis

In this study, the K-mean algorithm is used to cluster the countries in each region and select the optimum number of clusters using the functional one-way ANOVA test. Further, the cluster analysis has been done for all the countries considered in the analysis.

3.3.1 Asian continent

Under this subsection, 15 countries are considered, and initially two clusters are obtained. However, as in section 2, the optimum number of clusters should be found. For that, continue the above procedure. Finally, three clusters were obtained.

Cluster 1: India, Philippines, Malaysia, Maldives

Cluster 2: Vietnam, South Korea, Oman, Sri Lanka, Cambodia

Cluster 3: China, Thailand, Japan, Indonesia, Nepal, Jordan

To find the difference between group means, apply the functional one-way ANOVA test. As a result, a smaller p-value indicates that the null hypothesis can be rejected at a 0.05 significant level and conclude that the three mean functions are not all equal. The K-mean approach is used to choose

clusters, which follows the shape of curves and, based on figure experiences, the three clusters have distinct shapes.

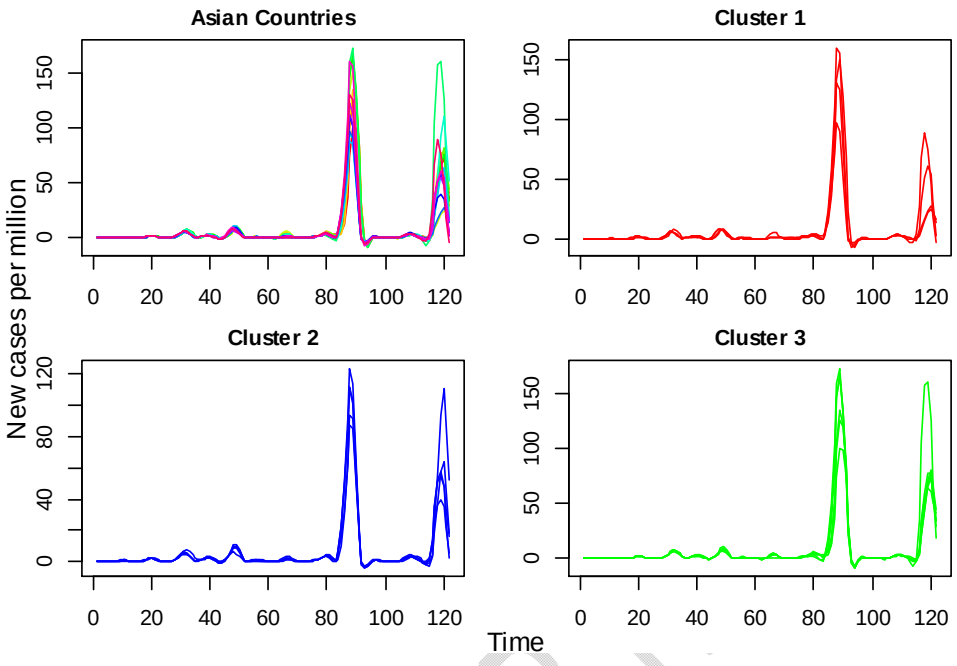

Fig.8. The curves of each clusters of Asian Continent

3.3.2 Europe continent

The 15 countries were selected from the continent and categorized according to the shape of their curves of the new cases per million. Cluster 1 and Cluster 2 are mentioned in the below.

Cluster 1: France, Ireland, Italy, Spain, Sweden, Switzerland, United Kingdom, Croatia, Portugal

Cluster 2: Austria, Germany, Greece, Iceland, Netherlands, Belgium

A functional one-way ANOVA test was performed after cluster analysis. Consequently, the p-value of 0.04 indicates that the null hypothesis can be rejected at a 0.05 significant level and conclude that the two mean functions are not all equal. Although the countries of the second cluster climbed up in the middle of March, certain curves in cluster 1 fluctuated at 0 during the given period, and the number of patients ascended to their maxima at the end of the first month. Cluster 2 is made up of the countries with the largest number of new cases.

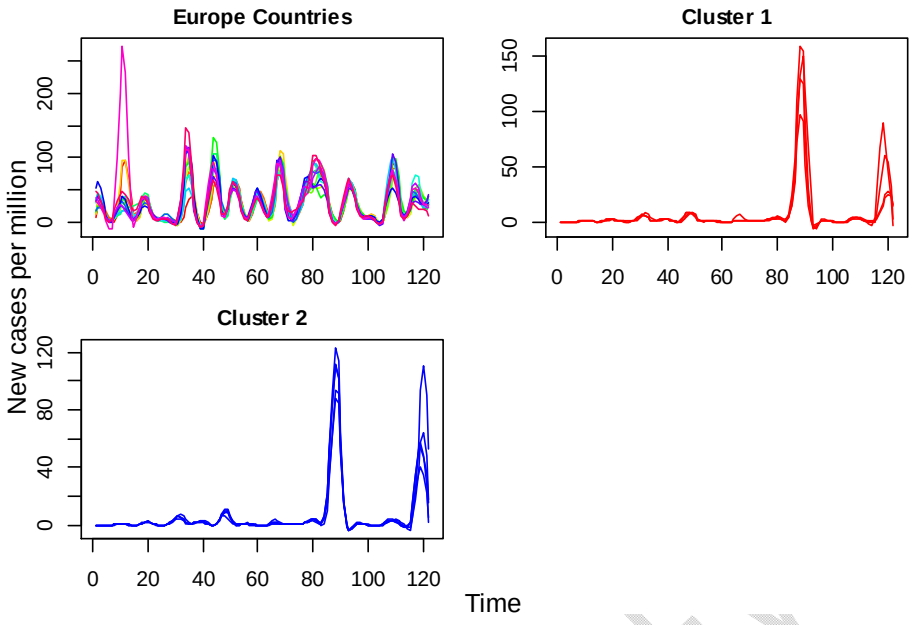

Fig.9. The curves of each clusters of Europe Continent

3.3.3 Oceania continent

Only six countries were chosen on this continent, and all six curves follow the same pattern. As a result, clusters cannot be discovered. Because many of the countries on the Oceania continent are surrounded by water, the rate of inflection is low in comparison to other continents. As a result, the various clusters cannot be distinguished.

3.3.4 African continent

Two clusters are identified using the k-mean cluster technique in R software package:

Cluster 1: Egypt, Zambia, Zimbabwe, Morocco, Rwanda, Namibia, Tunisia, Kenya, Malawi

Cluster 2: South Africa, Tanzania, Mozambique, Botswana, Mauritius, Uganda

A functional one-way ANOVA test was performed after cluster analysis. As a result, the p-value < 0.0001 indicated that the null hypothesis can be rejected at a 0.05 significant level and conclude that the two mean functions are not all equal. These clusters cannot be categorized further. The shape of the two cluster curves below can be used to identify the tiny difference.

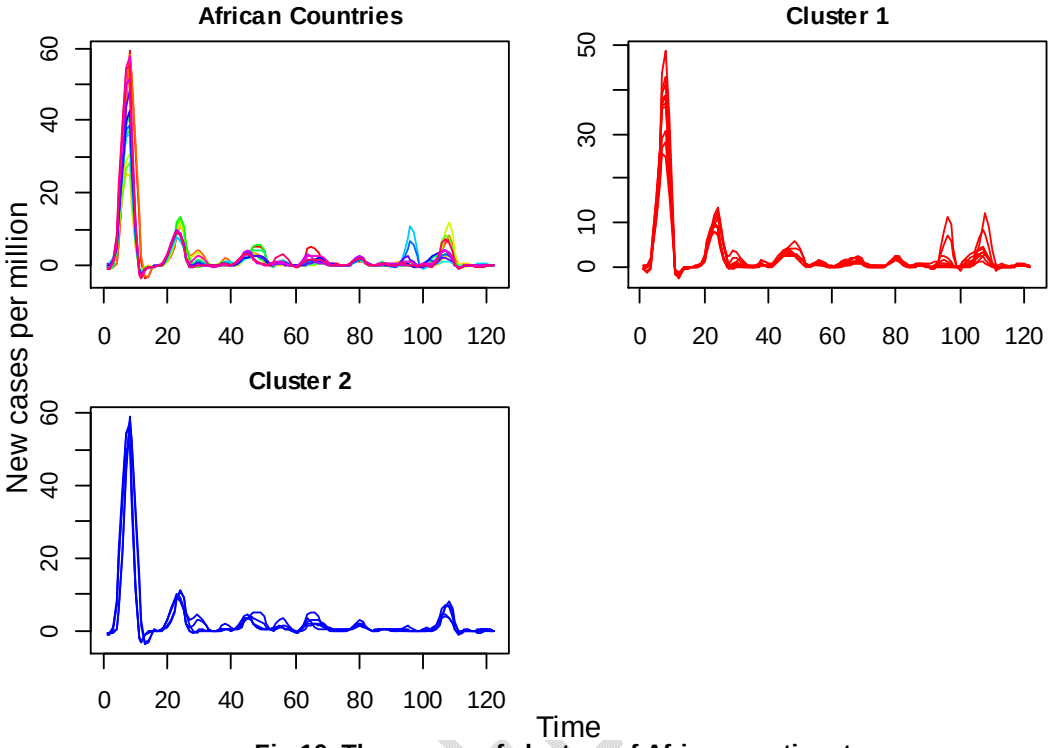

Fig.10. The curves of clusters of African continent

3.3.5 North American continent

North American countries are categorized into two clusters.

Cluster 1: Mexico, Dominican Republic, Honduras, El Salvador, Guatemala, Barbados, Jamaica, Trinidad and Tobago

Cluster 2: United States, Canada, Belize, Costa Rica, Cuba, Panama, Nicaragua

A functional one-way ANOVA test was performed after cluster analysis. Consequently, the null hypothesis can be rejected and infer that the two mean functions are not all equal because the p-value is 0.01(<0.05).

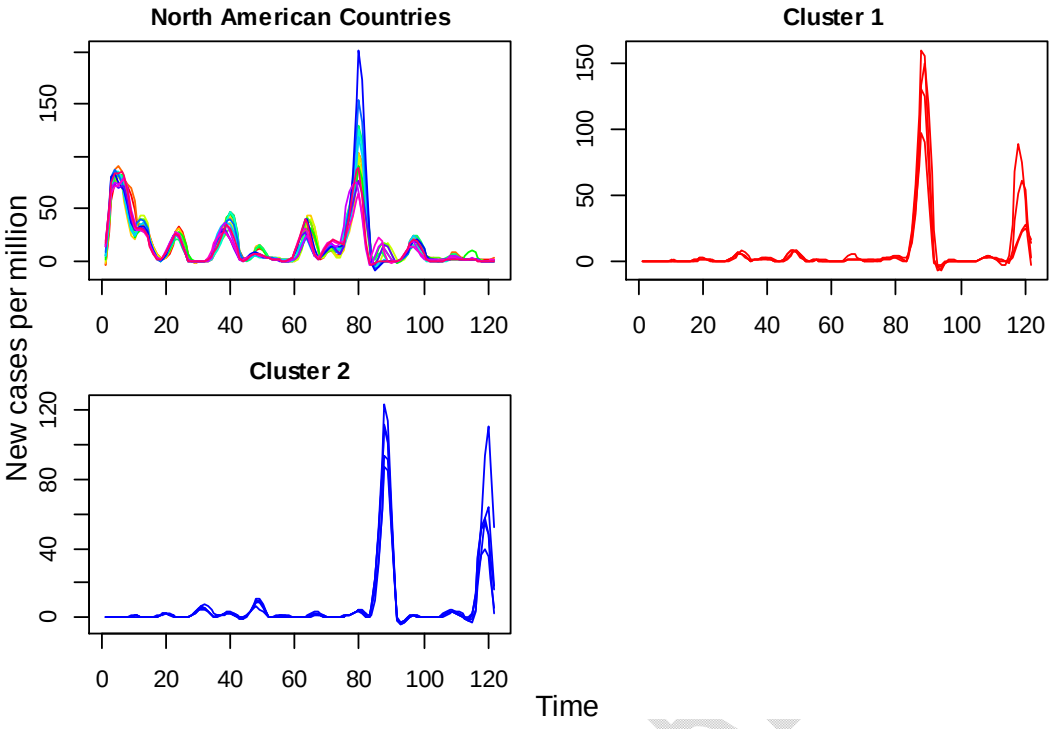

Fig.11. The curves of clusters of North American continent

3.3.6 South American continent

Only 13 countries are classified as being part of the South American continent, and they are as follows:

Cluster 1: Ecuador, Colombia, Paraguay, Falkland Islands, Guyana, Suriname

Cluster 2: Brazil, Chile, Argentina, Peru, Uruguay, Bolivia, Venezuela

A functional one-way ANOVA test was performed after cluster analysis. A p-value less than 0.05 indicates that the null hypothesis is rejected at a 0.05 significant level and conclude that the two means are not all equal. For the amount of new instances, Cluster 1 and Cluster 2 have different curve forms.

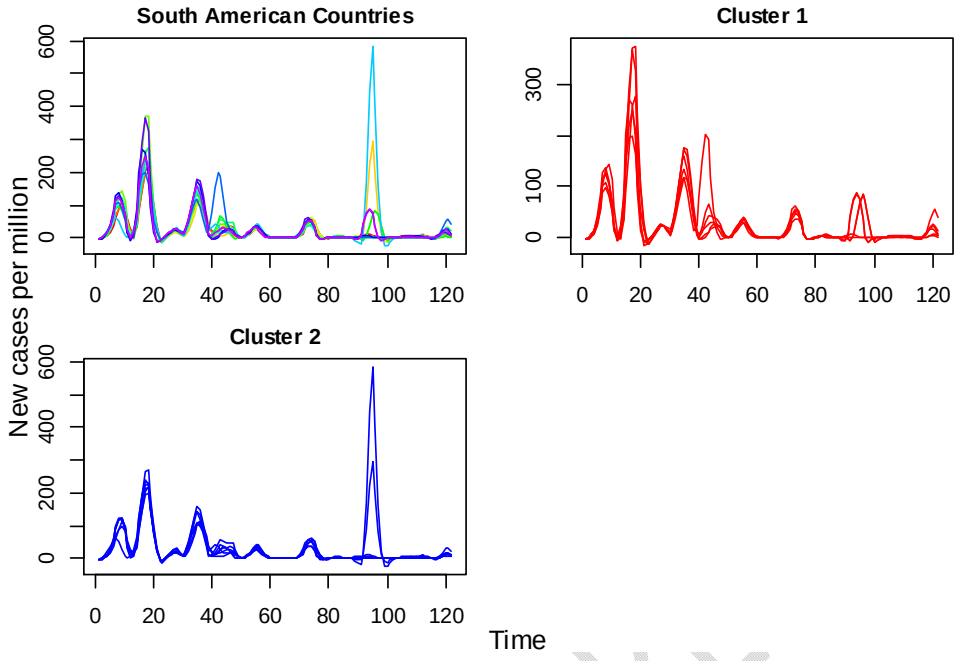

Fig.12. The curves of clusters of South American continent

3.3.7 The world

All the selected countries are first divided into two clusters: cluster 1 and cluster 2. The optimal number of clusters should be determined here. As a result, both of these clusters are attempting to be further classified. Finally, the world can be divided into three distinct groupings.

Cluster 1: Barbados, Belize, Botswana, Canada, Chile, China, Colombia, Costa Rica, Croatia, Cuba, Dominican Republic, Ecuador, Egypt, El Salvador, Falkland Islands, Fiji, France, French Polynesia, Germany, Greece, Guatemala, Guyana, Honduras, Iceland, India, Indonesia, Ireland, Italy, Jamaica, Japan, Jordan, Malawi, Malaysia, Maldives, Mauritius, Mexico, Mozambique, Namibia, Nepal, Netherlands, New Caledonia, New Zealand, Nicaragua, Oman, Paraguay, Peru, Rwanda, South Africa, Spain, Sri Lanka, Suriname, Sweden, Switzerland, Tanzania, Thailand, Trinidad and Tobago, Tunisia, United States, Venezuela, Vietnam, Zimbabwe

Cluster 2: Belgium, Bolivia, Brazil, Panama, Papua New Guinea, Philippines, Portugal, South Korea, Uganda, Uruguay, Zambia

Cluster 3: Argentina, Australia, Austria, Cambodia, Kenya, Morocco, United Kingdom

A functional one-way ANOVA test was performed after cluster analysis. The null hypothesis can be rejected at a 0.05 significant level with a p-value < 0.001, indicating that the three means are not all equal. The curves in Cluster 1 have a certain shape, and outliers exist in Clusters 2 and 3.

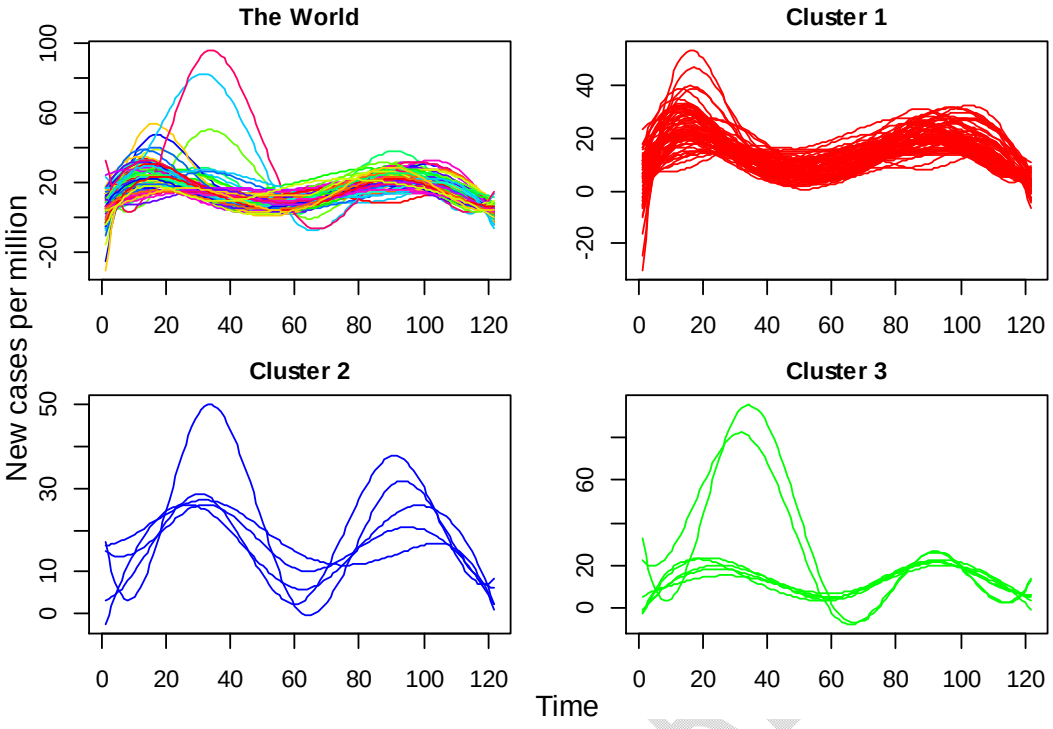

Fig.13. The curves of each clusters in the world

3.4 Logistic Regression models

Except for the Asian continent, every region has two clusters. As a result, depending on the number of nominal variables, BLR or MLR must be used. As a result, MLR is used to identify the regression model for Asian countries, while BLR is utilized to find other models. But here MLR is used only for Asian countries and the world.

The independent variables for estimating a MLR model include population density, median age, aged 65 or older, aged 70 or older, and female smokers. This information is obtained from a WHO official source. The five independent variables that exhibit a link with COVID-19 spread This is why these variables have been chosen. Libraries (nnet) in the R packages can be used to MLR. Because it does not need manipulating the data, the applicable function was picked. To begin, the level of the output should be specified the relevel function's baseline. Cluster 1 was selected as the reference group in this study. The model can be written as:

$$ln(P(Cluster\ 2)/P(Cluster\ 1)) = -1.4089 + 0.0009(Population\ density) - 0.0259(Median\ age) - 0.163(Aged\ 65\ older) + 0.2233(Aged\ 70\ older) + 0.0358(Female\ smokers)$$

$$ln(P(Cluster\ 3)/P(Cluster\ 1)) = -0.986 - 0.0032(Population\ density) - 0.042(Median\ age) - 0.4444(Aged\ 65\ older) + 0.7668(Aged\ 70\ older) + 0.0009(Female\ smokers)$$

After constructing the regression model, the next thing was to calculate p-values of the regression coefficients using Wald tests. The coefficients were considered to be of significance with a two-tailed value of $p < 0.05$, and no significance was attributed otherwise.

Relative risk is defined as the ratio of the likelihood of selecting one outcome category over the probability of choosing the baseline category (and it is sometimes referred to as odds, described in the regression parameters above). The exponentiated regression coefficients are relative risk ratios

for a unit change in the predictor variable since the relative risk is the right-hand side linear equation exponentiated. To have these risk ratios, the coefficients from the model can be exponentiated.

Table. 2. The relative risk of the world model

Cluster (over cluster 1)	Intercept	Population density	Median age	Aged 65 older	Aged 70 older	Female smokers
Cluster 2	0.2444	1.0009	0.9744	0.8496	1.2501	1.0365
Cluster 3	0.3731	0.9969	0.9589	0.6412	2.1529	1.0009

To validate the model, the accuracy of the model is being looked at. This accuracy can be calculated from the classification table. Then the model accuracy turned out to be 79.01% in the training dataset. Moreover, 80.08% accuracy could be achieved in the test dataset, and this number is very close to the trained one, so the model is concluded as good. The significance of predictors to the model can be determined using the results of likelihood ratio tests. All of the variables had considerable main effects on cluster selection at significant level 0.001, according to results. Consequently, Age 70 older variable tend to choose both cluster 2 against cluster 1 and cluster 3 against cluster 1.

Next, the regression mode for the Asian continent is built. since it has three clusters. MLR should be applied. The model can be obtained as follows:

$$\ln(P(Cluster\ 2)/P(Cluster\ 1)) = 225.0957 + 0.0370(Population\ density) - 8.3555(Median\ age) - 10.5595(Aged\ 65\ older) + 8.2223(Aged\ 70\ older) - 4.6949(Female\ smokers) - 0.2649(Hospital\ bed\ per\ thousand)$$

$$\ln(P(Cluster\ 3)/P(Cluster\ 1)) = 579.2922 - 2.6226(Population\ density) - 31.7475(Median\ age) - 375.9642(Aged\ 65\ older) + 793.683(Aged\ 70\ older) + 78.419(Female\ Smokers) + 32.1318(Hospital\ bed\ per\ thousand)$$

In this model, another variable is included to find the significant model, which is hospital bed per thousand. Further, the model summary can be interpreted as below:

Then p-values of the regression coefficients should be calculated, using Wald tests. The coefficients were considered to be of significance with a two-tailed value of $p < 0.05$, and no significance was attributed otherwise. All the coefficients except the hospital bed per thousand in the second equation are significant at 0.05 significant level. The relative risk ratio can be calculated as follows,

Table. 3. The relative risk of the Asian continent model

Cluster (over cluster 1)	Population density	Median age	Aged 65 older	Aged 70 older	Female smokers	Hospital bed per thousand
Cluster 2	< 0.0001	1.0377	3723.333	< 0.0001	< 0.0001	9.0089e+13
Cluster 3	0.0726	< 0.0001	< 0.0001	Inf	1.1400e+34	< 0.0001

To see if this is valid, the model's accuracy is being tested. This accuracy can be calculated using the classification table. After that, the model's accuracy in the training dataset was 100 percent. Furthermore, the model was able to achieve 100% accuracy in the test dataset, which is extremely close to training, indicating that the model was good.

All of the variables had considerable main effects on cluster selection, according to the likelihood ratio test at significant level 0.001. Moreover, Female smoker variable tends to choose cluster 2 against cluster 1 and Median ager variable tends to choose cluster 3 against cluster 1.

4. CONCLUSION

This study aimed to introduce the concept of functional data, the basis function approach for their representation as smooth functions, and smoothing techniques for estimation out of discretely observed data. The main properties of the most widely used basis expansions were discussed and presented in the context of Covid-19 data. Initially, the number of new cases per million was analyzed using descriptive statistical methods. Furthermore, the creation and computation of functional data models helped to study FDA as well as the data categorization into continents by considering the world data set, which contains 79 countries around the world. The concept behind the FDA technique is to convert discrete time intervals into functional data.

The visualizations indicated that the data is non-periodic, therefore B-spline basis is used to smooth data functions and a large number of basis functions are needed to describe the patterns of the daily recorded cases. The functional K-mean cluster analysis follows the method that categorized the similar shapes of mean functional curves in each region of the world separately. Further, all the regions except the Asian continent have two clusters, and the Asian continent and the world data can be categorized into three groups. The MLR can be utilized to find the variable that can be used to describe the cluster index. Consequently, MLR gives a model for the Asian continent and the whole world, while LR gives a model for other regions.

The concentration on how the FDA technique can be used for COVID-19 data analysis is an important benefit of this study. A functional form of data could be established rather than employing discrete data, which could be analyzed over any time interval and also have a functional relationship among the countries in each continent as well as the entire world. Moreover, K-mean cluster analysis is a significant feature of the FDA and cluster id is used to build models that follow multinomial logistic regression. This study paves the way for identifying the specific variables that are related to the categories of counties. As in the world, people over the age of 70 made a significant contribution to selecting cluster 2 over cluster 1 and cluster 3 over cluster 2. On the Asian continent, the female smoker variable preferred cluster 2 over cluster 1, while cluster 3 over cluster 1 could be determined by the median age variable.

Further studies could be carried out by considering the data from later waves of the disease spread, which will help identify the differences in functional changes with respect to the findings of the current study.

REFERENCES

1. Aabed, K., & Lashin, M. M. (2021). An analytical study of the factors that influence COVID-19 spread. Saudi Journal of Biological Sciencesjournal, 28(2), 1177-1195. doi:10.1016/j.rinp.2021.104578

2. Admasu, F. T. (2021). Knowledge and proportion of COVID-19 vaccination and associated factors among cancer patients attending public hospitals of Addis Ababa, Ethiopia, 2021: A Multicenter Study. Infection and Drug Resistance, 14, 4865-4876.

3. Alotaibi, N. (2021). Statistical and deterministic analysis of covid-19 spread in Saudi Arabia. Results in Physics, 28, 1-8. doi:10.1016/j.rinp.2021.104578

4. Chu, J. (2021). A statistical analysis of the novel coronavirus (COVID-19) in Italy and Spain. 16(3), 1-13. doi:10.1371/journal.pone.0249037

5. Gholipour, E., Vizvari, B., Babaqi, T., & Takacs, S. (2021). Statistical analysis of the Hungarian COVID-19 victims. Medical virology, 93(12), 6660-6670. doi:10.1002/jmv.27242

6. Giovanatti, A., Elassar, H., Karabon, P., Wunderlich-Barillas, T., & Halalau, A. (2021). Social Determinants of Health Correlating with Mechanical Ventilation of COVID-19 Patients: AMulti-Center Observational Study. International Journal of General Medicine, 14, 8521-8526. doi:10.2147/ijgm.s334593

7. Jess, A., A. D., Mrianne, E., Camila, G., C. E., Diana, H., & Maimuna, S. M. (2021, Octomber 14). Risk factors for increased COVID-19 case-fatality in the United States: A county-level analysis during the first wave. doi:10.1371/journal.pone.0258308

8. Kumar, S. (2020, May 13). Monitoring novel corona virus (COVID-19) infections in India by Cluster analysis. Retrieved from Springer link: https://link.springer.com

9. Pigoli, D., Aston, J., Ferraty, F., Mazumder, A., Richards, C., & Hall, M. (2017, September 02). Estimation of temperature-dependent growth profiles for the assessment of time of hatching in forensic entomology. Retrieved from https://arxiv.org/pdf/1709.00623.pdf

10. Santos, L. D., Stevanato, K. P., Roszkowski, I., & Pelloso, S. M. (2021). Impact of the Covid-19 Pandemic on Women's Health in Brazil. Journal of Multidisciplinary Healthcare, 14, 3205-3211. doi:10.2147/jmdh.s322100

11. Sera, F., Griffiths, L., Dezateux, C., Geraci, M., & Cortina-Borja, M. (2017, November 18). PLOS ONE. doi:10.1371/journal.pone.0187677

12. Umana, H., Fuente, M., Elortegui, G., & Fonseca, F. (2020). Multinomial logistic regression to estimate and predict the perceptions of individuals and companies in the face of the COVID-19 pandemic in the Nuble Region, Chile. Sustainability, 12(22), 1-20. doi:10.3390/su12229553

13. Wolkewitz, M., Lambert, J., von Cube, M., Bugiera, L., Grodd, M., Hazard, D., & Kaier, K. (2021). Statistical analysis of clinical COVID-19 data:A concise overview of lessons learned, common errors and how to avoid them. Clinical Epidemiology, 12, 925-928. doi:10.2147/CLEP.S256735

14. Devkota, J.U.(2022). Vector Auto regression in Forecasting COVID-19 Under-Reporting–Nepal as a Case Study. Journal of Nepal Mathematical Society, NepJol Double Star Journal, December 2022. DOI: https://doi.org/10.3126/jnms.v5i2.50016.

15. Devkota, J.U.(2022). Forecasting deterioration of mental health during COVID-19 pandemic and lockdown - examples from Nepal, Global Journal of Infectious Diseases and Immune Therapies, PUBTEXTO, Volume 4, Issue 3, October 2022. LINK: https://www.pubtexto.com/pdf/?forecasting-deterioration-of-mental-health-during-covid19-pandemic-and-lockdown--examples-from-nepal