Characterizing the epidemiological dynamics of COVID-19 using a non-parametric framework

Sujit Bebortta and Dilip Senapati*

Department of Computer Science, Ravenshaw University, Cuttack 753 003, India

The recently evolved family of coronaviruses known as the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) or COVID-19 has spread across the world at a critically alarming rate, thereby causing a global health emergency. Several nations have been adversely affected in terms of both social and economic aspects. Hence there is an utmost need to control the transmission rate of the virus by incorporating stringent control measures. In this article, a non-parametric framework for characterizing the epidemiological behaviour of COVID-19 is suggested. Several statistical analysis tests have been conducted on the time series data acquired for four countries to derive a relationship between the three considered cases, viz. the new incidence of COVID-19, new deaths, and new sample testing facilities. Further, considering the dynamical behaviour of the sample data, a smoothing spline approach is implemented to obtain a better analysis of the observations. Subsequently, autocorrelation function is used to study the degree of correlation for each considered case for specific time lags. Finally, the non-parametric kernel density estimate is adopted for obtaining a robust characterization of the underlying distributions pertaining to each case considered in this study. Hence these observations lead to the development of an efficient prediction framework that can be implemented for analysing the epidemiological behaviour of the COVID-19 pandemic.

Keywords: Autocorrelation function, COVID-19, epidemiological dynamics, non-parametric statistical tests.

IN recent times, emergence of the 2019 novel corona virus disease (COVID-19) has been observed to induce acute respiratory disorders caused by a family of coronaviruses[1]. On 11 February 2020, the International Committee on Taxonomy of Viruses (ICTV) named the virus as severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2)[2,3]. Since the initial onset of the virus in December 2019, it has intrinsically mutated and affected a large population across the globe. The virus has a highly rapid transmission potential and can easily be a source of infection across different settings like public places, hospitals or other contaminated surfaces. Conceivably, the virus has now spread across 213 countries throughout the world and has proven to be lethal to humans with comorbidities or other acute conditions[4]. The rapidly increasing spread of the virus has attracted attention of the scientific community globally for developing diversified models which can efficiently track and contain the epidemiological spread of COVID-19. Most epidemiological models require exhaustive analysis of empirical data to intrinsically understand the dynamics of a contagion.

To this end, understanding the role of proper statistical data analysis models becomes crucial for developing precise disease diagnostic frameworks[5–7]. Several works have extensively studied the use of statistical data analysis for comprehending the behaviour of diseases and their patterns of spread. However, in real-world scenarios, the distribution of such epidemiological data may largely deviate from some specific distribution models. Hence, non-parametric statistical tests could provide a precise idea regarding the underlying behaviour of a specific dataset. Most sophisticated prediction tools employ these non-parametric tests to obtain a plausible evidence of real-world datasets. After assessing the statistical behaviour of the datasets, several well-known mathematical tools can be implemented to study the relationship between observations from any given dataset[7–12].

According to some composite theories which have evolved recently, it has been observed that increase in rigorous testing facilities has also contributed towards the rise in single-day peaks of COVID-19 cases. However, this indicates a positive influence in the public healthcare domain, which also signifies more traceability of new confirmed COVID-19 cases. Thus, by considering similar cohorts in tracking the epidemiological spread of the virus, more tractable public safety frameworks can be developed by policy makers and health-care organizations, which will aid in concurrently managing and tracking the prevalence of COVID-19 (refs 13–15).

In this article, we consider the reported daily new cases, new deaths and new testing services for COVID-19 pertaining to four worst-affected countries across the world, viz. the United States, the United Kingdom, Russia and India. We performed non-parametric statistical tests like the χ^2 and analysis of variance (ANOVA) to assess statistical behaviour of the considered dataset. It was observed that ANOVA indicated acceptance of the null hypothesis under 95% confidence interval (CI). Thus, our assumptions hold true that the number of cases significantly vary across

*For correspondence. (e-mail: senapatidillip@gmail.com)

individual countries. The smoothing spline function has been adopted for our observations to obtain numerical stability. The respective results for R^2 values have been computed for the corresponding countries along with the individual smoothing parameters. We then report the results using auto-correlation function (ACF) within specific confidence bounds over the time-series data obtained for the four countries considered in this study. From the obtained correlograms, it was observed that the data are in excellent agreement in terms of their correlation over successive intervals of time. Further, non-parametric kernel-based distribution analysis for specifically two kernel types, i.e. Gaussian and Epanechnikov kernels has been done to characterize statistical dynamics of the observed data. Thus, from the above observations it can be deduced that this work leads to a highly efficient prediction framework by deriving a relationship between some critical parameters, which can be used for characterizing the epidemiological impact of the COVID-19 pandemic.

After performing a qualitative analysis of some recent works which focused on assessing the impact of COVID-19 through conventional learners and parameterized probabilistic models, we reached the conclusion that non-parametric models leverage the capability of providing better fit towards characterizing a larger population of functional forms. It was also observed that the use of non-parametric models led to the avoidance of poor assumptions corresponding to the underlying function. Thus, the present work focused on analysing statistical behaviour of reported daily new COVID-19 cases, deaths and testing services through non-parametric tests like the χ^2 and ANOVA. With respect to four worst-affected countries of the world, it was analytically observed that the number of cases showed significant variance across all four countries and depicted acceptance of the null hypothesis under 95% CI. The non-parametric smoothing spline function was employed to capture the nonlinearity and temporal behaviour of the considered COVID-19 dataset. The efficacy of the non-parametric smoothing spline model was observed by computing the respective R^2 values subject to different smoothing parameters. The results for ACF over the time-series data for the considered dataset were illustrated through correlograms for successive time lags. Finally, the results for non-parametric kernel-based distribution models were obtained to characterize the frequency histograms for the datasets of different countries.

Background studies

The cases and fatalities associated with COVID-19 have exacerbated at a highly alarming rate across the world. In this section, different studies employing various predictive disease risk models for determining the spread of COVID-19 are examined. Dimri *et al.*[16] performed a comparative analysis for predicting the characteristics of COVID-19 spread in India. They employed three models, viz. susceptible, infected and recovered (SIR) model, exponential model and a quadratic model for studying the epidemiological outbreak of COVID-19. Chimmula and Zhang[17] employed long short-term memory (LSTM) network to predict the possible trends for COVID-19 cases in Canada and a decline in future cases. It was inferred from their study that the incidence of the virus witnessed a linearly growing trend in the daily new cases, unlike many more prominent nations. Tang *et al.*[18] presented a transmission risk analysis framework by employing the susceptible, exposed, infectious, and recovered (SEIR) model[19]. This is a time-dependent, dynamic transmission control model, which can effectively quantify the strategic evolution of precautionary measures and their impact on reducing COVID-19 transmission. Adhikary *et al.*[20] predicted the outbreak of COVID-19 infection by employing artificial intelligence (AI) techniques subject to different demographics. They proposed the use of Bayesian ridge regressor technique over a wide range of COVID-19 data for predicting future trends of the infection.

Considering the large-scale spread of COVID-19 globally, several anti-contagion measures like social distancing during public interactions, restricting mobility of people, avoiding mass gatherings, using face masks or shields, etc. have largely shown a positive impact in curbing the transmission rate of the virus. The adoption of these policies by local communities as well as national bodies has exceedingly assisted in circumventing the adverse consequences of the virus. Hsiang *et al.*[21] examined the use of reduced form econometric models for analysing the endogenous behaviour of COVID-19 over six localities. The authors assessed empirical impact of the infection rate in light of precautionary measures imposed by national and state administrations. Robertson[22] proposed a Poisson regression model for predicting new cases and deaths associated with COVID-19. This analysis can also be employed for prioritizing high-risk locations and managing critical clinical requirements. Block *et al.*[23] modelled the network of infection spread using a stochastic social network framework. They introduced three social-distancing strategies which assist in flattening the transmission curve and at the same time strengthen the communities. This model can be fundamentally adopted for stabilizing psychosocial behaviour of individuals in a post-lockdown world. Liu *et al.*[24] explored the impact of humidity, temperature and migration index on transmission rate of COVID-19 in 30 provinces of China. The association between virus transmission rate and the above evaluation parameters was determined using a nonlinear regression analysis. Further, data obtained for different cities pertaining to confirmed cases of COVID-19 were fitted with generalized linear models to achieve a city-specific impact on the number of new cases. Table 1 provides a comprehensive account of the works surveyed in this article.

Table 1. Comparative summary of background studies

Author(s)	Features	Model(s)	Outcomes
Dimri *et al.*[16]	Comparative analysis for predicting the characteristics of COVID-19 spread in India	SIR model, exponential model and quadratic model	To contain the spread and monitor the rise in trend of COVID-19 cases
Chimmula and Zhang[17]	Predicted the possible trends for COVID-19 cases in Canada and declination in future cases	LSTM network	Forecast the declination point of COVID-19 outbreak
Tang *et al.*[18]	Presented a transmission risk analysis framework	SIER model	Time-dependent dynamic transmission control model
Adhikary *et al.*[20]	Predicted COVID-19 outbreak by employing artificial intelligence (AI) techniques subject to different demographics	Bayesian ridge regressor and Gaussian ridge regressor	Predicting future trends of infection spread
Hsiang *et al.*[21]	Examined the use of reduced-form econometric models for analysing the endogenous behaviour of COVID-19 over six localities	Econometric models	Develop anti-contagion policy for averting infection rate
Robertson[22]	Proposed a Poisson regression model for predicting new cases and deaths associated with COVID-19	Poisson regression model	Prioritization of high-risk locations and managing critical clinical requirements
Block *et al.*[23]	Modelled the network of infection spread using a stochastic social network framework. The authors introduced three social-distancing strategies	Stochastic social network model	Flattening the infection transmission curve and strengthen communities
Liu *et al.*[24]	Explored the impact of humidity, temperature and migration index on transmission rate of COVID-19 in 30 provinces of China	Nonlinear regression analysis	Provide city-specific impact on the number of new COVID-19 cases

Table 2. Summary of various cases corresponding to four different countries

Country	New cases	New deaths	New tests
USA	5,573,847	174,255	69,576,538
UK	322,280	41,403	12,076,636
India	2,905,823	54,849	32,526,364
Russia	942,106	16,099	33,096,427

In view of the above studies, here we propose a non-parametric framework to provide better convergence in characterizing large datasets and make predictions-based on temporal characteristics of the COVID-19 dataset. Initially we employ non-parametric tests like χ^2 and ANOVA to observe the performance of the distribution models and assess the significance of the sample data for the mean between them. The non-parametric smoothing spline function was employed, which reduces the complexity of using high-order polynomials as encountered with other conventional regression models and leverages flexible fits for the considered COVID-19 dataset. ACF was provided pertaining to time-series COVID-19 data over successive time lags to understand the temporal behaviour on new cases, deaths and tests performed. Further, kernel-based density estimation technique was adopted for the characterization of different density histograms.

Methodology

Data source

The dataset was obtained from https://ourworldindataorg/, which consists of country-wise COVID-19 dataset. The dataset when acquired consisted of temporal data between 31 December 2019 and 21 August 2020. In this study, we have considered datasets for four countries, namely the US, the UK, Russia and India. For our analysis, we have extracted data between 1 March 2020 and 21 August 2020. The reason for this is because the cases of COVID-19 started to become prominent only from March 2020. We have considered three fundamental attributes for analysis, namely incidence of new cases, new deaths and the number of samples tested each day. We further performed extensive statistical analysis on the data to illustrate the correlation between these attributes. Table 2 provides a descriptive summary of the three case types for the four different countries considered in this study.

Proposed framework

The contributions of this work are multifold and lead to efficient characterization of COVID-19 incidence. The COVID-19 dataset collected from public repositories has been selectively analysed for specifically four countries across the world. The dataset has been tested for its statistical significance using non-parametric statistical tests such as the χ^2 and ANOVA. A detailed description of the test hypothesis and respective test statistics are presented in later sections. Here, the hypothesis corresponding to ANOVA is accepted for our assumption that the number of cases corresponding to each country is independent. We then incorporate the smoothing spline function for analysing the observations corresponding to different

groups such as new cases, new deaths and new testing facilities reported for each country. Further, the use of autocorrelation analysis has been reported for different critical parameters like number of daily cases, new deaths and testing facilities for individual countries over different lags. This strategy is essential for policy makers and clinicians to understand the relationship between individual parameters and predict their consequent outcomes for successive lags. Finally, we fit the datasets for all the respective attributes with the non-parametric kernel density model pertaining to different countries. The model is fitted over the dataset which reports the new cases, deaths and testing facilities from COVID-19 at individual country level for different kernel types. This provides precise observation for the statistical dynamics of each attribute for characterizing the relationship between their growth rate and interdependencies, and may lead to forecasting of future trends in monitoring epidemiological behaviour of the virus. Figure 1 illustrates a high-level workflow analysis of the proposed framework for processing and characterization of COVID-19 datasets.

Statistical significance tests

χ^2 non-parametric test

The χ^2 non-parametric test is a well-known statistical test which is used to study the dependencies among features of a dataset[6,25]. It has found extensive applicability in biomedical systems for hypothesis testing and evaluation of possible underlying dependencies. Hence, the χ^2 test for the purpose of the study can be stated as follows

$$\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \tag{1}$$

where O_{ij} and E_{ij} represent the observed and expected number of cases against the four countries respectively. Here, i varies from 1 to 4, representing the four countries which we have considered for our analysis and j varies from 1 to 3 which depicts the three attributes, namely new cases, new deaths, and new samples tested per day.

Null hypothesis (H_0): The number of cases is independent corresponding to the four countries. The calculated χ^2 value is greater than the tabulated χ^2 value under 95% confidence (i.e. 5 level of significance) with $(4-1)(3-1) = 6$, and the degree of freedom is given as $\chi^2_{6,0.05} = 1.635$. Hence, the null hypothesis is rejected, indicating that the number of various cases is dependent on individual countries.

Analysis of variance

ANOVA is an extensively used non-parametric statistical test which has found extensive applications in a variety of scientific domains[6,26–28]. It is a robust statistical test which is imperatively used for analysing observations with distinct distributions.

Let the variate c_{ij} represent different cases corresponding to the four countries. The sum of squared deviations for all cases of various countries is defined as

$$Q = \sum_{i=1}^{4}\sum_{j=1}^{3}c_{ij}^2 - \frac{T^2}{N},\ T = \sum_{i=1}^{4}\sum_{j=1}^{3}c_{ij}, \tag{2}$$

where N is the total number of variates and n_i is the number of cases within the ith country.

The sum of squared deviations between countries is defined as

$$Q1 = \sum_{i=1}^{4}\frac{T_i^2}{n_i} - \frac{T^2}{N}, T_i = \sum_{j=1}^{4}c_{ij}. \tag{3}$$

The sum of squared deviations within countries can be defined as

$$Q2 = Q - Q1. \tag{4}$$

Null hypothesis (H_0): The number of cases is significant with respect to each country.

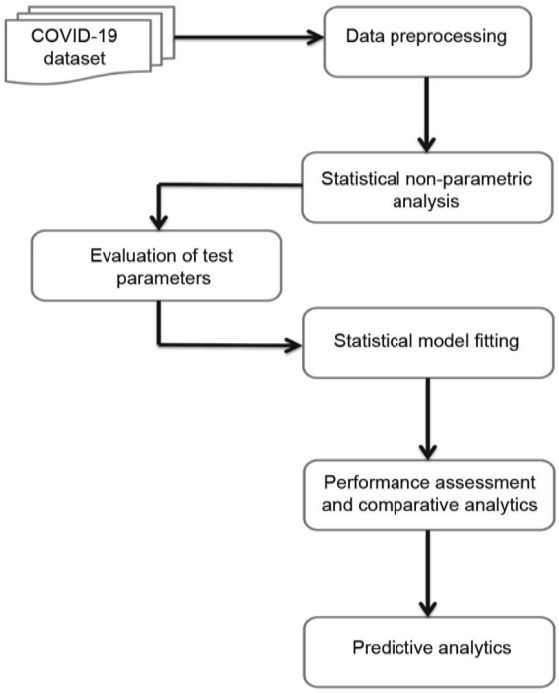

Figure 1. Proposed framework for processing and analysing the COVID-19 dataset.

Table 3. ANOVA table for classification

Source of variance	Sum of squares	d.f.	Mean square	Variance ratio
Between countries	$Q1 = 6.9E + 18$	$h - 1 = 3$	2.3E + 17	5.5E + 18/2.3E + 17 = 2.42
Within countries	$Q2 = 4.4E + 19$	$N - h = 8$	5.5E + 18	–

The degrees of freedom between countries is defined as the number of countries subtracted by 1 (i.e. $\gamma_1 = h - 1 = 4 - 1 = 3$). Similarly, the degrees of freedom within countries is defined as the number of variates subtracted from the number of countries (i.e. $\gamma_2 = N - h = 12 - 4 = 8$). The F tabulated value corresponding to the degrees of freedom ($\gamma_1 = 3$, $\gamma_2 = 8$), under 95% CI is obtained as $F_{0.05}$ ($\gamma_1 = 3$, $\gamma_2 = 8$) = 8.84, which is greater than the calculated F-value = 2.42 (Table 3). Hence, the null hypothesis is accepted, indicating that the number of cases varies significantly with respect to individual countries.

Non-parametric smoothing spline model

The data used for predicting the actual trend pertaining to various cases such as new cases, new deaths and new tests are noisy and aperiodic. Therefore, the conventional trend analysis models are driven by many parameters to capture the nonlinearity pattern in the data. It is also a promising challenge to estimate the fitting parameters or perfect weights of the model using real datasets[29]. This article uses non-parametric smoothing spline fitting approach to capture the temporal behaviour of the COVID-19 data. Here, the model only depends on the smoothing parameter α, which establishes a smooth line between the two intervals of a dataset.

Let (x_i, y_i), $i = 0(1)n - 1$ be the n set of data points. A smoothing cubic spline $S(x)$, which satisfies the following conditions is used

- $S(x_i) = y_i$, $i = 0(1)n - 1$.
- $S(x)$, $S'(x)$ and $S'''(x)$ are continuous at defined intervals.
- $S(x)$ is a cubic polynomial over each defined interval.

The smoothing spline minimizes the following expression

$$L(\alpha, w_i) = \alpha \sum_i w_i [y_i - s(x_i)]^2 + (1 - \alpha) \int (\nabla^2 s)^2 \, \mathrm{d}x, \quad (5)$$

where $\nabla^2 = \mathrm{d}^2/\mathrm{d}x^2$ is a second-order differential operator.

The value of the smoothing parameter lies between $0 \le \alpha \le 1$. If $\alpha = 0$, the smoothing splines produce least-square straight line, whereas for $\alpha = 1$, it acts as a cubic interpolant. The optimal range of α is defined at the near of order $(1 + h^3/6)^{-1}$, where h is the average spacing between the two consecutive data points.

Autocorrelation function

ACF provides a measure of stochastic process memory which can be used to evaluate the correlation between certain variables over specific time lags[30–32]. Autocorrelation provides the degree of randomness and fluctuation in a spatio-temporal dataset with respect to its past memory lag-l. The autocorrelation structure helps select proper models corresponding to different datasets and it satisfies the bias-variance trade-off (i.e. type-I error rate (falsely accepting) and type-II error rate (true rejecting))[33,34].

Let C_t represent stochastic time series for the number of new cases, new deaths or new tests for the COVID-19 scenario. It is important to define the dependence between C_t and C_{t-l}, where l is the time lag of the time-series data. The correlation coefficient between C_t and its lag series C_{t-l} is called autocorrelation. It is defined as follows

$$\rho_l = \frac{\mathrm{cov}(C_t, C_{t-l})}{\sqrt{\mathrm{var}(C_t)\,\mathrm{var}(C_{t-l})}}. \quad (6)$$

For $l = 0$, $\rho_0 = 1$ and $\rho_l = \rho_{-l}$. The value of ρ_l lies between –1 and 1. Autocorrelation plays an important role for evaluation of the power spectrum of the time-series, i.e. for calculation of the frequency content of the time series. In the present COVID-19 scenario, the number of new cases, new deaths or new tests intrinsically depends on the past few days. Therefore, ρ_l converges to zero for large l, indicating that the dependency of C_t diminishes for past long memories of time series.

Non-parametric kernel estimation

As the data used for analysis of the dynamics of COVID-19 are highly unsystematic and unpredictable, they are not well characterized and captured by the popular conventional probability density functions[35–37]. However, a kernel-based non-parametric distribution can characterize and explain such data, where the estimation of parameters for the known probability distribution cannot be properly done. A kernel-based non-parametric distribution has more advantage over complicated and complex mixture distributions whose parameter estimations are quite impossible through the method of moment matching and log-likelihood estimation procedures. The kernel density

over the cases from an unknown distribution with n sample points is defined as

$$\hat{f}_h(c) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{c-c_i}{h}\right), \tag{7}$$

where $K(.)$ is the kernel smoothing function of bandwidth size h. The kernel $K(.)$ satisfies conditions such as

$$\frac{1}{h}K\left(\frac{c-c_i}{h}\right) \geq 0 \text{ and } \int \frac{1}{h}K\left(\frac{c-c_i}{h}\right)\mathrm{d}c = 1. \tag{8}$$

The average of smoothing kernel $K(.)$ function helps define a new probability density function as $\hat{f}_h(c)$. This article uses two well-known kernel types, Gaussian and Epanechnikov, to describe the characteristics of the data. The Gaussian kernel is defined as follows

$$K_h(c) = \frac{1}{h\sqrt{2\pi}}\mathrm{e}^{-((1/2)(c/h)^2)}. \tag{9}$$

The Epanechnikov kernel is defined as follows

$$K_h(c) = \frac{3}{4h}\left(1-\left(\frac{c}{h}\right)^2\right) I\left(\left|\frac{c}{h}\right| \leq 1\right), \tag{10}$$

where $I(x)$ is an indicator function such that $I(x) = 1$ for $|x| \leq 1$ and $I(x) = 0$ otherwise. Following Tasy[38], the optimal bandwidth $(\hat{h})$ for the non-parametric kernel density models which provide proper trade-off between overfitting and underfitting is defined as

$$\hat{h} = \begin{cases} 1.06 s\tau^{-(1/5)} & \text{for the Gaussian kernel} \\ 2.34 s\tau^{-(1/5)} & \text{for the Epanechnikov kernel,} \end{cases} \tag{11}$$

where s and τ are standard error of the sample and sample size respectively.

Results and discussion

The acquired data were pre-processed employing column average method to substitute the missing values in them[39]. This is essential as most models do not support missing values and hence cause difficulty in convergence. We first deal with the results obtained for smoothing spline fit over the COVID-19 data considered in this study along with the respective smoothing parameters (P) and R^2 values for each country. Then the results for autocorrelation analysis are obtained and correlation between the observations is discussed. Finally, the empirical datasets for all four countries are fitted with a non-parametric kernel density model for Gaussian and Epanechnikov kernels to provide a characterization of statistical distribution of the new cases, new deaths and number of samples tested per day.

Non-parametric smoothing spline

The non-parametric smoothing spline model is a widely used technique for handling nonlinear observations. In Figure 2 *a*, the smoothing spline fit for new COVID-19 cases in the US are presented. The smoothing parameter (P) here is considered as 0.9000, whereas the R^2 value is given as 0.9995 and the adjusted R^2 value is 0.9987. The goodness-of-fit pertaining to this model is indicated by high R^2 values, where R^2 value closer to 1 indicates very low difference between the fitted value and observable data points. Figure 2 *b* gives the smoothing spline fit for new COVID-19 deaths reported each day within the considered time stamp. Here, the R^2 value is found to be 0.9588 and the adjusted R^2 value is 0.8974 for smoothing parameter $P = 0.9000$. Figure 2 *c* shows smoothing spline fit over the new samples tested per day. The R^2 value is observed to be 0.9930 and the adjusted R^2 value is obtained as 0.9825 for $P = 0.9000$. Table 4 provides a detailed summary of the statistics.

Figure 2 *d* provides a smoothing spline fitted over the daily new cases reported in the UK. The R^2 value is observed to be 0.9904, whereas the adjusted R^2 value is obtained as 0.9761 with smoothing parameter $P = 0.9000$. In Figure 2 *e*, the death cases per day in the UK fitted with smoothing spline model for the considered time interval are reported. Here, the R^2 value is obtained as 0.9624 and the adjusted R^2 value is 0.9063 for $P = 0.9000$. Figure 2 *f* provides the COVID-19 samples tested per day with the fitted smoothing spline for $R^2 = 0.9919$ and adjusted $R^2 = 0.9797$.

Figure 2 *g* shows the new COVID-19 cases reported per day in India along with the fitted smoothing spline model. The model obtains R^2 value of 0.9908 and adjusted R^2 value 0.9772 with smoothing parameter $P = 0.9000$.

Table 4. Summary for goodness of fit

Country	Details	R^2	P
USA	New cases	0.9995	0.9000
	New deaths	0.9588	0.9000
	New tests	0.9930	0.9000
UK	New cases	0.9904	0.9000
	New deaths	0.9624	0.9000
	New tests	0.9919	0.9000
India	New cases	0.9908	0.9000
	New deaths	0.9339	0.9000
	New tests	0.9918	0.9000
Russia	New cases	0.9956	0.9000
	New deaths	0.9746	0.9000
	New tests	0.9956	0.9000

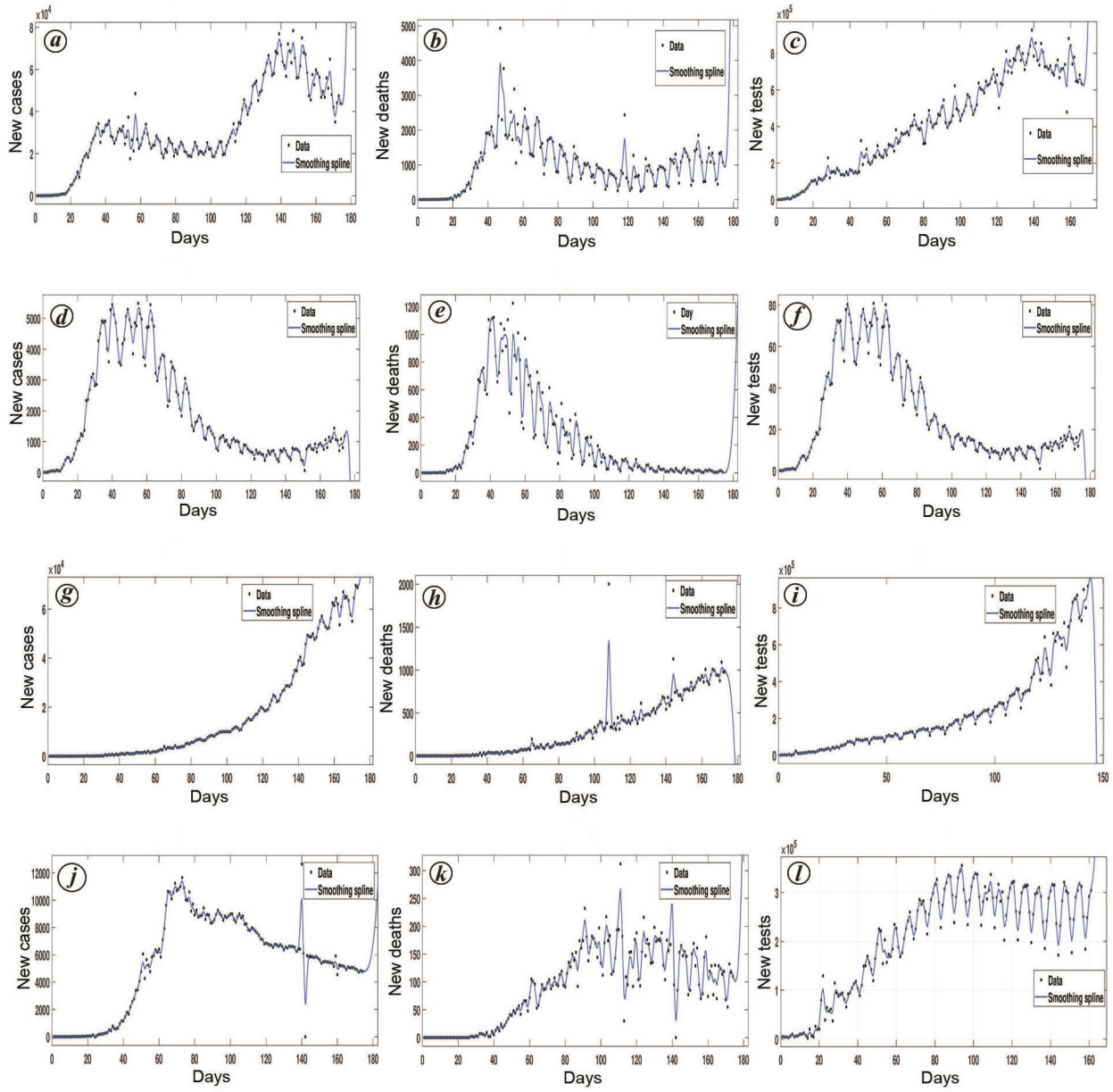

Figure 2. ***a*–*c***, Smoothing spline fitting over the US. ***a***, New cases; ***b***, new deaths; ***c***, new tests. ***d*–*f***, Smoothing spline fitting over the UK. ***d***, New cases; ***e***, new deaths; ***f***, new tests. ***g*–*i***, Smoothing spline fitting over India. ***g***, New cases, ***h***, new deaths; ***i***, new tests. ***j*–*l***, Smoothing spline fitting over Russia. ***j***, New cases; ***k***, new deaths; ***l***, new tests.

Figure 2 *h* shows the smoothing spline fitted with new death cases reported in a day. The R^2 and adjusted R^2 values are obtained as 0.9339 and 0.8352 respectively for the same P value. In Figure 2 *i*, data for new COVID-19 samples tested in a day in India are fitted with the smoothing spline model. For $P = 0.9000$, the R^2 value is obtained as 0.9918 and the adjusted R^2 value was 0.9795.

In Figure 2 *j*, the new COVID-19 cases observed per day in Russia are fitted with the smoothing spline model. The R^2 value here is observed to be 0.9956, whereas the adjusted R^2 value is 0.9891 for smoothing parameter $P = 0.9000$. As observed from Figure 2 *k*, the R^2 value for new death cases reported in a day fitted with the smoothing spline model is 0.9746 and the adjusted R^2 value is 0.9367 for the same P value. In Figure 2 *l*, data for new samples tested in a day for the specified time interval fitted with the smoothing spline model are provided. Here, the R^2 value is observed to be 0.9956 and adjusted R^2 value is 0.9891 for the same P value.

Autocorrelation function analysis

ACF analysis provides the characterization of future events in compliance with the past events. Hence, it is empirically observed that ACF exhibits some memory behaviour with respect to time series. This tends to

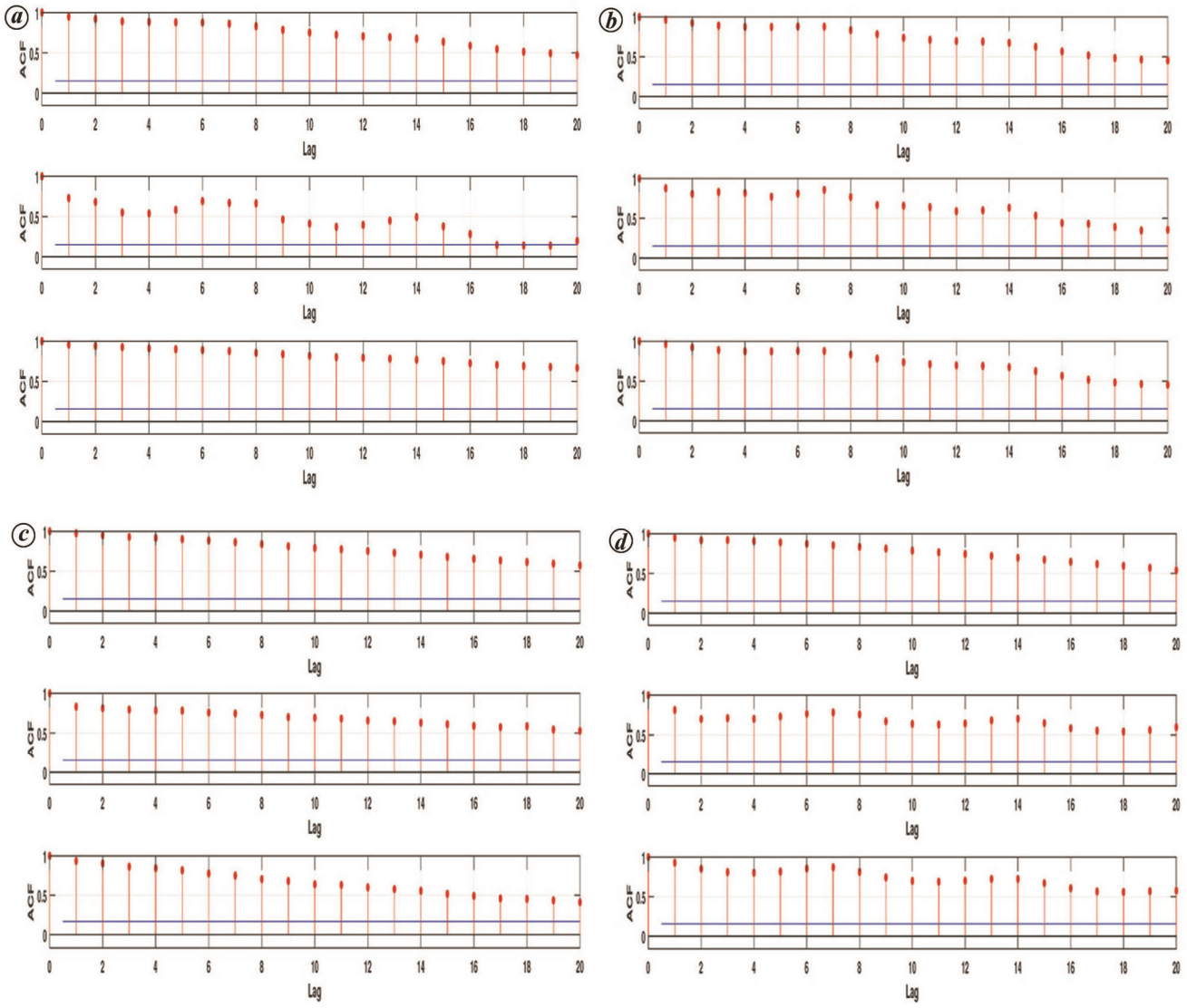

Figure 3. ***a***, Autocorrelation function (ACF) of the US dataset corresponding to new cases, new deaths and new tests. ***b***, ACF of the UK dataset corresponding to new cases, new deaths and new tests. ***c***, ACF of India dataset corresponding to new cases, new deaths and new tests. ***d***, ACF of Russia dataset corresponding to new cases, new deaths and new tests.

influence future predictions in terms of the previously made observations in an observation space. Figure 3 *a* shows the observations made from ACF pertaining to the three attributes considered in this study, viz. new COVID-19 cases, newly reported deaths, and new samples tested each day for the US. It can be observed that the new cases obtain high correlation at lags 0, 1, 2, 4, 6 and 8, thereby indicating positive autocorrelation. In the case of newly reported deaths, high correlation is observed at lags from 0 to 6. Further, for the new COVID-19 samples tested, high correlation is observed at lags 0 to 8.

The correlograms in Figure 3 *b* refer to the newly reported cases, deaths and samples tested for COVID-19 in the UK for ACF analysis. For the new cases, there is high correlation at lags from 0 to 6. For new deaths, it can be observed from the respective correlogram that there is high correlation for lags 0, 1, 3, 4 and 6. Finally, for the ACF analysis performed upon the new tests, the function is observed to obtain high correlation at lags from 0 to 7.

Figure 3 *c* provides the correlograms obtained for ACF analysis on new cases, deaths and new samples tested in India for COVID-19. The ACF for new cases in India shows high correlation at lags 0 to 4. The ACF for new deaths shows high correlation at lags 0 to 6. Lastly, the ACF analysis for new samples tested in India shows high correlation at lags 0 to 5.

In Figure 3 *d*, the correlograms for Russia are provided corresponding to new cases, new deaths, and new samples tested for the ACF analysis. Here, the ACF for new cases

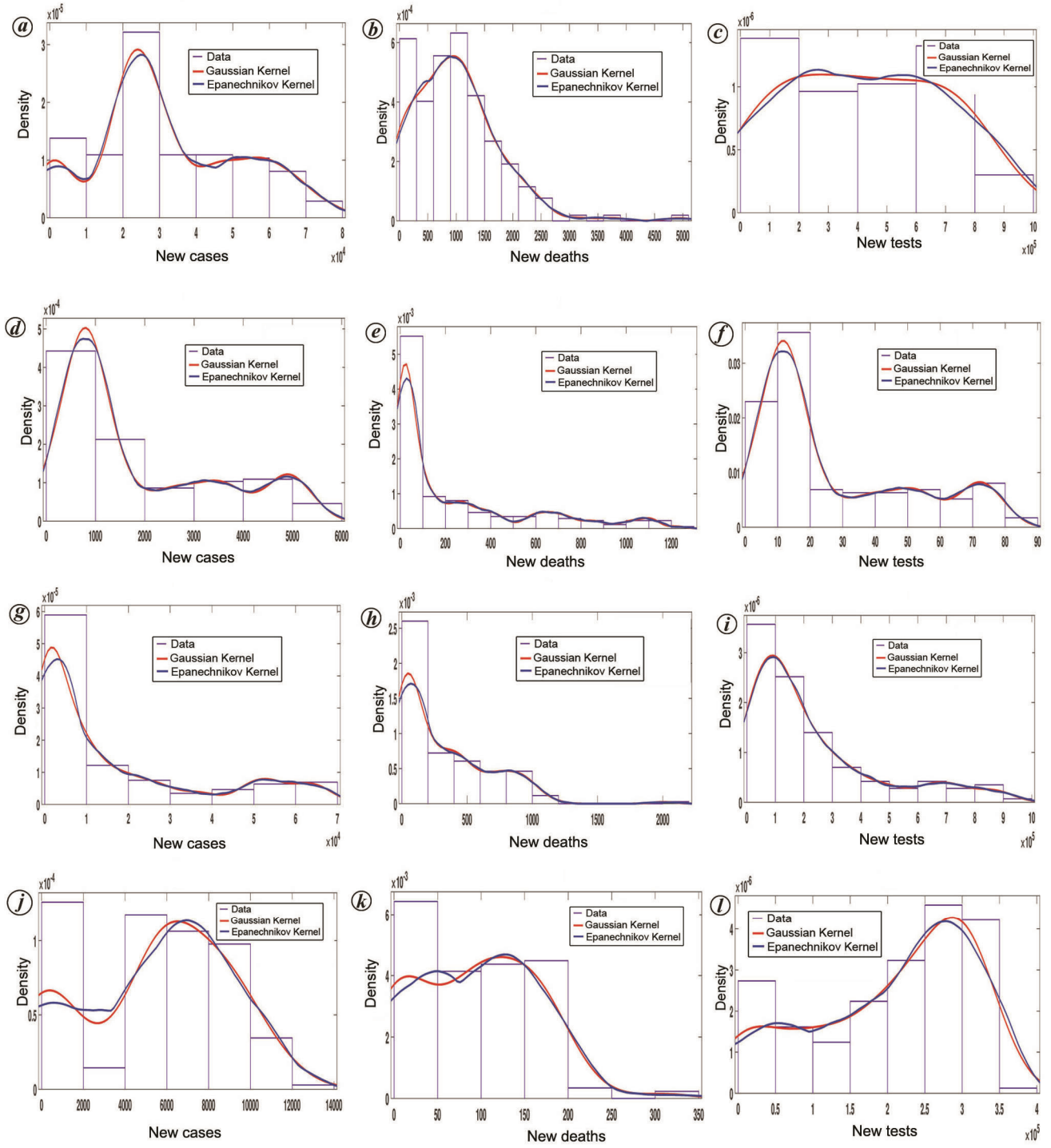

Figure 4. ***a*–*c***, Density estimation of the US data using Gaussian and Epanechnikov kernels. ***a***, New cases; ***b***, new deaths; ***c***, new tests. ***d*–*f***, Density estimation of the UK data using Gaussian and Epanechnikov kernels. ***d***, New cases; ***e***, new deaths; ***f***, new tests. ***g*–*i***, Density estimation of India data using Gaussian and Epanechnikov kernels. ***g***, New cases; ***h***, new deaths; ***i***, new tests. ***j*–*l***, Density estimation of Russia data using Gaussian and Epanechnikov kernels. ***j***, New cases; ***k***, new deaths; ***l***, new tests.

shows high correlation at lags 0 to 5. However, in of case of new deaths, there is high correlation at lags 0, 1 and 7. The ACF for new samples tested shows high correlation at lags 0, 1, 2 and 6.

Non-parametric kernel density estimation

The non-parametric kernel density estimation model is a well-known statistical model which considers finitely

sampled data to draw statistical inference over a population. For our analysis, we have considered two kernel types, i.e. Gaussian and Epanechnikov kernels, which are fitted over the data acquired for the four countries. Figure 4 *a* provides the density estimation over the new cases of COVID-19 in the US for both kernel types. Figure 4 *b* presents the density analysis of new death cases in the US corresponding to kernel density model. Further, the density of new samples tested is characterized by employing the non-parametric kernel density model as shown in Figure 4 *c*.

The distribution of new COVID-19 cases in the UK is characterized using the kernel density estimate for Gaussian and Epanechnikov kernels. Figure 4 *d* shows corresponding distribution plot along with the fitted data. Figure 4 *e* shows data corresponding to new deaths reported per day along with the fitted non-parametric kernel density model. Lastly, as shown in Figure 4 *f*, the distribution of new tests is fitted with the kernel density model.

Figure 4 *g* provides the distribution of new incidence of COVID-19 in India along with the kernel density model for the Gaussian and Epanechnikov kernel types. In Figure 4 *h*, the distribution of samples for new deaths reported in India are presented, where the kernel density model is fitted over the sample data for the above-mentioned kernel types. Figure 4 *i* presents the density of new samples tested in a day over the specified time period along with the kernel density model. This provides precise knowledge to the health authorities and policy makers to understand the underlying distribution pattern of the data points. Hence such characterization is crucial for understanding the statistical behaviour of epidemiological data.

The distribution of data corresponding to new cases, new deaths and new tests carried out on samples of patients in Russia are shown in Figure 4 *j*–*l* respectively. Figure 4 *j* presents the kernel density model fitted over sample distribution for Gaussian and Epanechnikov kernel types representing the number of new cases. Figure 4 *k* represents the density of new deaths characterized by kernel density model for the above kernel types. Figure 4 *l* shows the distribution of new COVID-19 samples tested, which is fitted with the non-parametric kernel density model. The article uses weighted average kernel methods to fit the PDFs of the corresponding non-parametric kernel-based models, where the centre of each histogram bin provides more weight to points close to the bin at the centre and less weight to points far away from the bin of the centre. Therefore, the model may not capture the catastrophic (i.e. extreme outlier) events in the considered dataset[38].

Conclusion and future work

The incidence of COVID-19 across the world has caused a global health emergency. This has critically impacted the public health sector as well as the economic and societal growth of several countries. Thus, it has become inevitable to manage the transmission of the COVID-19 virus and impose stringent control policies like effective contact tracing and increasing number of testing facilities across the world. This article proposes a non-parametric framework for precise characterization of the epidemiological dynamics of COVID-19 and analysing its correlation with other crucial factors such as sample testing facilities, influence on the number of daily confirmed cases, mortality rate, etc. We considered the impact of COVID-19 and its associated factors on four countries, namely the US, the UK, India and Russia. Non-parametric statistical tests such as the χ^2 and ANOVA were performed to analyse the dependency of individual countries with the incidence of COVID-19 cases. It was observed that the ANOVA test accepted the null hypothesis with a confidence level of 95%, thereby indicating that our assumptions hold true. Further, non-parametric smoothing spline approach was implemented for robustly inferring the highly nonlinear behaviour of the sample data. The respective R^2 and adjusted R^2 values were presented for each country pertaining to different cases. ACF analysis was performed to obtain the correlation for each country for specific time lags. Finally, the non-parametric kernel density estimate for two kernel types, viz. Gaussian and Epanechnikov kernels was incorporated to characterize the underlying distribution for the sample data. Hence, the observations made in this study lead towards the development of a highly efficient prediction system which can be used to observe the variability in future trends of the COVID-19 pandemic.

1. Wang, Li-Sheng, Wang, Yi-Ru, Ye, Da-Wei and Qing-Quan, L., A review of the 2019 novel coronavirus (COVID-19) based on current evidence. *Int. J. Antimicrob. Agents*, 2020, **61**(2), 105,948.
2. Rademaker, M., Baker, C., Foley, P., Sullivan, J. and Wang, C., Advice regarding COVID-19 and use of immunomodulators, in patients with severe dermatological diseases. *Australiasm J. Dermatol.*, 2020, **61**(2), 158–159.
3. Yang, P. and Wang, X., COVID-19: a new challenge for human beings. *Cell. Mol. Immunol.*, 2020, **17**(5), 555–557.
4. Kampf, G., Todt, D., Pfaender, S. and Steinmann, E., Persistence of coronaviruses on inanimate surfaces and their inactivation with biocidal agents. *J. Hosp. Infect.*, 2020, **104**(3), 246–251.
5. Hebel, J. R. and McCarter, R., *Study Guide to Epidemiology and Biostatistics*, Jones & Bartlett Publishers, Burlington, Massachusetts, 2011.
6. Sahoo, H., Senapati, D., Thakur, I. S. and Naik, U. C., Integrated bacteria–algal bioreactor for removal of toxic metals in acid mine drainage from iron ore mines. *Bioresour. Technol. Rep.*, 2020, **11**, 100,422.
7. Senapati, D. and Karmeshu, Generation of cubic power-law for high frequency intra-day returns: maximum tsallis entropy framework. *Digit. Signal Process.*, 2016, **48**, 276–284.
8. Bebortta, S. *et al.*, Evidence of power-law behavior in cognitiveiot applications. *Neural Comput. Appl.*, 2020, **32**, 1–13.
9. Bebortta, S., Kumar Singh, A. K., Mohanty, S. and Senapati, D., Characterization of range for smart home sensors using tsallis'

entropy framework. In *Advanced Computing and Intelligent Engineering*, Springer, Singapore, 2020, pp. 265–276.
10. Nayak, G., Singh, A. K. and Senapati, D., Computational modeling of non-Gaussian option price using non-extensive tsallis' entropy framework. *Comput. Econ.*, 2020, **57**, 1–19.
11. Mukherjee, T., Singh, A. K. and Senapati, D., Performance evaluation of wireless communication systems over Weibull/*q*-lognormal shadowed fading using tsallis' entropy framework. *Wireless Pers. Commun.*, 2019, **106**(2), 789–803.
12. Fernandes, A. A. R. *et al.*, Comparison of curve estimation of the smoothing spline nonparametric function path based on PLS and PWLS in various levels of heteroscedasticity. In *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, 2019, vol. 546, p. 052,024.
13. Schultz, M. B., Vera, D. and Sinclair, D. A., Can artificial intelligence identify effective COVID-19 therapies? *EMBO Mol. Med.*, 2020, **12**(8), e12817.
14. Chen, L. *et al.*, Disease progression patterns and risk factors associated with mortality in deceased patients with COVID-19 in Hubei Province, China. *Immunity Inflamm. Dis.*, 2020, **8**(4), 584–594.
15. Bebortta, S., Panda, M. and Panda, S., Classification of pathological disorders in children using random forest algorithm. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), IEEE, 2020, pp. 1–6.
16. Dimri, V. P., Ganguli, S. S. and Srivastava, R. P., Understanding trend of the COVID-19 fatalities in India. *J. Geol. Soc. India*, 2020, **95**(6), 637.
17. Reddy, V. K. R. and Zhang, L., Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons Fractals*, 2020, p. 109864.
18. Tang, B., Bragazzi, N. L., Li, Q., Tang, S., Xiao, Y. and Wu, J., An updated estimation of the risk of transmission of the novel coronavirus (2019-nCoV). *Infect. Dis. Model.*, 2020, **5**, 248–255.
19. Tang, B., Wang, X., Li, Q., Bragazzi, N. L., Tang, S., Xiao, Y. and Wu, J., Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. *J. Clin. Med.*, 2020, **9**(2), 462.
20. Adhikary, S., Chaturvedi, S., Chaturvedi, S. K. and Banerjee, S., COVID-19 spreading prediction and impact analysis by using artificial intelligence for sustainable global health assessment. In *Advances in Environmental Engineering Management*, Springer, Cham, 2021, pp. 375–386.
21. Hsiang, S. *et al.*, The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature*, 2020, **584**(7820), 262–267.
22. Robertson, L. S., COVID-19 confirmed cases and fatalities in 883 US counties with a population of 50,000 or more: predictions based on social, economic, demographic factors and shutdown days. *medRxiv*, 2020.
23. Block, P. *et al.*, Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. *Nature Hum. Behav.*, 2020, 1–9.
24. Liu, J. *et al.*, Impact of meteorological factors on the COVID-19 transmission: a multi-city study in China. *Sci. Total Environ.*, 2020, **26**, 138513.
25. Yuan, Z., Ji, J., Zhang, T., Liu, Y., Zhang, X., Chen, W. and Xue, F., A novel chi-square statistic for detecting group differences between pathways in systems epidemiology. *Stat. Med.*, 2016, **35**(29), 5512–5524.
26. Lin, D. *et al.*, Evaluations of the serological test in the diagnosis of 2019 novel Coronavirus (SARS-COV-2) infections during the COVID-19 outbreak. *Eur. J. Clin. Microbiol. Infect. Dis.*, 2020, **39**, 1–7.
27. Magagnoli, J. *et al.*, Outcomes of hydroxychloroquine usage in United States veterans hospitalized with COVID-19. *Med*, 2020.
28. Hu, S. *et al.*, Weakly supervised deep learning for COVID-19 infection detection and classification from CT images. *IEEE Access*, 2020, **8**, 118869–118883.
29. Farrow, D. C. *et al.*, A human judgment approach to epidemiological forecasting. *PLOS Comput. Biol.*, 2017, **13**(3), e1005248.
30. Hudson, J., Fielding, S. and Ramsay, C. R., Methodology and reporting characteristics of studies using interrupted time series design in healthcare. *BMC Med. Res. Methodol.*, 2019, **19**(1), 137.
31. Walczak, B., *Wavelets in Chemistry*, Elsevier, 2000.
32. Satija, U., Ramkumar, B. and Sabarimalai Manikandan, M., Automated ECG noise detection and classification system for unsupervised healthcare monitoring. *IEEE J. Biomed. Health Informat.*, 2017, **22**(3), 722–732.
33. Kühn, I., Incorporating spatial autocorrelation may invert observed patterns. *Divers. Distribut.*, 2007, **13**(1), 66–69.
34. Dormann, C. F. *et al.*, Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 2007, **30**(5), 609–628.
35. Irvine, M. A. and Hollingsworth, T. D., Kernel-density estimation and approximate bayesian computation for flexible epidemiological model fitting in python. *Epidemics*, 2018, **25**, 80–88.
36. Pereira-Franchi, E. P. L. *et al.*, Molecular epidemiology of methicillin-resistant *Staphylococcus aureus* in the Brazilian primary health care system. *Trop. Med. Int. Health*, 2019, **24**(3), 339–347.
37. Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J. and Rosenfeld, R., Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLOS Comput. Biol.*, 2018, **14**(6), e1006134.
38. Tsay, R. S., *Anal. Financial Time Series*, John Wiley, 2005, vol. 543.
39. Choong, M. K., Charbit, M. and Yan, H., Autoregressive model-based missing value estimation for DNA microarray time series data. *IEEE Trans. Inform. Technol. Biomed.*, 2009, **13**(1), 131–137.

ACKNOWLEDGEMENT. We thank OURIIP, Odisha, India for extending their support to initiate the research under Grant No. 28Seed/2019/COMP.SG and Engg.-5 and provide the adequate research environment.

Received 24 February 2021; revised accepted 25 January 2022

doi: 10.18520/cs/v122/i7/790-800