ISSN: 2754-4745

Journal of Physical Medicine Rehabilitation Studies & Reports

Research Article

Open Access

Predictive Modeling for Coronavirus Pandemic: A Time Series Analysis

Suresh Kumar Sharma*, Preeti, Gurveer Kaur, Shalini and Arshika

Department of Statistics, Panjab University, Chandigarh, Pin Code-160014, India

ABSTRACT

The COVID-19 pandemic has led to a dramatic loss of human life worldwide. India, with a population of more than 1.34 billion - the second largest population in the world have faced acute difficulty in controlling the transmission of Coronavirus among its population, particularly during the second wave. This results into serious repercussions on mortality and morbidity in India. In this Study, the secondary data which is available on public domain of Government of India and Other Countries was used. This data is extracted from different websites from 1 April, 2020 to 30 April, 2021 for a total of 395 days. This data consists of – cumulative confirmed cases, active cases, recovered cases, the actual deaths per day and cumulative deaths. The association between daily confirmed cases and mortality was established using the generalized additive model (GAM) with natural and penalized spline smoothers at (6,2,2) degrees of freedom in R software with mortality as a dependent variable. Smoothers for day of the week, active cases, actual active cases were also included in the model. In the corresponding period mortality rate on an average 533 is deaths per day. The association between daily confirmed cases and daily mortality was found to be statistically significant. The relative risk has been computed for increase in every 1000 number of daily confirmed cases. For increase in 1000 number of daily confirmed cases, the expected number of deaths amounts to 1.006702 (approximately). The entire data has been divided into seven zones and for each zone GAM Model was fitted to make future prediction. The analysis box plots, smoothing plots along with PACF plots, residual plots and predicted plots.

The daily confirmed cases were significantly associated with mortality. The study shows that there is need to improve awareness of Coronavirus and infrastructure facilities in hospitals in India. Moreover, there is also need to improve and make a transparent registration system of deaths due to this killer virus.

***Corresponding author**

Suresh Kumar Sharma, Department of Statistics, Panjab University, Chandigarh, Pin Code-160014, India.

Received: February 17, 2025; **Accepted:** February 21, 2025; **Published:** March 05, 2025

Introduction

The COVID-19 global pandemic has generated an abundance of research quickly following the outbreak. In early December 2019, an outbreak of coronavirus disease 2019 (COVID-19), caused by a novel severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2), occurred in Wuhan City, Hubei Province, China. On January 30, 2020 the World Health Organization (WHO) declared the outbreak as a Public Health Emergency of International Concern. Recent studies reveal that more impact of Coronavirus was more on elderly and those suffering from underlying medical conditions such as circulatory and respiratory diseases [1-4].

Every continent in the world has been affected by this highly contagious disease, with nearly a million cases diagnosed in over 200 countries worldwide. Coronaviruses are a family of viruses that can cause mild to moderate upper-respiratory tract illnesses such as the common cold, severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS). When an infected person coughs or sneezes, the new Coronavirus may be transmitted through expelled droplets. These droplets can enter a person's system through "contact routes," such as the mouth, eyes, or nose. It is also possible for the droplets to be inhaled into the lungs. It is also important to note the incubation period of the coronavirus. The "incubation period" means the time between catching the virus and the time when symptoms of the disease begin to emerge. This is a critical period of time because when people do not know they have the disease; they may not be as vigilant in being careful not to spread it. Most estimates of the incubation period for COVID-19 range from 1-14 days, though the virus most commonly surfaces with symptoms around day five. Most people with COVID-19 have mild illness and are able to recover at home without medical care. He first case of COVID-19 in India, was reported on 30 January 2020. As of May 2021, India has the second-highest number of confirmed cases in the world (after the United States) with 26,752,447 cases of COVID-19. In India, acute effects also need to be studied so that a comprehensive policy can be evolved for the prevention of the health effects of Coronavirus. Therefore, a time-series study was carried out to determine the association of daily confirmed cases with mortality.

Materials and Methods

In this Study, the secondary data available on public domain was used and their website links are given are given in references.

The entire data contains 28 States of India, 7 Union Territories, and one National Territory (Delhi). The zone-wise distribution of States and Union Territories has been listed in Table-1. This data consists of variables– total confirmed cases, daily active cases, active cases, total recoveries, actual active cases, total mortality and mortality per day.

Table 1: Zone wise Distribution of States and Union Territories

Region	States
East	Bihar, Orrisa, Jharkhand, West Bengal
West	Rajasthan, Gujarat, Goa, Maharashtra
North	Himachal Pradesh, Punjab, Uttar Pradesh, Haryana, Uttarakhand
South	Kerala, Karnataka, Tamil Nadu, Andhra Pradesh, Telangana
North East	Meghalaya, Manipur, Nagaland, Sikkim, Mizoram, Arunachal Pradesh, Tripura, Assam
Central	Madhya Pradesh, Chhattisgarh
Union Territories	Chandigarh, Andaman and Nicobar Islands, Ladakh, Lakshadweep, Puducherry, Dadra and Nagra Haveli, Jammu and Kashmir
Delhi	Delhi

In the sequel, we will use the following Notation:
Tcc - Total Confirmed cases
Dcc - Daily Confirmed cases
Ac - Active cases
Tr - Total Recoveries
Acc - Actual Active cases
Total Mortality - Cumulative Deaths
Mortality - Daily Mortality (Number of deaths per day).

Statistical Analysis
The data were examined for quality and consistency by cross checking different websites. Descriptive statistics i.e. mean, minimum, maximum, and range were computed. A generalized additive model (GAM) was fitted in R statistical software with the Quasi-Poisson link function for mortality as the dependent variable and daily confirmed cases as the independent variable with penalized or natural spline with (6,2,2) degrees of freedom. The smoothers were added for day effect, active cases and actual active cases. The day of the week was coded as 1 to 7 (Monday through Sunday). The base model includes smoothers at initial stage:

Log (E(mortality)) = s (day effect, df) + s (active cases, df) + s (actual active cases, df)

The optimal degrees of freedom (df) were determined using approach suggested by R. Aggarwal, S. K. Sharma and K. Jain (2014). After this, the main variables daily confirmed cases (dcc) were added to the model.

The adopted model was
Log (E(mortality)) =daily confirmed cases +s (day effect) + s (active cases) + s (actual active cases)

It has been noticed that the generalized cross-validation (GCV) score did not vary much between the natural spline at (6,2,2) df and penalized spline at (6,2,2) df.

Results
During a 395-day period between April 1,2020 and April 30, 2021 there were 2,08330 deaths registered in India which stands third highest in the world. In this period mortality rate on an average was 533 deaths per day. The summary statistics which includes mean, minimum, maximum and range are shown in Table 2 and monthly averages of all variables are shown in Table 3.

Table 2: Summary Table

Variables	Mean	Minimum	Maximum	Range
tcc	6335703	1635	18762976	18761341
dcc	48254	278	386452	386174
ac	468177	1464	3170228	3168764
tr	5775463	133	15384418	15384285
Total mortality	92066	38	208330	208292
mortality	533	13	3645	3632
acc	560240	1502	3378558	3377056

Table 3: Monthly Average Table

Month	tcc	dcc	ac	tr	Total mortality	Mortality/day	acc
APRIL_20	13250	1010	10521	2345	421	36	10905
MAY_20	93112	4583	53826	36393	2893	132	56719
JUNE_20	337795	12450	141211	185936	10648	391	151859
JULY_20	1021261	34808	351144	644520	25597	609	376741
AUGUST_20	2604826	63948	665008	1889706	50112	927	715120
SEPTEMBER_20	4970459	86817	935915	3953098	81446	1101	1017361
OCTOBER_20	7315171	61657	789945	6413691	111536	779	901481
NOVEMBER_20	8818547	43152	481770	8206972	129806	517	611575
DECEMBER_20	9909046	28104	338040	9427220	143786	385	481826

JANUARY_21	10531460	18585	207666	10171886	151908	206	359573
FEBRUARY_21	10915330	12791	147722	10611914	155694	102	303416
MARCH_21	11496155	39310	283881	11053064	159211	196	443091
APRIL_21	14688501	220570	1677524	12833006	178010	1529	1855495
GRAND AVERAGE	6335703	48254	468177	5775463	92066	533	560240

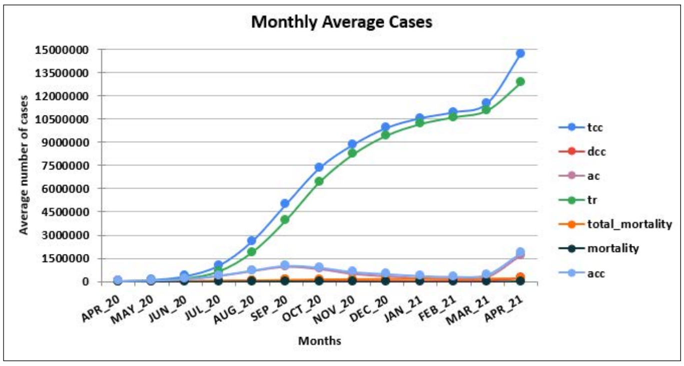

Figure 1: Monthly Average Cases in India

It is evident from the graph (figure 1) that cumulative confirmed cases have shown an increasing trend till the end of January 2021 and showed a slight declining trend till the beginning of March 2021 but in April 2021 peak has reached to the maximum. Total recoveries are also following the similar pattern as total confirmed cases. However, there was large spike in mortality particularly in September 2020 when around 1101 deaths per day were recorded which is much more than the previous five-month's average. With decline of the first wave by September 2020, mortality began to fall again. From March 2021(because of the onset of second wave), this mortality began to rise and in April 2021, astonishing number i.e. almost 1529 deaths per day were reported. The confirmed cases also jumped to around 2.205 lakh per day in April 2021 which were more than what was observed at the beginning of second wave.

Zone-wise Box plots show overall patterns of Coronavirus with respect to total confirmed cases (tcc), total recoveries (tr) and total mortality (tr). It gives a comparison of different zones and provide useful information in terms of median, quartiles, minimum and maximum values of different aspects of Coronavirus.

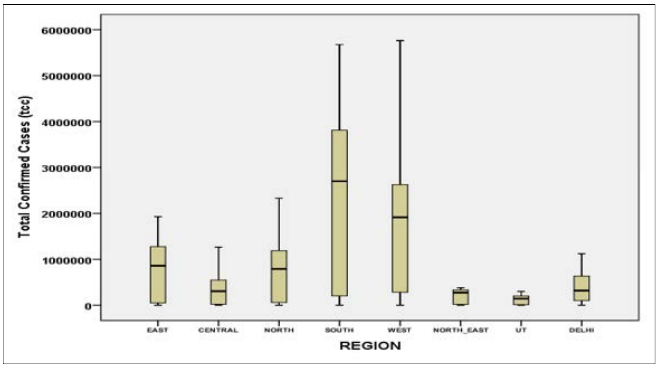

Figure 2: Box Plots for Total Confirmed Cases

The box plot of North-East and Union Territory (UT) is relatively short because in these regions total confirmed cases were less as compared to the other regions. Both of them are highly negatively skewed which signifies that maximum confirmed cases are in the last quarter only. There were more confirmed cases in South and West regions the box plots are relatively high. Moreover, for both regions data is negatively skewed which means more confirmed cases fall in the last 6-months. Also, we can see that Central region is almost symmetric and it is normally distributed. East, North and west are also negatively skewed and Delhi is positively skewed.

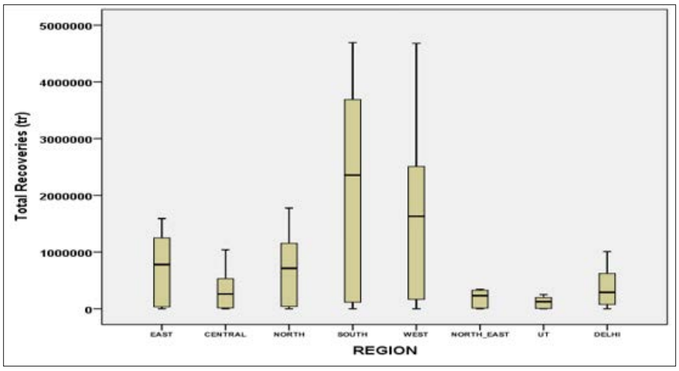

Figure 3: Box Plots for Total Recoveries

In this total recovery box plot, we again see that plot of South and West are tall which means the data is spread out and the box plot of Union Territory (UT) and North-East is relatively small because of the smaller number of cases. Delhi is positively skewed and rest all are negatively skewed except Central region which is almost symmetric.

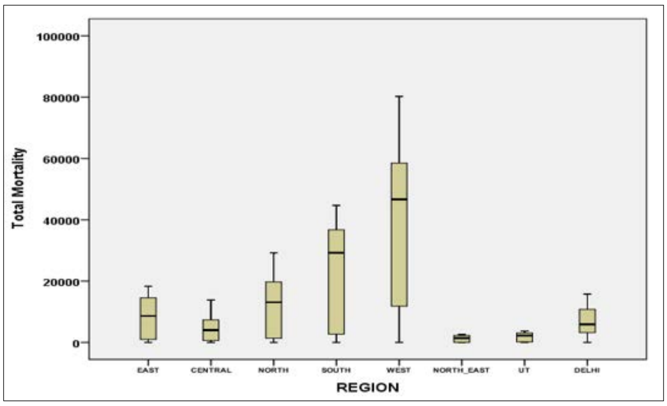

Figure 4: Box Plots for Total Mortality

In mortality box plot the same pattern is followed for South and West, however, mortality is high in West and from this we can conclude more people are dying in west due to Coronavirus. This Region contains Maharashtra state which is the top most highly affected state in India. North East Region has lower mortality compared to all the regions. Delhi is again positively skewed here and rest all are negatively skewed except central region which is almost symmetrical.

Development of GAM Model

Additive model generalizes the linear model by modelling the expected value of Y as

$$E[Y|\, X_1, \dots, X_k] = S_0 + S_1(X_1) + \dots + S_k(X_k) = S_0 + \sum_{i=1}^{k} S_i(X_i).$$

This model is a non-parametric model where the usual regression coefficients β_0,β_1,...,β_k in linear regression model are replaced by smoothers S_1,...,S_k (where S is an unspecified non-parametric function). A Smoother is an arbitrary function that appears as a smooth curve through a scatter plot of points and is estimated in a nonparametric fashion. Sometimes, an additive model may contain a combination of parametric linear functions and non-parametric smooth functions.

GAM Model

If we combine additive models and generalized linear models, we have the notion of Generalized Additive Model (GAM) written as

$$g(E[Y|\, X_1, \dots, X_k]) = S_0 + \sum_{i=1}^{k} S_i(X_i).$$

Generalized Additive Model consists of a random component, an additive component and a link function relating these two components. The response Y, the random component is assumed to possess a density in the exponential family. The quantity

$$\eta = S_0 + \textstyle\sum_{i=1}^{k} S_i(X_i).$$

Note that S_1,...,S_k are smooth functions and they define the additive component. Finally, the relationship between mean μ of the response variable and η is defined through a link function g(μ) = η where $\mu = E[Y|\, X_1, \dots, X_k]$.

Generalized Cross Validation

Cross Validation (CV) works by leaving points (xi, yi) and estimating the smoother at xi based on remaining (n-1) points. The cross-validation sum of squares is

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \{y_i - \hat{f}_\lambda^{-i}(x_i)\}^2$$

where $\hat{f}_\lambda^{-i}(x_i)$ indicates the fit at x_i computed by leaving out the ith data point. There is a simple way to define $\hat{f}_\lambda^{-i}(x_i)$ given only smoother matrix $\boldsymbol{S}_\lambda$. The Generalized Cross Validation (GCV) can be computed from the expression

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - tr(\boldsymbol{S}_\lambda)/n} \right\}^2,$$

where $\hat{f}_\lambda(x_i) = \sum_{j=1, j \neq i}^{n} S_{ij}(\lambda) y_j$ and $\mathbf{S} = \{S_{ij}\}$ is an n x n matrix called smoother matrix.

Fitting of the BASE model is carried out with Penalized Splines (PS) and Natural Splines (NS) at (6,2,2) df by considering day effect, active cases (ac) and actual active cases (acc) as covariates. After that, the main variable, daily confirmed cases are added to the model. In the sequel, we work out the PACF and Residual plots for Lag0 along with predicted plot of mortality in R for both the models. Although statistical models have been developed for each Zone, but for the want of space the predicted model were presented for the entire data, however, Zone-wise brief results are presented in table 8.

Smooth Plots

In this section smoothing plots for total confirmed cases, total recoveries, actual active cases, daily confirmed cases, total mortality, mortality/day & Active cases have been worked out. The smoothers generally act as covariates and after controlling their effect, the actual relationship between dependent and independent variables can be made. They also help us to identify outliers present in the data. Generally, box-plot is used to identify outliers, however, in case of higher values of skewness and kurtosis smooth plot considered to be a better method to identify outliers. Since our data was skewed for most of the zones (see figures 2-4), hence application of smoothing plot is required in analysis of data. In these plots red bubbles plotted are the data values and blue line shows how the data fits the model. The smoother plots work well with 20 degrees of freedom (df), which are optimal degrees of freedom suggested by the data using R-program.

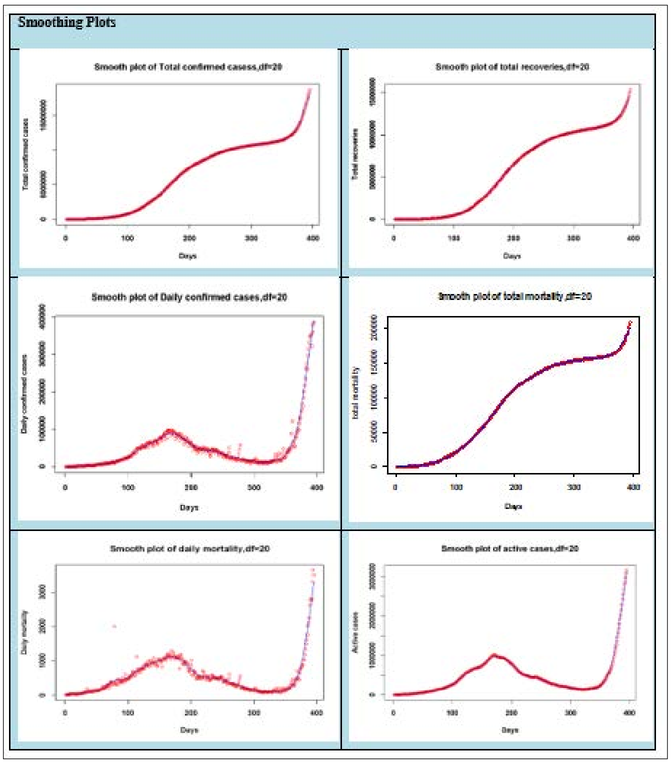

Figure 5: Smoothing plots of total confirmed cases, total recoveries, daily confirmed cases, total mortality, mortality/day & Active cases

There are couples of reasons behind the outliers present in smoothing plots for daily confirmed cases and daily mortality, one of them being that government often revise data or report a single-day large increase in cases or deaths from unspecified days without, historical revisions, which causes an irregular pattern in daily reported cases like, Maharashtra reported 1409 deaths on 17 June, 2020, in which nearly 95% was from the backlog. Actual mortality rate, on that day in Maharashtra was 81; however, with the state health department adding 1328 backlog fatalities, Maharashtra's death toll surged to 5537. The same thing is happening with other states as well. Another reason is the onset of the second wave of Covid-19 in India. During this period, infections and mortality began to rise around mid-March and increased rapidly, which leads to spike in mortality and confirmed cases. However, they are justifiable outliers and we cannot ignore them or remove them from our data. These outliers have been retained while developing the predictive model.

From the above figure, corresponding to 78th day and 115th day it is clearly visible there was sudden spike of 1444 deaths and 1129 deaths respectively in daily mortality smooth plot, and it is because of backlog of corona cases that have been reported and added on these days. The analysis has been performed for the entire data India as well as Zone-wise Firstly, Natural Spline (NS) and Penalized Spline (PS) at (6,2,2) models were fitted to the entire data and the results are presented below. Secondly, zone-wise analysis was also performed and brief results with important information is mentioned in table 8. It has been observed that models PS (6,2,2) and NS (6,2,2) fits well to the completer data as well zone-wise data.

Model fitting for Entire Data
Model: NS (6,2,2)
Family: quasipoisson; Link function: log
mortality ~ dcc + s (day, k = 6 + 1, fx = T, bs = “cr”) + s (acc, k = 1 + 1, fx = T, bs = “cr”) + s (ac, k = 1 + 1, fx = T, bs = “cr”)

Table 3: Significance of Main Variable

Significance of Main Variable				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	5.488178453	0.068468811	80.156	< 0.0002 ***
dcc	0.000006753	0.000001353	4.992	0.00091 ***

On the basis of values in Table 3 it is evident that effect of daily confirmed cases on mortality is highly significant (p<0.0005)

Table 4: Significance of Smooth Terms

Approximate significance of smooth terms:				
	Edf	**Ref.df**	**F**	**p-value**
s(day)	6	6	79.16	< 0.0002 ***
s(acc)	2	2	21.14	0.000195 ***
s(ac)	2	2	20.24	0.000439 ***
R-sq.(adj) = 0.956 Deviance explained = 96.2% GCV = 16.603 Scale est. = 24.162 n = 395				

It is concluded on the basis of values in Table 4 that the smoothers for day effect (p< 0.0005), acc(p<0.0005) and ac(p<0.0005) are highly significant. The deviance explained by the model is 96.2% with GCV score of 16.603.

Model: PS (6,2,2)
Family: quasipoisson; Link function: log

mortality ~ dcc + s(day, k = 6 + 1, fx = F, bs = “cr”) + s(acc, k = 1 + 1, fx = F, bs = “cr”) + s(ac, k = 1 + 1, fx = F, bs = “cr”)

Table 5: Significance of Main Variable

Significance of main variable				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	5.491332976	0.068459701	80.213	< 0.0002 ***
dcc	0.000006702	0.000001353	4.953	0.0000011 ***

The value of daily confirmed cases (dcc) in Table 5 is highly significant (p<0.0001). Hence dcc has significant impact on mortality.

Table 6: Significance of Smooth Terms

Approximate significance of smooth term				
	Edf	Ref.df	F	p-value
s(day)	5.978	6	78.48	<0.0002 ***
s(acc)	1.997	2	20.63	<0.0002 ***
s(ac)	1.998	2	19.72	<0.0002 ***
R-sq.(adj) = 0.956 Deviance explained = 96.2% GCV = 16.602 Scale est. = 24.226 n = 395				

It is concluded on the basis of values in Table 6 that the smoothers for day effect (p< 0.0002), acc(p<0.00002) and ac(p<0.0002) are highly significant. The deviation explained by the model is 96.2% with GCV score 16.602. Although, both NS and PS models have almost same GCV *score, however, GCV Score (16.602) of PS (6,2,2) is slightly less than that of GCV Score (16.603) of NS (6,2,2), therefore PS (6,2,2) model better fits our data*. The Relative Risk along with 95% confidence intervals has been presented in Table 7.

Table 7: Relative risk along with 95% confidence intervals

dcc(daily confirmed cases)			**95% confidence interval (RR**	
Beta-coefficient	**Relative risk (RR)**	**Standard error**	**Lower**	**Upper**
0.000006702	1.006702	0.0001353	0.77222	1.312443

The relative risk has been computed for increase in every 1000 number of daily confirmed cases. For increase in 1000 number of daily confirmed cases, the expected number of deaths amounts to 1.006702 (approximately).

Residuals and Pacf Plots for PS (6,2,2) Model along with Predicted Plot for Mortality

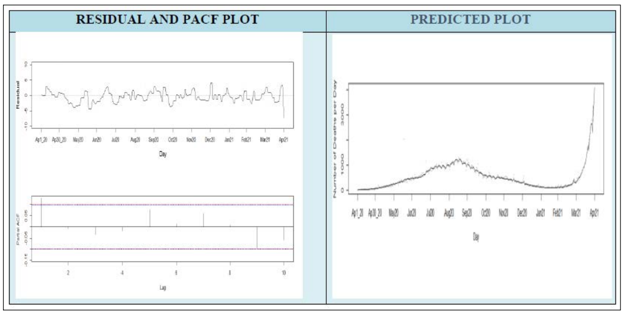

Figure 6: Residuals, PACF and Predicted plots for PS (6, 2, 2) df

There is a significant effect of daily confirmed cases on mortality after smoothing the effects of day, active cases and actual active cases. The points in the residual plots are symmetrically distributed over and above the baseline zero, leading to justification of the model. The predicted plot of mortality at lag0 can be used to determine episodes of mortality during a particular day. A prediction for episodes of mortality seems to be reasonably well and it can be depicted from the predicted plot. The effect of daily confirmed cases (on particular day) on mortality can be seen not only on that particular day but also effects of virus can last for weeks. The PACF plots shows that mortality is high (i.e. significant) during the first day at lag0 and after that it is comes out to be significant at 5th day which means the effect of daily confirmed cases (on first day) on mortality have severe effect for the first five days. From figure 6, it's clearly visible that predicted plot closely fits the actual data. Hence, the PS (6, 2, 2) model fits well to our data.

Zone Wise Brief Results

Table 8: Region-wise Modeling Results

1.	**East Region**			
Significance of Main Variable: Model: PS(6,2,2)				
	Estimate	**Std.Error**	**t-value**	**Pr(>|t|)**
(Intercept)	2.97863250	0.09766155	30.500	< 0.0002 ***
dcc	0.00005916	0.00001867	3.169	0.00165 **
Approximate significance of smooth terms				
	E .df	Ref.df	F	p-value
s(day)	5.783	5.976	61.963	<0.0002 ***
s(acc)	1.925	1.990	2.986	0.0431 *
s(ac)	1.997	2.000	3.439	0.0315 *
R-sq.(adj) = 0.954 Deviance explained = 94.5% **GCV = 2.5707 Scale est. = 5.501 n = 395**				

2.	**West Region**			
Significance of Main Variable: Model: PS(6,2,2)				
	Estimate	**Std.Error**	**t-value**	**Pr(>|t|)**
(Intercept)	4.603e+00	8.032e-02	5 7.307	< 2e-16 ***
dcc	2.270e-05	4.911e-06	4.622	5.21e-06 ***
Approximate significance of smooth terms				
	E .df	Ref.df	F	p-value
s(day)	5.852	5.992	33.623	<2e-16 ***
s(acc)	2.000	2.000	5.532	0.0044 **
s(ac)	1.001	1.001	9.585	0.0021 **
R-sq.(adj) = 0.846 Deviance explained = 88.9% **GCV = 18.374 Scale est. = 29.961 n = 395**				

3.	**North Region**			
Significance of Main Variable: Model: PS(6,2,2)				
(Intercept)	**Estimate**	**Std.Error**	**t-value**	**Pr(>|t|)**
dcc	3.77E+00	4.17E-02	90.43	< 2e-16 ***
	-1.64E-05	6.09E-06	-2.7	0.00724 **
Approximate significance of smooth terms				
	E .df	Ref.df	F	p-value
s(day)	5.981	6	96.27	<2e-16 ***
s(acc)	1.999	2	51.13	<2e-16 ***
s(ac)	2	2	46.5	<2e-16 ***
R-sq.(adj) = 0.971 Deviance explained = 96.7% GCV = 2.6684 Scale est. = 2.5703 n = 395				

4.	**South Region**			
Significance of Main Variable: Model: NS(6,2,2)				
	Estimate	**Std.Error**	**t-value**	**Pr(>|t|)**
(Intercept)	3.816705794	0.05797091	65.838	<0.0002***
dcc	0.000018472	0.00000361	5.109	0.000513 ***
Approximate significance of smooth terms				
	E .df	Ref.df	F	p-value
s(day)	6	6	98.39	<0.0002***
s(acc)	2	2	18.68	0.000181***
s(ac)	2	2	19.23	0.000110 ***
R-sq.(adj) = 0.942 Deviance explained = 96.6% **GCV = 3.8865 Scale est. = 4.4907 n = 395**				

5.	**Central Region**			
Significance of Main Variable: Model: PS(6,2,2)				
	Estimate	**Std.Error**	**t-value**	**Pr(>|t|)**
(Intercept)	2.81870550	0.04963182	56.792	<2e-16 ***
dcc	0.00004644	0.00001349	3.441	0.000642 ***
Approximate significance of smooth terms				
	E .df	Ref.df	F	p-value
s(day)	6	6	43.012	<2e-16 ***
s(acc)	2	2	4.319	0.01397 *
s(ac)	2	2	4.696	0.00966 **
R-sq.(adj) = 0.973 Deviance explained = 95.2% **GCV = 2.4917 Scale est. = 2.5906 n = 395**				

6.	**North East Region**			
Significance of Main Variable: Model: NS(6,2,2)				
	Estimate	**Std.Error**	**t-value**	**Pr(>|t|)**
(Intercept)	1.71516334	0.05027335	34.117	<0.0002 ***
dcc	0.00005752	0.00003388	1.698	0.0906
Approximate significance of smooth terms				
	E .df	Ref.df	F	p-value
s(day)	6	6	18.988	<0.0002 ***
s(acc)	2	2	4.522	0.0116 *
s(ac)	2	2	4.395	0.0132 *
R-sq.(adj) = 0.823Deviance explained = 84.8% **GCV = 1.098 Scale est. = 1.0732 n = 304**				

7. Union Territory (UT)				
Significance of Main Variable: Model: PS(6,2,2)				
	Estimate	**Std.Error**	**t-value**	**Pr(>|t|)**
(Intercept)	1.71174844	0.05184730	33.015	<0.0002 ***
dcc	0.00007461	0.00006027	1.238	0.216
Approximate significance of smooth terms				
	E .df	Ref.df	F	p-value
s(day)	5.984	5.997	56.423	<0.0002 ***
s(acc)	1.569	1.700	0.453	0.696
s(ac)	1.729	1.812	0.405	0.724
R-sq.(adj) = 0.892 Deviance explained = 90.1% **GCV = 0.94636 Scale est. = 0.87992 n = 395**				

8. Delhi				
Significance of Main Variable: Model: PS(6,2,2)				
	Estimate	**Std.Error**	**t-value**	**Pr(>|t|)**
(Intercept)	2.955e+00	6.328e-02	46.693	< 2e-16 ***
dcc	-1.058e-05	1.606e-05	-0.659	0.51
Approximate significance of smooth terms				
	E .df	Ref.df	F	p-value
s(day)	5.944	5.998	44.834	<2e-16 ***
s(acc)	1.937	1.996	24.464	<2e-16 ***
s(ac)	1.000	1.000	2.364	0.125
R-sq.(adj) = 0.0.883 Deviance explained = 91.9% **GCV = 4.8619 Scale est. = 6.3162 n = 395**				

The region-wise results are presented only for best fitted model with appropriate parameters and degrees of freedom. Out of the 8 regions NS (6,6,2) model fits well to only 2 regions (South and North East) and for all other regions model PS (6,2,2) works well. It may be noted that for North-East region, data was available for n=304 days, while complete data was available for other regions. The main variable daily confirmed cases (dcc) were significant in 5 regions but it was non-significant in three Regions-North East, Union Territory (UT) and Delhi. One on the possible reasons for non-significant could be underreporting of data in these regions. Soothers were significant in almost all regions except Union Territory (UT) and Delhi.

Discussion

India is among the 5 topmost corona-affected country of the world with 2,08,330 total deaths with an average of 48254 confirmed cases and 533 deaths per day. After looking at the graph (trend analysis) of India, it's clearly visible that number of cases in April 2020 to June 2020 is low, from June 2020 onwards; it started increasing day by day. There was a large spike in deaths particularly in month of September 2020 (first wave of corona) when approximately 1101 deaths per day were recorded. With decline of first wave, mortality began to fall. From March 2021(because of onset of second wave), this mortality again began to rise and in April 2021 where an astonishing number i.e. almost 1529 deaths per day were recorded.

For the entire data, Generalized Additive Model has been applied to see the impact of daily confirmed cases on Mortality. Fitting of the BASE model is carried out with Penalized Splines (PS) and Natural Splines (NS) at (6,2,2) df by considering day effect, active cases and actual active cases as covariates. After that, the main variable, daily confirmed cases are added to the model. We work out the smoothing plots, the parameter estimates for all models along with PACF, Residual plots and Predictive plots of mortality for Lag0 in R for both the models. The Relative Risk (RR) has also been computed for increase in every 1000 number of daily confirmed cases.

For NS (6,2,2) and PS (6,2,2) model, the effect of daily confirmed cases on mortality by smoothing the day effect, active cases and actual active cases are found to be statistically significant. Analysing the Generalized Cross-validation, we noticed GCV Score of PS (6,2,2) is less than that of GCV Score of NS (6,2,2). Hence, PS (6,2,2) fits well to the data. With increase in every 1000 number of daily confirmed cases, the expected number of deaths amounts to 1.007(approximately).

It is evident from PACF and Residual plots that there is significant effect of daily confirmed cases on mortality after smoothing the effects of day, active cases and actual active cases. Residual plots show difference between actual and predicted value is evenly distributed around zero. From the predicted plots it is visible that actual data closely fits to the model. Therefore, the PS (6,2,2) model fits our data well. The Region-wise results are also presented in Table 8 [5-12].

References

1. Nyberg T, Twohig KA, Harris RJ, Shaun R Seaman, Joe Flannagan, et al. (2021) Risk of hospital admission for patients with SARS-CoV-2 variant B.1.1.7: cohort analysis. BMJ 373: n1412.
2. Bager P, Wohlfahrt J, Fonager J, Morten Rasmussen, Mads Albertsen, et al. (2021) Risk of hospitalisation associated with infection with SARS-CoV-2 lineage B.1.1.7 in Denmark: an observational cohort study. Lancet Infect Dis 21: 1507-1517.
3. Novel Coronavirus MOHFW Home (2020) Ministry of Health and Family Welfare. GOI https://mohfw.gov.in/.
4. Chin AWH, Chu JTS, Perera MRA, Hui KPY, Yen HL, et al. (2020) Stability of SARS-CoV-2 in different environmental conditions. The Lancet Microbe 1: e10.
5. R Aggarwal, SK Sharma, K Jain (2014) Optimal Choice of Splines and Knots in TPSPLINE and TRANSREG Procedures. IOSR Journal of Mathematics 10: 42-52.
6. De Boor C (1978) A Practical Guide to Splines. Mathematics of Computation 27.
7. Eilers PHC, Marx BD (1996) Flexible Smoothing with B-splines and Penalties. Statistical Science 11: 89-121.
8. Green PJ, BW Silverman (1994) Nonparametric Regression and Generalized Linear Models. London: Chapman & Hall 184.
9. Gu C (2002) Smoothing Spline ANOVA Models. New York: Springer Verlag.
10. Hastie T, Tibshirani R (1986) Generalized Additive Models. Statistical Science 3: 297-318.
11. Hasties TJ, Tibshirani RJ (1990) Generalized Additive Models. New York: Chapman and Hall, New York 352.
12. Trevor Hastie, Robert Tibshirani (1987) Non-Parametric Logistic and Proportional Odds Regression. Applied Statistics 36: 260-276.