COVID -19 – TIME SERIES (CLASSICAL DECOMPOSITION METHOD) MODELLING FOR CONFIRMED, DEATH AND RECOVERY OF CASES FOR GLOBAL, INDIA AND OTHER GEOGRAPHIES

Vikram R Sreedhar
Principal Consultant & Lead
Olives and Berries Consulting LLP, Bangalore, India
www.onbconsulting.com

24th March 2020

Contents

2. Background on Time Series

A Time-Series is a *time-stamped* data set in which each data point corresponds to a set of observations made at a relevant time instance. Unlike the regression where the attempt was to express a response variable in terms of some of the other attributes (explanatory variables), time-series analysis assumes the system to be a black-box — we just attempt to predict what is coming based on the past behavior patterns.

For instance, treating stock prices as a timeseries would mean creating a model that could predict say the stock prices over the next one month if we knew the prices over the past year. We don't really care what factors affect stock prices, macroeconomic policies, oil prices, Brexit, etc., we just want to predict, given we know how it has changed in the past.

A time-series view is often taken when we suspect a 'white-box' model based on the underlying factors will be too complex, we do not know the underlying factors that influence the response variable, or may be just that we are not as interested in understanding why something behaves the way it does, as we are in just predicting what will happen next and move on.

Examples of time-series arise in diverse domains and are almost ubiquitous. An illustrative list of time-series data across a variety of domains is given below.

- Daily Stock market figures
- Sales / Revenue / Profitability figures of companies by year
- Demographic / Development data (population, birthrate, infant mortality figures, literacy,
- per-capita income, school enrollment figures) by year
- Blood Pressure and other body vitals by time units (in hours, minutes,) appropriate to the severity of the patient
- Ecology / Geology data — progression of the intensity of the tremors of an earthquake with time, pollution figures by year, annual average temperature, thickness of the glacial ice-sheet etc. by year indicating global warming, annual rainfall figures, etc.,
- Epidemiology data—spread (number of people affected) of an epidemic with time
- Speech data (signal parameters evolving over time)—these need to be modeled for speech recognition and of course to build applications such as the popular Siri on iPhones [1]

Most time series can be analyzed by examining some typical characteristics exhibited by their distinctive data sets:

1. Trends: Time varying trends (eg., increasing / decreasing linearly). A typical stock market bull run is an example of an upward trend — the index moving up steadily with time. A trend is typically seen over a (relatively) long stretch of time.

2. Seasonality: Exhibiting a regular repeating pattern of some sort. Sales volume in a typical departmental store shows regular seasonal patterns, increased volumes during weekends, increased volumes around festivals, etc. Seasonal patterns happen with a periodicity that is known and predictable.

3. Long-Term Cyclicity: Long-term cyclic behavior. Economic inflation-recession cycles are a typical example.

4. Autocorrelation: Dependence of the next value on the past. The phenomenon of some video on YouTube for instance going 'viral' is a manifestation of autocorrelation. If the number of views over the last 5 days is say 10, 11, 13, 15, 15 for one video and is 10, 100, 500, 5000, 50000 for another video. That the first one hasn't 'picked up steam' in the last 5 days is a pretty strong indication that it is not likely to be too far from say 15 even on the 10th day. The second video however is the one going 'viral'. The volumes over the last 5 days is a strong indicator of what we can expect on the 6th day. In general degree-k autocorrelation is a where the following value is highly correlated with the history over the last k time instances. We are of course assuming discrete time here — that observations are made once in (or in multiples of) a fixed unit of time.

5. Stochasticity: Almost all real time-series data is noisy due to the inherent unpredictability
in the behaviour of the data sources and instruments used to collect the data. Real timeseries
data is therefore a combination of some expected behaviour (the deterministic, predictable ideal behaviour) and a random stochastic noisy component. [2]

3. Standard Methodology followed in Timeseries Modelling

1. Visualization of data as a line or scatter plot. From visualization we tend to observe the trends, patterns and outliers in the underlying data.

2. *Stationarizing* the data. A stationary time series can be defined as whose statistical metrics viz., mean, covariances between observations at varied time intervals are all constant over a period. It refers to *time invariant* relationship it exhibits between observations at varied time periods. Characteristics like trend, pattern, cyclicity, seasonality are all non-stationary. Therefore, Stationarizing the data is de-seasoning, de-trending the data through series of mathematical transformations. The residual series or the residue is the one which is obtained after removing trend, patterns and seasonality in the data. It can be presented as

$$(Y0, Y1 \dots Yn) \ \ where\ (Yt = S(Xt))$$

3. Test for the stationarity of the residual series. If not, then iterate the step 2, till residual series is stationary

4. Develop a final model by applying the inverse of mathematical transformations S^{-1} on the residual model.

5. Check for the stochasticity of the model developed. Once modeled the residue post predictions should be white noise which is a pure noise, it has no coordinated behaviour that can be modeled. This is known as the stochastic part of the time series data.

4. Data Selection and Exploratory Data Analysis

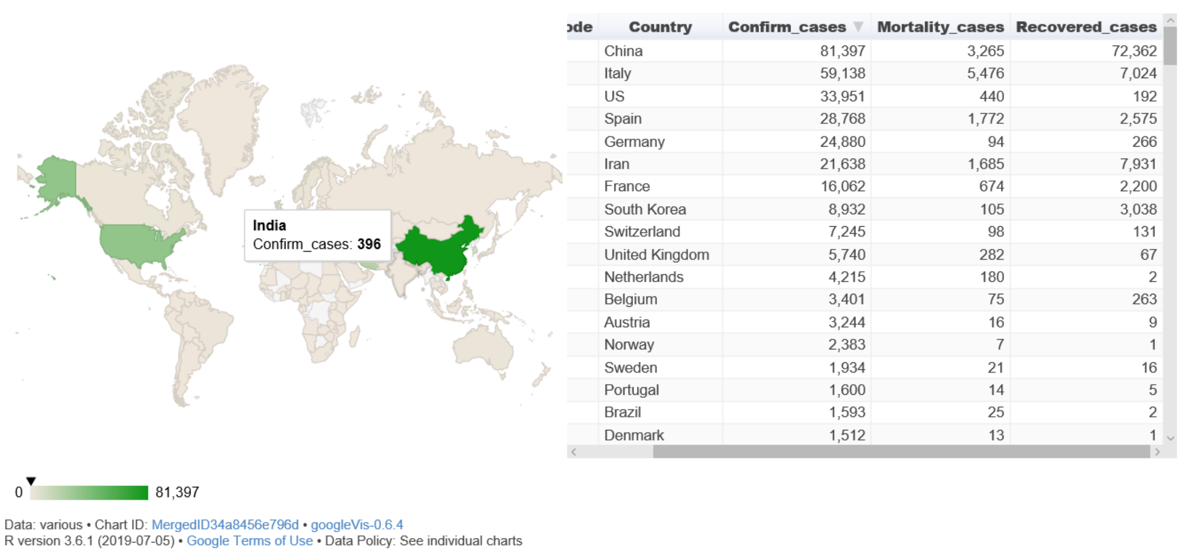
Data is from repository of 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL). Source of data being World Health Organization and availability of the data is on GitHub. [3]

Data has situation reports of Inception till date of confirmed cases, mortality and recovered cases of persons infected with COVID 19 virus with details of Country, Province, date updated and Latitude and Longitude of the country from where the cases are reported. World Health Organization defines the confirmed case as "a person with laboratory confirmation of 2019-nCoV infection, irrespective of clinical signs and symptoms.

Data is updated daily at midnight with daily reports country wise on confirmed, mortality and recovered cases.

Exploratory data analysis – Bivariate Analysis

Global Scenario of confirmed, mortality and recovered cases as of 22nd March 2020 of COVID 19 Pandemic – Chart 1



ode	Country	Confirm_cases	Mortality_cases	Recovered_cases
	China	81,397	3,265	72,362
	Italy	59,138	5,476	7,024
	US	33,951	440	192
	Spain	28,768	1,772	2,575
	Germany	24,880	94	266
	Iran	21,638	1,685	7,931
	France	16,062	674	2,200
	South Korea	8,932	105	3,038
	Switzerland	7,245	98	131
	United Kingdom	5,740	282	67
	Netherlands	4,215	180	2
	Belgium	3,401	75	263
	Austria	3,244	16	9
	Norway	2,383	7	1
	Sweden	1,934	21	16
	Portugal	1,600	14	5
	Brazil	1,593	25	2
	Denmark	1,512	13	1

Tabular Data of total of confirmed, mortality and recovered cases country wise as of 22nd March 2020

Table 1

country_code	Country	Confirm_cases	Mortality_cases	Recovered_cases
CN	China	81,397	3,265	72,362
IT	Italy	59,138	5,476	7,024
US	US	33,951	440	192
ES	Spain	28,768	1,772	2,575
DE	Germany	24,880	94	266
IR	Iran	21,638	1,685	7,931
FR	France	16,062	674	2,200
KR	South Korea	8,932	105	3,038
CH	Switzerland	7,245	98	131
GB	United Kingdom	5,740	282	67
NL	Netherlands	4,215	180	2
BE	Belgium	3,401	75	263
AT	Austria	3,244	16	9
NO	Norway	2,383	7	1
SE	Sweden	1,934	21	16
PT	Portugal	1,600	14	5
BR	Brazil	1,593	25	2
DK	Denmark	1,512	13	1
CA	Canada	1,494	21	14
AU	Australia	1,318	7	88
MY	Malaysia	1,306	10	139
TR	Turkey	1,236	30	0
CZ	Czech Republic	1,120	1	6
JP	Japan	1,086	40	235
IL	Israel	1,072	1	37
IE	Ireland	906	4	5
LU	Luxembourg	798	8	6
EC	Ecuador	789	14	3
PK	Pakistan	776	5	5
Others	Others	757	7	325
PL	Poland	634	7	1
CL	Chile	632	1	8
FI	Finland	626	1	10
GR	Greece	624	15	19
TH	Thailand	599	1	44
IS	Iceland	568	1	36
ID	Indonesia	514	48	29
SA	Saudi Arabia	511	0	16
QA	Qatar	494	0	33
GB	UK	456	8	18
SG	Singapore	455	2	144
RO	Romania	433	3	64
SI	Slovenia	414	2	0
IN	India	396	7	27
PH	Philippines	380	25	17
RU	Russia	367	0	16
PE	Peru	363	5	1
BH	Bahrain	332	2	149
EG	Egypt	327	14	56
EE	Estonia	326	0	2

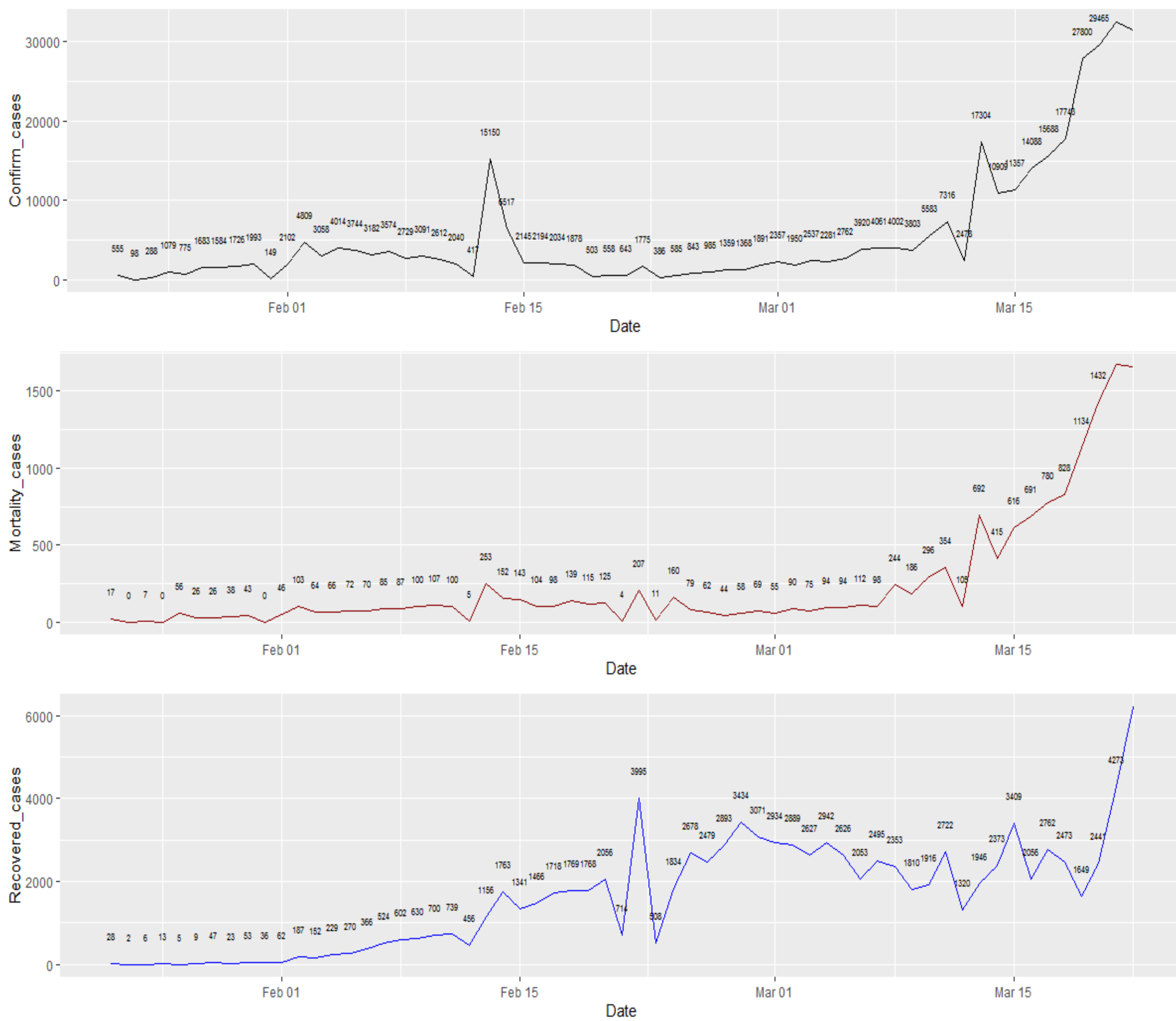
ZA	South Africa	274	0	0
HR	Croatia	254	1	5
MX	Mexico	251	2	4
LB	Lebanon	248	4	8
PA	Panama	245	3	0
IQ	Iraq	233	20	57
CO	Colombia	231	2	3
AR	Argentina	225	4	3
RS	Serbia	222	2	1
DO	Dominican Republic	202	3	0
DZ	Algeria	201	17	65
AM	Armenia	194	0	2
KW	Kuwait	188	0	27
BG	Bulgaria	187	3	3
SK	Slovakia	185	1	7
SM	San Marino	160	20	4
AE	United Arab Emirates	153	2	38
LV	Latvia	139	0	1
CR	Costa Rica	137	2	2
UY	Uruguay	135	0	0
HU	Hungary	131	6	16
LT	Lithuania	131	1	1
BA	Bosnia and Herzegovina	126	1	2
MA	Morocco	115	4	3
MK	North Macedonia	114	1	1
AD	Andorra	113	1	1
VN	Vietnam	113	0	17
JO	Jordan	112	0	1
TW	Taiwan*	102	1	8
CY	Cyprus	95	1	3
MD	Moldova	94	1	1
MT	Malta	90	0	2
AL	Albania	89	2	2
BN	Brunei	88	0	2
KH	Cambodia	84	0	1
LK	Sri Lanka	82	0	3
BY	Belarus	76	0	15
BF	Burkina Faso	75	4	5
TN	Tunisia	75	3	1
UA	Ukraine	73	3	1
VE	Venezuela	70	0	15
SN	Senegal	67	0	5
TW	Taiwan	67	1	20
NZ	New Zealand	66	0	0
AZ	Azerbaijan	65	1	10
KZ	Kazakhstan	60	0	0
GP	Guadeloupe	56	0	0
OM	Oman	55	0	17
GE	Georgia	54	0	3
TT	Trinidad and Tobago	50	0	1

RE	Reunion	47	0	0
UZ	Uzbekistan	43	0	0
AF	Afghanistan	40	1	1
CM	Cameroon	40	0	0
LI	Liechtenstein	37	0	0
MQ	Martinique	37	1	0
CU	Cuba	35	1	0
TZ	Tanzania	35	0	0
CD	Congo	33	1	0
NG	Nigeria	30	0	2
BD	Bangladesh	27	2	3
HN	Honduras	26	0	0
BO	Bolivia	24	0	0
GH	Ghana	24	1	0
MC	Monaco	23	0	1
PS	Palestine	22	0	0
PY	Paraguay	22	1	0
ME	Montenegro	21	0	0
GB	Republic of Ireland	21	0	0
GT	Guatemala	19	1	0
RW	Rwanda	19	0	0
GF	French Guiana	18	0	6
MU	Mauritius	18	1	0
JM	Jamaica	16	1	2
TG	Togo	16	0	1
KE	Kenya	15	0	0
BB	Barbados	14	0	0
CI	Cote d'Ivoire	14	0	1
KG	Kyrgyzstan	14	0	0
MV	Maldives	13	0	0
ET	Ethiopia	11	0	4
MO	Macau	11	0	10
YT	Mayotte	10	0	0
MN	Mongolia	10	0	0
GY	Guyana	7	1	0
SC	Seychelles	7	0	0
SO	Somalia	7	0	0
GQ	Equatorial Guinea	6	0	0
GA	Gabon	5	1	0
SR	Suriname	5	0	0
AW	Aruba	4	0	0
SZ	Eswatini	4	0	0
BS	The Bahamas	4	0	0
CF	Central African Republic	3	0	0
SV	El Salvador	3	0	0
MG	Madagascar	3	0	0
NA	Namibia	3	0	0
MD	Republic of Moldova	3	0	0
ZM	Zambia	3	0	0
ZW	Zimbabwe	3	0	0

AO	Angola	2	0	0
BJ	Benin	2	0	0
BT	Bhutan	2	0	0
FO	Faroe Islands	2	0	0
FJ	Fiji	2	0	0
GN	Guinea	2	0	0
HT	Haiti	2	0	0
XK	Kosovo	2	0	0
MR	Mauritania	2	0	0
NP	Nepal	2	0	1
NI	Nicaragua	2	0	0
NE	Niger	2	0	0
ZA	Saint Lucia	2	0	0
SX	Sint Maarten	2	0	0
SU	Sudan	2	1	0
AG	Antigua and Barbuda	1	0	0
KY	Cayman Islands	1	1	0
TD	Chad	1	0	0
GB	Channel Islands	1	0	0
CW	Curacao	1	0	0
DJ	Djibouti	1	0	0
ER	Eritrea	1	0	0
GD	Grenada	1	0	0
Others	Holy See	1	0	0
MZ	Mozambique	1	0	0
GB	North Ireland	1	0	0
PG	Papua New Guinea	1	0	0
FR	Saint Barthelemy	1	0	0
VC	Saint Vincent and the Grenadines	1	0	0
GM	The Gambia	1	0	0
TL	Timor-Leste	1	0	0
UG	Uganda	1	0	0
VA	Vatican City	1	0	0
GL	Greenland	0	0	0
GU	Guam	0	1	0
GG	Guernsey	0	0	0
JE	Jersey	0	0	0
IL	occupied Palestinian territory	0	0	0
CV	Praia,Praia,Cape Verde	0	0	0
CO	Puerto Rico	0	1	0
CG	Republic of the Congo	0	0	0

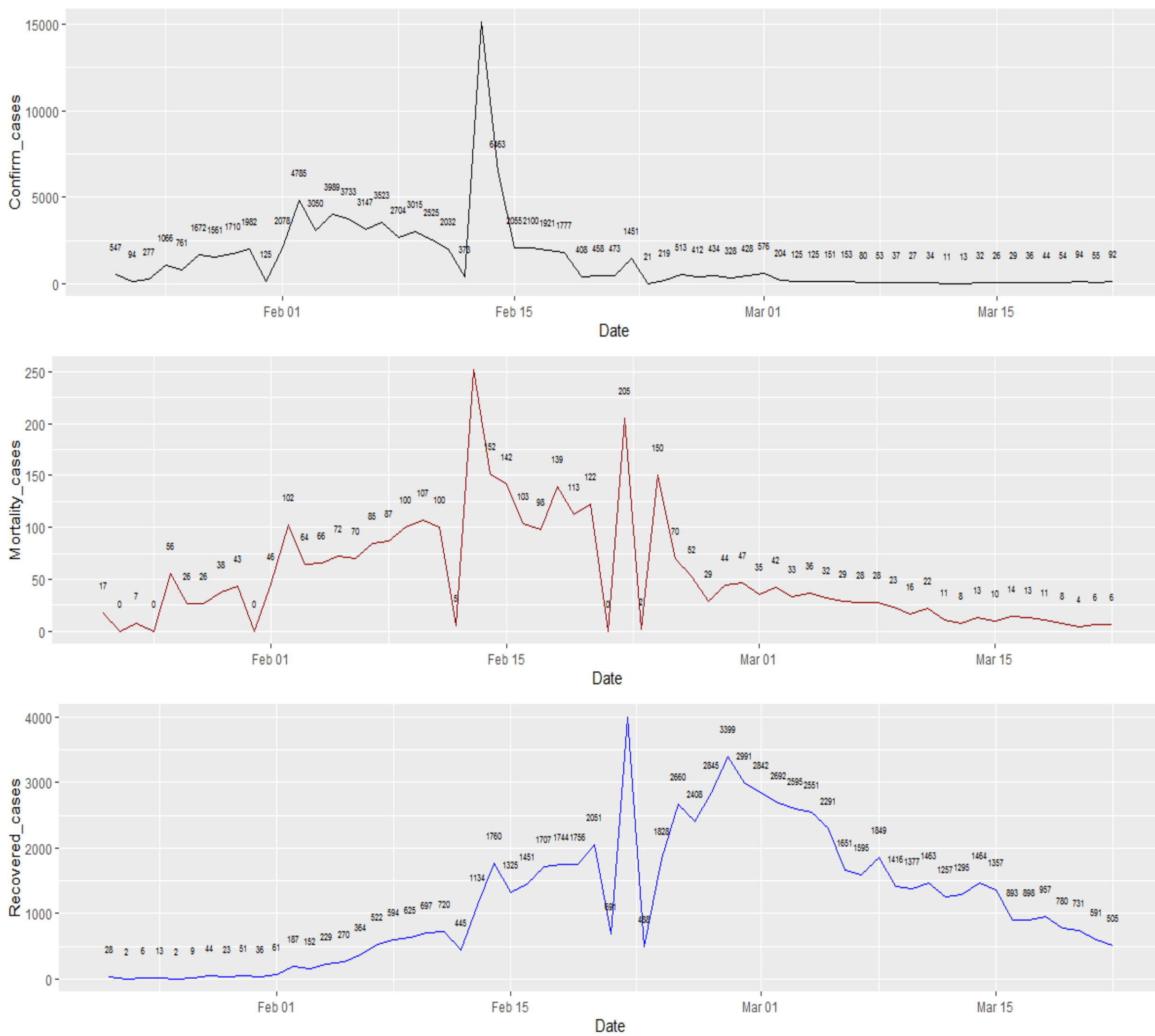
Data: df2 • Chart ID: TableID34a82da95717 • googleVis-0.6.4
R version 3.6.1 (2019-07-05) • Google Terms of Use • Documentation and Data Policy

Day wise – Confirmed Cases, Mortality and recovered cases – Worldwide -upto 22nd March

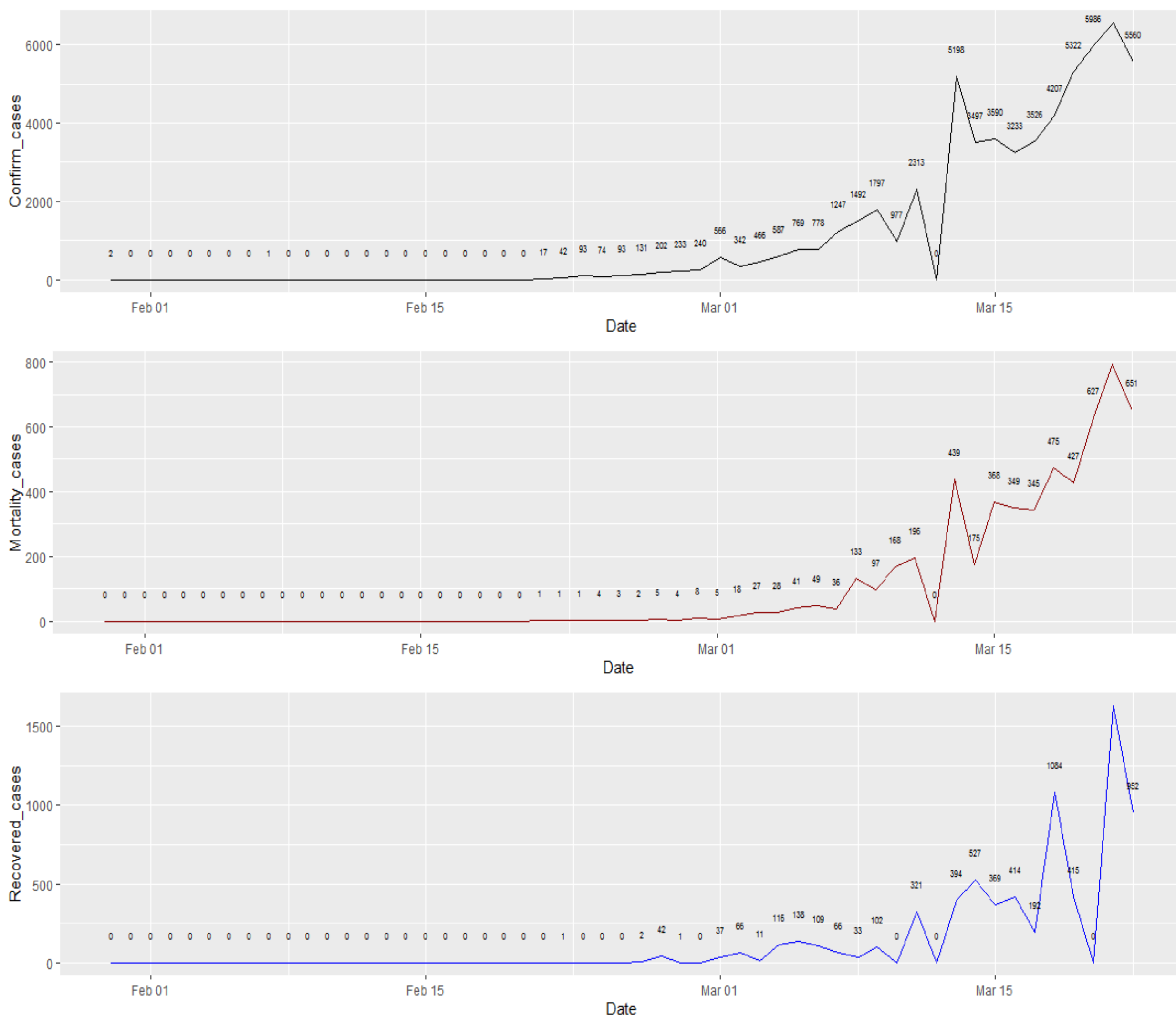


Day wise Confirmed, Mortality and recovered cases – India – Upto 22nd March

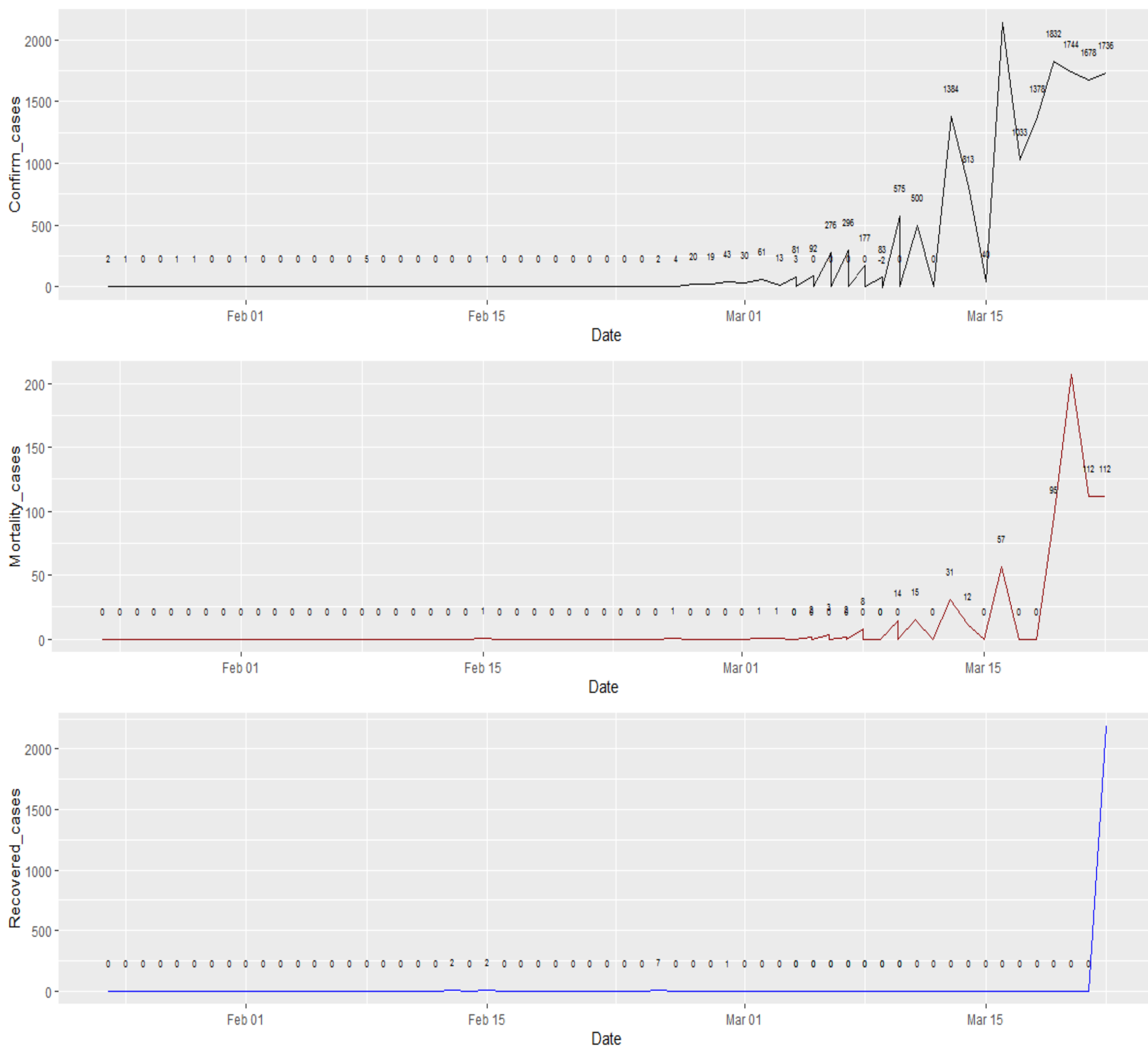
Day wise Confirmed, Mortality and recovered cases – China – Upto 22nd March



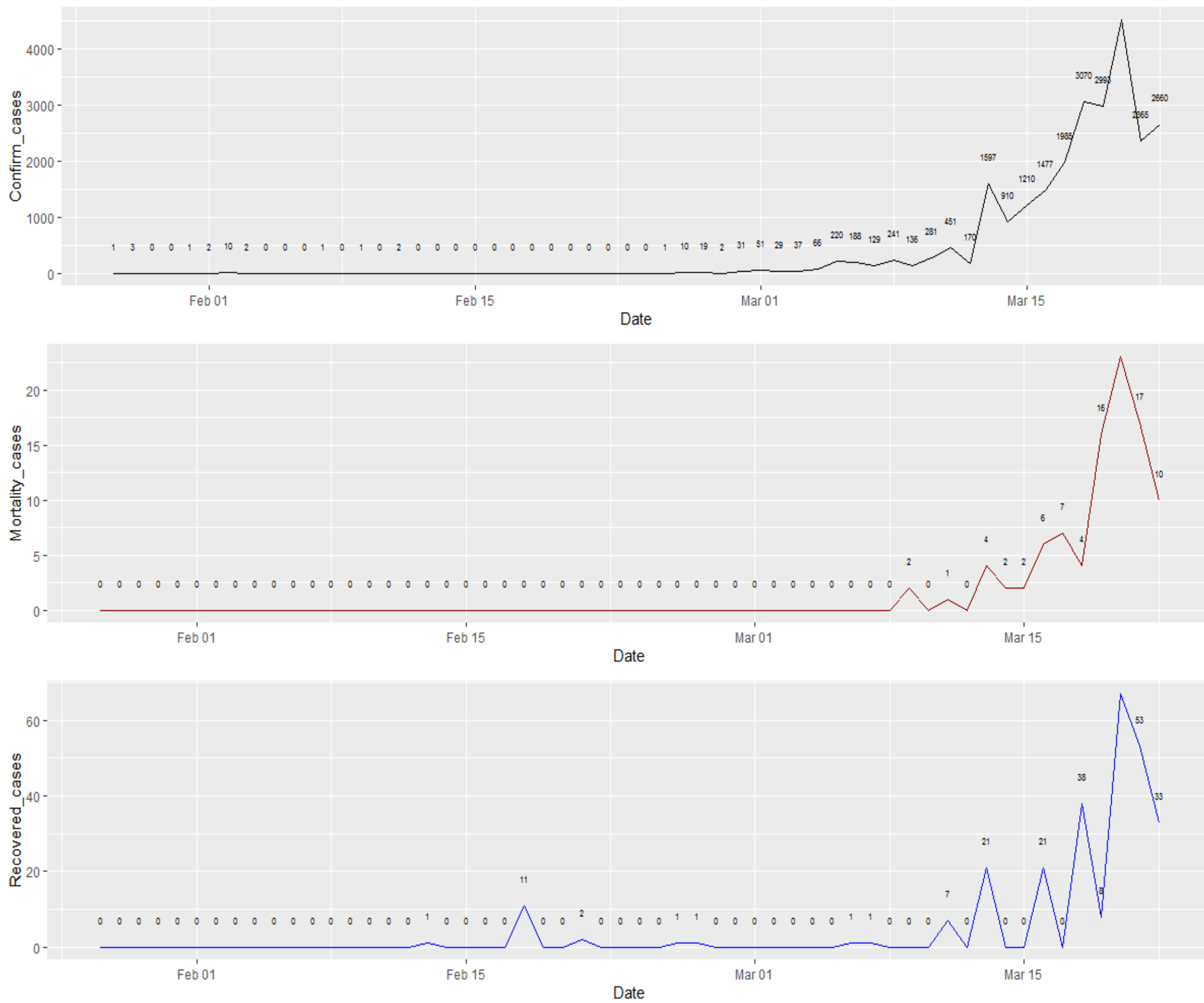
Day wise Confirmed, Mortality and recovered cases – Italy – Upto 22nd March



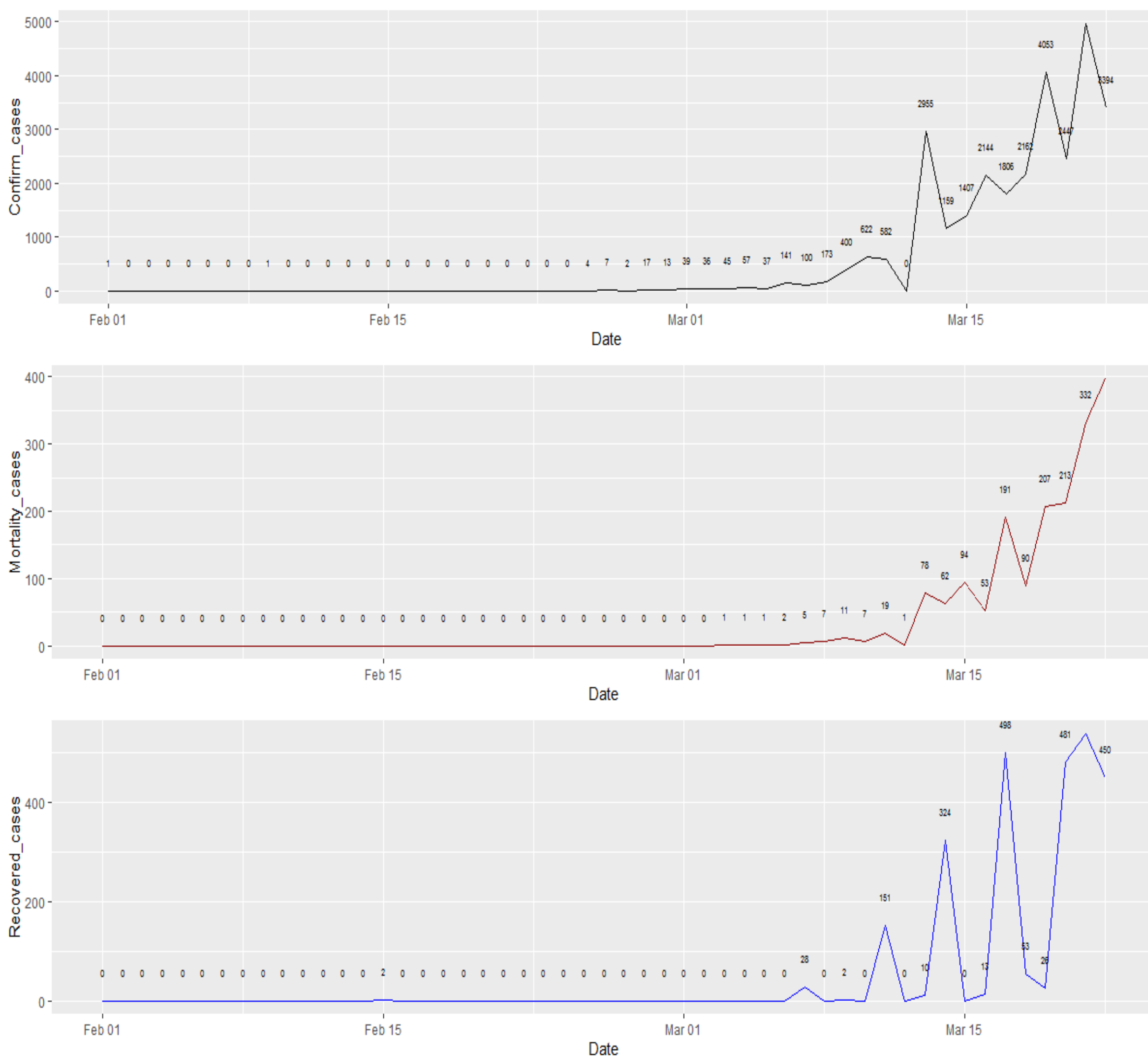
Day wise Confirmed, Mortality and recovered cases – France – Upto 22nd March



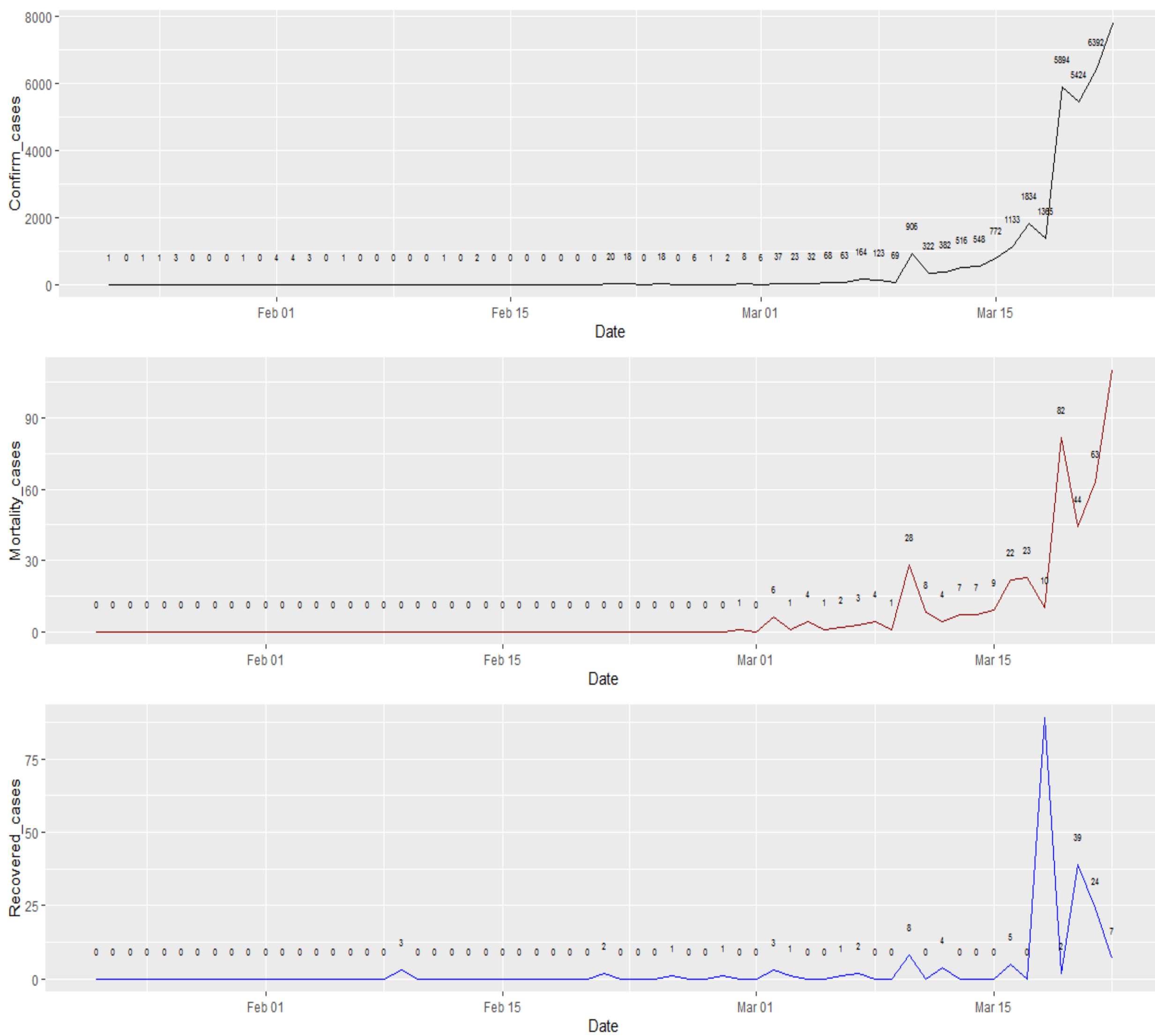
Day wise Confirmed, Mortality and recovered cases – Germany – Upto 22nd March



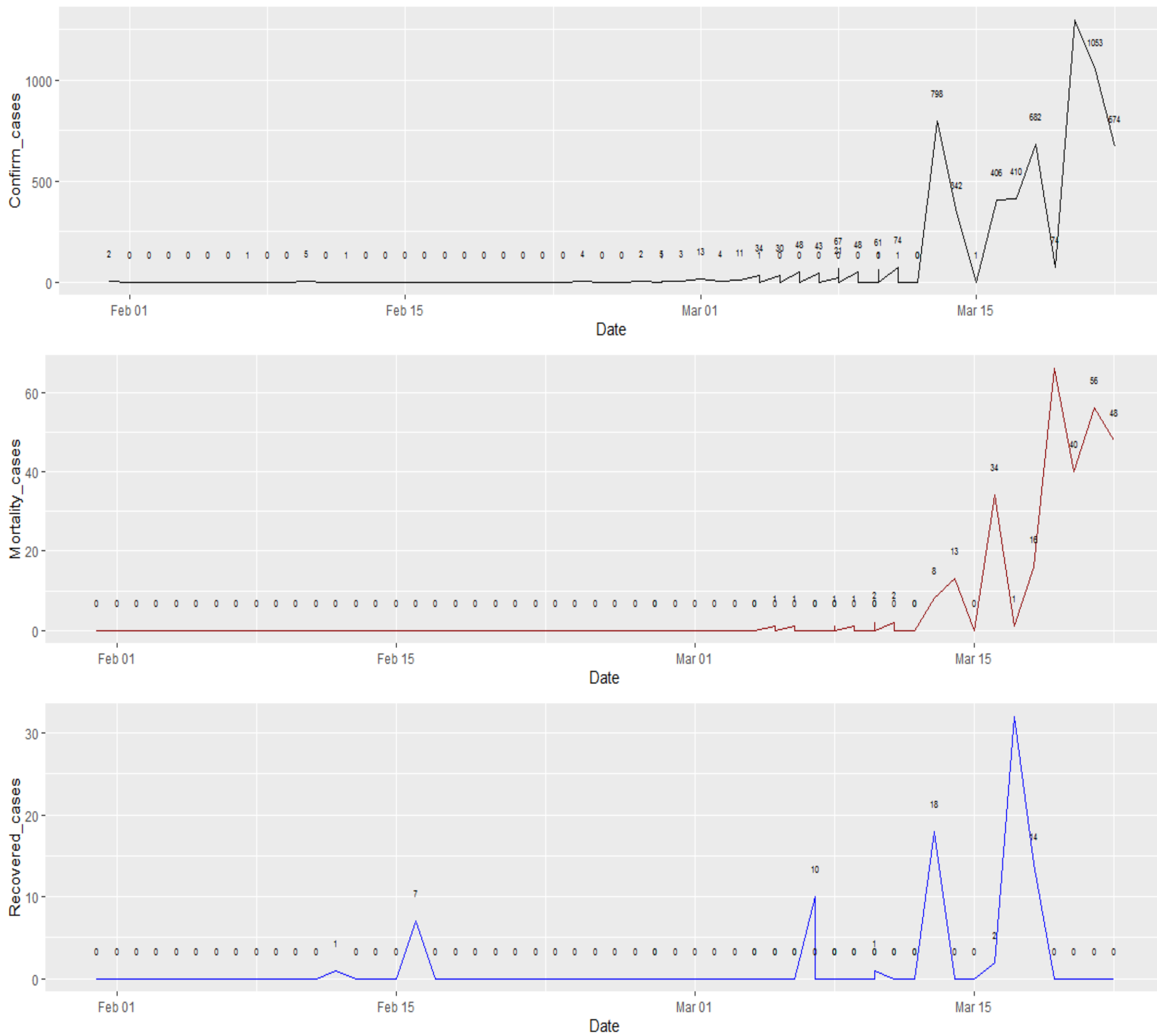
Day wise Confirmed, Mortality and recovered cases – Spain – Upto 22nd March



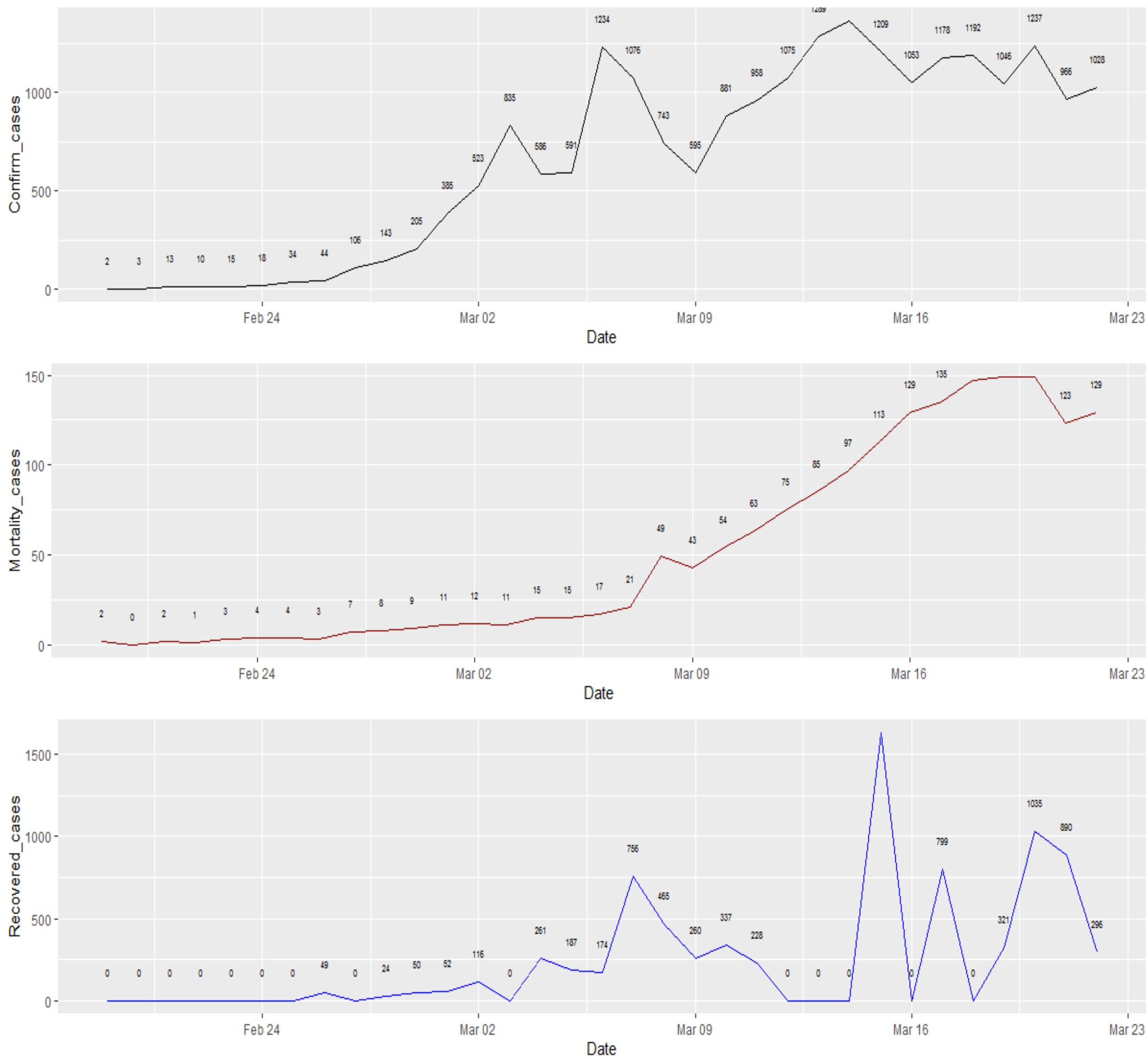
Day wise Confirmed, Mortality and recovered cases – US – Upto 22nd March



Day wise Confirmed, Mortality and recovered cases – UK – Upto 22nd March



Day wise Confirmed, Mortality and recovered cases – Iran – Upto 22nd March



Univariate Analysis – Weekday wise reporting of cases upto March 22nd 2020

Sl.No.	Weekday	Confirm cases	Mortality cases	Recovered cases
1	FRIDAY	63,228	2,542	12,917
2	MONDAY	29,813	1,422	11,168
3	SATURDAY	59,058	2,738	18,225
4	SUNDAY	60,401	2,950	17,698
5	THURSDAY	55,380	1,868	12,121
6	TUESDAY	34,167	1,561	12,767
7	WEDNESDAY	35,270	1,585	13,161

5. Time Series Model Architecture

Summary of Prediction Algorithm for India Confirmed cases prediction

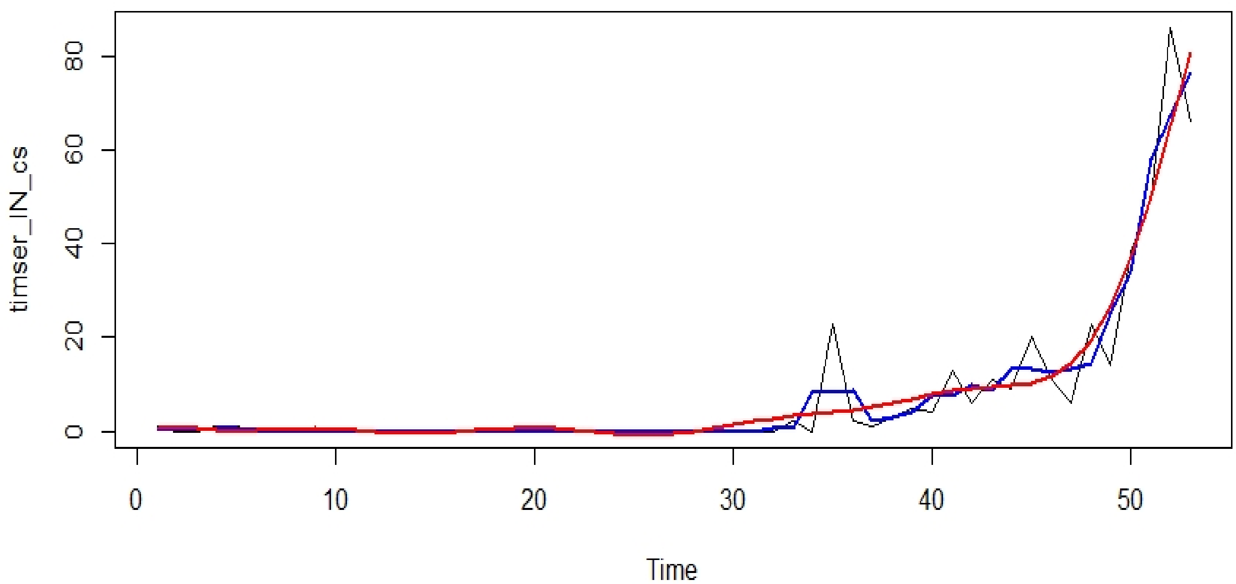
```
Call:
lm(formula = Confirmed_cases ~ sin(0.3885 * day1) * poly(day1,
    3) + cos(0.3795 * day1) * poly(day1, 3) + day1, data = smootheddf1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9404 -0.7021 -0.1761  0.5868  7.9904

Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                         7.5504     0.3917  19.274  < 2e-16 ***
sin(0.3885 * day1)                  5.1336     0.6349   8.085 5.07e-10 ***
poly(day1, 3)1                     84.9758     7.4975  11.334 3.28e-14 ***
poly(day1, 3)2                     54.9056     3.5546  15.447  < 2e-16 ***
poly(day1, 3)3                     44.6303     7.8959   5.652 1.35e-06 ***
cos(0.3795 * day1)                 -1.6913     0.6699  -2.525 0.015537 *
day1                                    NA         NA      NA       NA
sin(0.3885 * day1):poly(day1, 3)1  33.7104     6.2715   5.375 3.33e-06 ***
sin(0.3885 * day1):poly(day1, 3)2  44.3826     5.3729   8.261 2.92e-10 ***
sin(0.3885 * day1):poly(day1, 3)3  17.0469     7.6131   2.239 0.030634 *
poly(day1, 3)1:cos(0.3795 * day1) -31.1286     6.3042  -4.938 1.37e-05 ***
poly(day1, 3)2:cos(0.3795 * day1) -22.2399     5.6049  -3.968 0.000285 ***
poly(day1, 3)3:cos(0.3795 * day1) -34.5725     7.0414  -4.910 1.50e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.502 on 41 degrees of freedom
Multiple R-squared:  0.9816,   Adjusted R-squared:  0.9767
F-statistic:   199 on 11 and 41 DF,  p-value: < 2.2e-16
```

Model is optimized as intercept and all co-ordinates have Probability of less than 0.05

6. Results and Discussion

Time series, trend and forecasting of above Sinusoidal Polynomial algorithm to the confirmed cases of timeseries data.

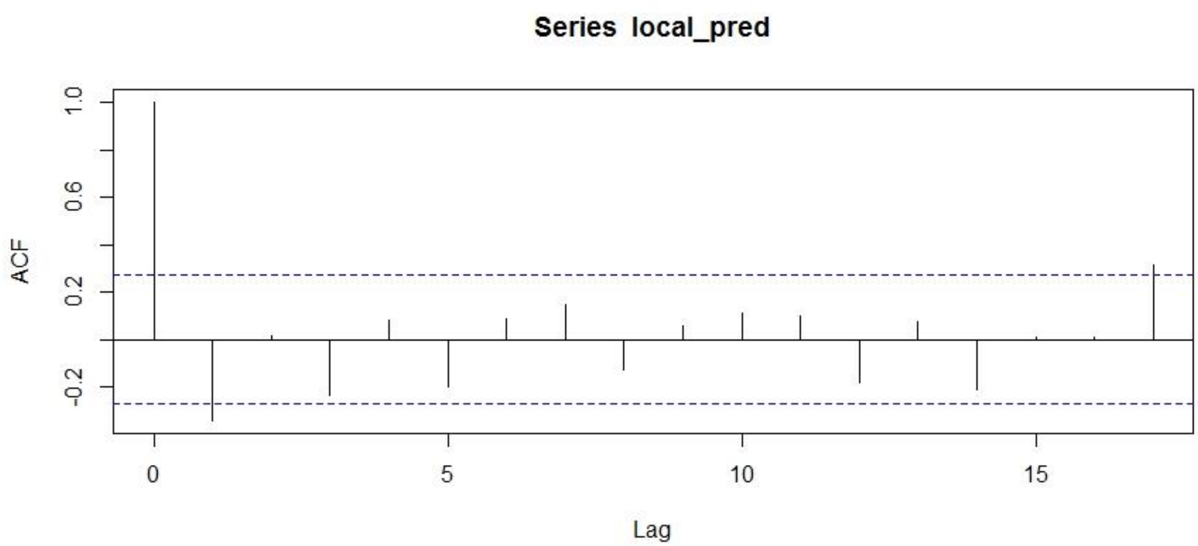
Time is in days from 30th of Jan 2020 to 22nd March 2020. Black line is the actual data of confirmed cases, blue line is the trend fitting and red line is the forecast using the algorithm. Difference between red and black line is regarded as white noise. White noise is a discrete signal whose samples are regarded as a sequence of serially uncorrelated random variables with zero mean and finite variance; a single realization of white noise is a random shock

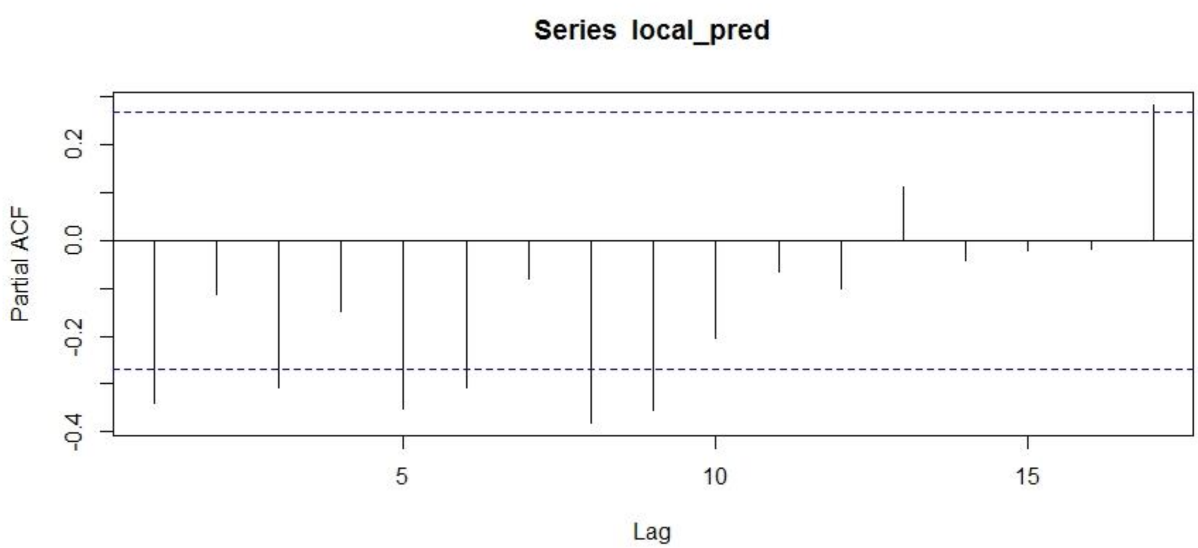


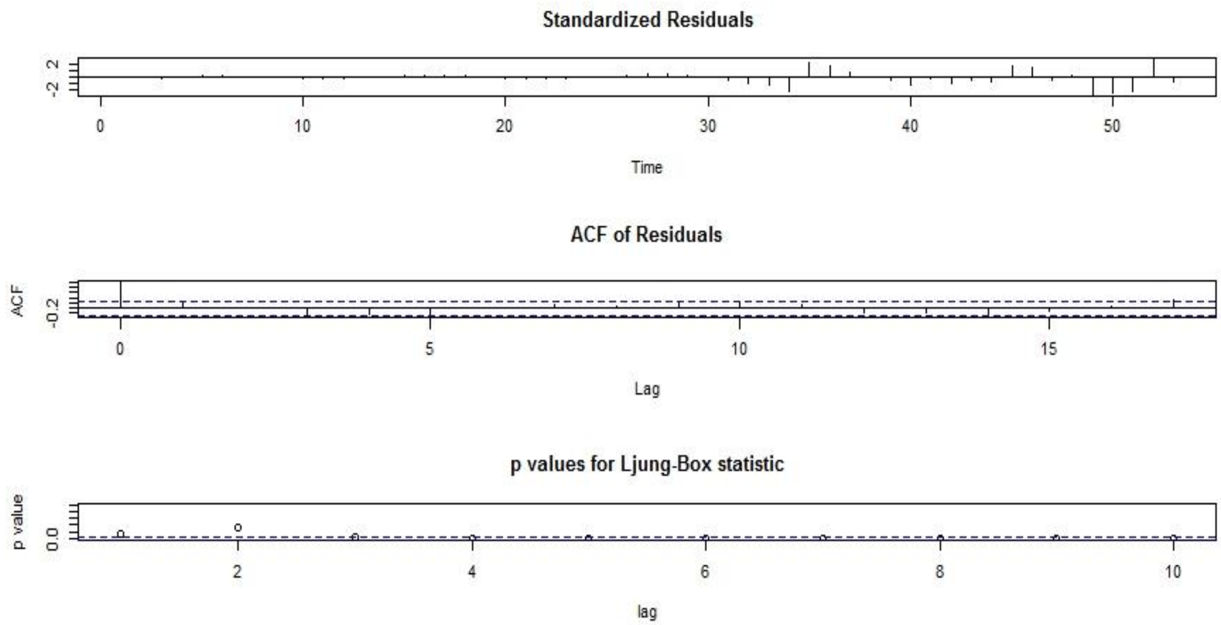
ACF and PACF Plots

ACF or Autocorrelation function plot will help one to understand how a given time series is correlated within its own values. It tries to find the relation between the latest values of the series with the historical past values. A time series data as stated in section 2 can have characteristics of cyclicity, seasonality, trend and residual. ACF incorporates all these aspects in a time series and find the correlation coefficient between present and past values. [4]

PACF or Partial autocorrelation function plot at lag k is the correlation that results after removing the effects which are already explained by the earlier lags(s), hence partial as we remove already found variations before we find the next correlation. So, if there is any hidden information in the residual which can be modeled by the next lag, we might get a good correlation and we will keep that next lag as a feature while modeling. Remember while modeling we don't want to keep too many features which are correlated as that can create multicollinearity issues. Hence, we need to retain only the relevant features. [5]







Forecast and Actual confirmed cases are shown below. We also have forecast up to 5th April 2020. MAPE of the model is at 35.27 % with accuracy of the model at 64.73%. Model will be more accurate once we retrain the model with actual confirm cases numbers. Similarly, model has been developed for worldwide model, country specific models for confirmed, mortality and recovery of COVID19 cases. These models are not discussed in this paper.

Date	day1	Actual Confirm_cases	Forecast of Confirm Cases
11-03-20	42	6	9
12-03-20	43	11	9
13-03-20	44	9	10
14-03-20	45	20	10
15-03-20	46	11	12
16-03-20	47	6	14
17-03-20	48	23	19
18-03-20	49	14	27
19-03-20	50	38	37
20-03-20	51	50	50
21-03-20	52	86	65
22-03-20	53	66	81
23-03-20		159	96
24-03-20		37	108
25-03-20		121	117
26-03-20			120
27-03-20			118
28-03-20			110
29-03-20			98
30-03-20			86
31-03-20			75
01-04-20			70
02-04-20			74
03-04-20			90
04-04-20			118
05-04-20			158

References

[1] Prof. Sreenivasa Raghavan, *Lecture on Time Series in IIIT-B,* Bangalore, 2017.

[2] P. S. Raghavan, *PGDDA-Lecture Notes/PA-II/3 Time Series Analysis,* Banglaore: IIIT - B, 2017.

[3] "Github," March 2020. [Online]. Available: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports. [Accessed 16 March 2020].

[4] Sangarshanan, "Towards Datascience," 3 October 2018. [Online]. Available: https://towardsdatascience.com/time-series-forecasting-arima-models-7f221e9eee06.

[5] J. Salvi, "Towards Datascience," Medium, 27 March 2019. [Online]. Available: https://towardsdatascience.com/significance-of-acf-and-pacf-plots-in-time-series-analysis-2fa11a5d10a8.

Appendix

[1] R – code for Time series ARIMA – Classical decomposition method modelling, EDA and data preparation

https://github.com/vikram-sreedhar/Timeseries-for-Covid19-for-India/blob/master/corona_pred_TS.R