International Journal of Future Generation Communication and Networking
Vol. 14, No. 1, (2021), pp. 814–823

Forecasting Covid-19 Confirmed Cases Deaths Andrecoveredusing Vector Autoregression

Dr.N.Poompavai
Department of computer science.
Bishop packiamArokiaswamy college of Education,
Pudukkottai -622303.
Email Poompavai84@gmail.com

Abstract

This article is about the recent complications faced by the people and government of the entire world. Thecomplication is all about an infectious disease caused by Anewlydiscoveredviruscalledcoronavirus (covid- 19). This has caused heavy damage to people in the society and also to the government. It took over lakhsofpeople'slivesanddreams.Insuchcasesthegovernmentneedstotakeprecaution.Theyshould be able to predict the future using the statistical report. Here in this article we have used a VectorAutoRegression algorithm for time series data to predict the future which shows how many people ofIndia are going to be affected, recovered and etc in the next five days. We have done this using the statistical report of India. We have proved this with good accuracy. This article is highly useful to predict the future and using this government can take precaution before the outbreak of an pandemic may cause damage or loss to the society.

Keywords:*Infectious disease,Coronavirus,Precauton,Pandemic*

1.Introduction

First of all let us know what Vector Auto Regression is. Vector Auto Regression (VAR) which is mentioned as VAR in many cases is an multivariate forecasting algorithm and is a scholastic process model used to capture the linear among the multiple time series variables. We call it as autoregressive because each variable is represented as a consequence of a past value,that is the predictors are time delayed values of the series. This is used when one wants to predict multiple series variables. This was developed by Sims in the year 1980. This algorithm is used in many industry,supply,economic,constructions,consumption,savingandetc.So,ifweuseanautoregressive algorithm in cases of Covid-19 it will be very appropriate. Here in this article we have collected the data of covid-19 in India from the website kaggle.com. So now by implementing the algorithm in the collected data we will reach an accuracy which may help us to predict the future, that is we are going to predict the number of fatality, recovery, and also the new cases to be affected due toCovid-19.

2.VectorautoRegression:

Vector autoregression(VAR) model is an augmentation of univariateautoregression model to multiple time series data. Here all the variables are dependent on itself, that is they are treated as endogenous. This is the most successful, flexible and easy model used for the analysis of multivariate time series data. It often provides superior prediction to those from univariate time series. In addition to data illustration and forecasting, this model is also used for analysis and structural intervention. The struc- ture of this model allows us to test restriction around multiple equations and it also tests whether the coefficients on all regression of the lag**p**are zero. This corresponds to testing the null value in the lag order.

ISSN: 2233-7857 IJFGCN

Before adding the variables to the VAR model we should recheck it carefully because adding unrelatedvariablestothe VAR modelreducesaccuracybyincreasingtheestimationerrors.Thetypical AR model equation is givenbelow

Table 1. Data Format

Date	Confirmed	Deaths	Recovered
2020-04-21	20080	645	3975
2020-04-22	21370	681	4370
2020-04-23	23077	721	5012
2020-04-24	24530	780	5498
2020-04-25	26283	825	5939

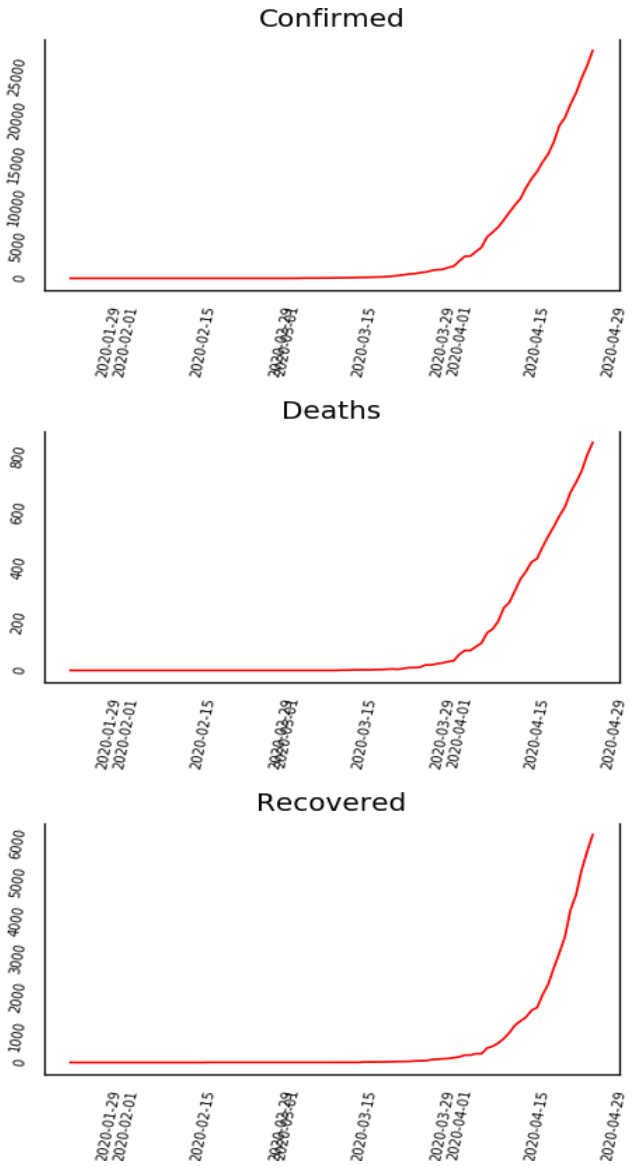

Figure 1. Graphical Representation of Covid Data

ISSN: 2233-7857 IJFGCN

3.Experimentalresults:

Thissectionisintendedtodiscuss the resultswhichisgeneratedusingthemethodsgiveninsection4.1 and4.2.

3.1 Granger's causality test:

This is a statistical hypothesis test in deciding whether one time series is useful in forecasting the other. This is a technique that seeks the direction of causality between the import and export files. According to the Granger causality test, if a signal in X_1"Granger-causes" (g-causes) signals X_2then thepastvaluesofX_1shouldcontaininformationthathelpstopredictX_2.Thistechniquewasdeveloped by Granger in 1960. Its formula is based on the linear regression model of scholarly process. The illustration of formula is given below,Consider a bivariate linear autoregressive models whose variables are X1 and X2 :

$$X_1(t) = \sum_{j=1}^{p} A_{11,jX_1}(t-j) + \sum_{j=1}^{p} A_{12,jX_2}(t-j) + E_1(1)$$

$$X_2(t) = \sum_{j=1}^{p} A_{21,jX_1}(t-j) + \sum_{j=1}^{p} A_{22,jX_2}(t-j) + E_2(2)$$

,

Here,

a. p is the maximum number of lagged observation included in themodel

b. Matrix A contains the coefficients of themodel

c. $X_1$and X_2arevariables

d. $E_1$and E_2areresiduals

WhereifthevarianceofE1isreducedbytheinsertionoftheX2termsinthefirstequationthenitis said that X_2g-causes X_1. This can also be said as X_2g-causes X_1if the coefficients in matrix $A_{12}$are different fromzero.

We have tested the Granger's causality for the data collected by us and the resulted output is given in Table 2

Table 2. Granger's Casuality Matrix

	Confirmed_x	**Deaths_x**	**Recovered_xConfirmed_y** 1.0	0.0
	0.0			
Deaths_y	0.0	1.0	0.0	
Recovered_y	0.0	0.0	1.0	

ISSN: 2233-7857 IJFGCN

3.2 Cointegration Test:

Cointegration test is a statistical property of a time series variable. To know more about cointegration first let us know what is 'order of integration'. The order of integration means the differentiation done to convert the non-stationary time series to stationary. Now, in multiple series there exists a linear combination of multiple series that has order of integration less than that of individual series, and hencethecollectionofsuchseriesiscalledcointegration.Whenmultivariateseriesiscointegratedthat means they will have a long runsignificantly.

Now by implementing the formula of cointegration by Johansen in the data collected we will get the result as shown in Table 3.

Table 3. Johansen cointegration test

Name	TeststatC(95%)	Signif	Confirmed
129.4	24.2761	True	
Deaths	52.74	12.3212	True
Recovered	9.21	4.1296	True

3.3 Splitting The Series Into Test And Train Data

Toreducethecomplicationtheseriesissplitintotrainandtestdata.Herewehavesplittedthedatafrom 2020-01-22 to 2020-04-20 as train data and data from 2020-04-21 to 2020-04-25 as test data.After splitting the series into train and test data is fitted and the algorithm is implemented. Here we have dividedtheseriesintotrainandtest.Wehavegot90trainingdataand5testingdata.Andthereforethe shape of the train and test data are (90, 3) and (5,3).

3.4 Test For Stationary To Make The Series Stationary:

The time series we have collected should be stationary in case of the VAR model as it is going to be forecasted. It is complicated to check all the series according to stationarity. So, to reduce this com- plication we have this technique of test for stationary. Now to check stationarity without complication we have used a popular method namely Augmented Dickey-Fuller Test(ADF), these methods are also called unit root tests or suite of tests.

If a series is found non-stationary we make use of the suite of tests. ADF tests the null value inthe time series. This is the large and complicated set of time series models. The ADF test statistic usesthenegativenumber.

Formula of ADF test is given in Equation 4

$$Oy_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 Oy_{t-1} + \cdots + \delta_{p-1} Oy_{t-p+1} + \varepsilon t \quad (3)$$

Here,

a. y(t) =variable

b. y(t-1) = lag1 of timeservice

c. Δy(t-1) = first difference of the time series at time(t-1)

d. β = coefficient of the timetrend

e. α =constant

f. p = lag order of autoregressiveprocess

g. Y = nullhypothesis

Implementing ADF test in the collected data we will get the result shown in Table 4, Table 5, Table6 respectively.

Table 4.Augmented Dickey-Fuller Test forConfirmed.

Augmented Dickey-Fuller Test on "Confirmed"
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level = 0.05
Test Statistic = 2.7015
No. Lags Chosen = 11
Critical value 1% =-3.517
Critical value 5% =-2.899
Critical value 10% = -2.587
=> P-Value = 0.9991. Weak evidence to reject the Null Hypothesis.
=> Series is Non-Stationary.

Table 5. Augmented Dickey-Fuller Test for Death

Augmented Dickey-Fuller Test on "Deaths"
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level = 0.05
Test Statistic = -2.5725
No. Lags Chosen = 12
Critical value 1% = -3.518
Critical value 5% = -2.9
Critical value 10% = -2.587
=> P-Value = 0.0988. Weak evidence to reject the Null Hypothesis.
=> Series is Non-Stationary.

Table 6. Augmented Dickey-Fuller Test forRecovered

Augmented Dickey-Fuller Test on "Recovered"
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level = 0.05
Test Statistic = 8.2093
No. Lags Chosen = 9
Critical value 1% =-3.515
Critical value 5% =-2.898
Critical value 10% = -2.586
=> P-Value = 1.0. Weak evidence to reject the Null Hypothesis.
=> Series is Non-Stationary.

So this shows us that none of our time series is stationary so to convert these non-stationary series to stationary we need to differentiate the series.

On differentiating the series for the first time we got all the output as non-stationary so now we havetwochoiceseithertopursuewith1stdifferenceorweneedtodifferentiatetheseriesoneormore times until the series may get converted tostationary.On differentiating the series for the fifth time we will get the following output,

Table 7. Augmented Dickey-Fuller Test on"Confirmed" after fifth difference

Augmented Dickey-Fuller Test on "Confirmed"
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level = 0.05
Test Statistic = -4.3026
No. Lags Chosen = 12
Critical value 1% =-3.525
Critical value 5% =-2.903
Critical value 10% = -2.589
=> P-Value = 0.0004. Rejecting Null Hypothesis
=> Series is Stationary.

Table 8. Augmented Dickey-Fuller Test on "Death"after fifth difference

Augmented Dickey-Fuller Test on "Deaths"
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level = 0.05
Test Statistic = -10.1675
No. Lags Chosen = 8
Critical value 1% = -3.519
Critical value 5% = -2.9
Critical value 10% = -2.587
=> P-Value = 0.0. Rejecting Null Hypothesis
=> Series is Stationary.

Table 9. Augmented Dickey-Fuller Test on "Recovered"after fifth difference

Augmented Dickey-Fuller Test on "Recovered"
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level = 0.05
Test Statistic = -10.6075
No. Lags Chosen = 12
Critical value 1% =-3.521
Critical value 5% =-2.903
Critical value 10% = -2.589
=> P-Value = 0.0. Rejecting Null Hypothesis
=> Series is Stationary.

And the above stated output is the resultant output of the series in differentiating it for the fifth time.

3.5 Selection Of The Order [P] VarModel :

Now, we need to select the right order of the VAR model. We fit increasing orders of VAR models and pick any one order which gives a model with least Akaike information criterion (AIC). We can also check other best fitting comparison estimations and they are Bayesian Information Criterion (BIC),Final Prediction Error (FPE) and Hannan-Quinn Information Criterion (HQIC).
Given below is the resultant output we got by fitting the correct order,

Table 10. Best Fitting Comparison Estimations

	AIC	BIC	FPE	HQIC
0	22.03	22.12	3.7e+09	22.07
1	21.19	21.55	1.5e+09	21.33
2	20.37	21.00	7.0e+08	20.62
3	20.14	21.04	5.1e+08	20.50
4	19.48	20.65	2.9e+08	19.95
5	18.21	19.65	8.2e+07	18.78
6	15.83	17.54	7.6e+06	16.51
7	13.61	15.59	8.4e+05	14.40
8	12.59	14.84	3.1e+05	13.49
9	11.56	14.08	1.1e+05	12.57

4.Check For Serial Correlation Of Residual [Errors] Using Durbin Watson Statistic:

Thisisusedtocheckiftherareanyleftoverpatternsintheerrors.Thatmeanstherearesomepatterns that are still left to be explained by the model. In such cases, we have three choices that can be done either we can increase the order of the model or we can induce more predictors or we can look for differentalgorithmstomodelthetimeseries.TocheckcorrelationoferrorswecanuseDurbinWatson Statistics.

Formula we use to check the correlation is mentioned below,

$$DW = \frac{\sum_{t=2}^{T} ((e_t - e_{t-1})^2)}{\sum_{t=1}^{T} e_t^2}$$

The resultant value from the implemented formula may vary between 0 to 4. If the value is closer to 0 that means it contains positive correlation, if the value is closer to 4 that means it has negative correlation and if the value is closer to 2 that means there is no serial correlation in theseries.

Now let us implement the formula in the data we have collected,

Confirmed : 2.18

Deaths : 2.17

Recovered : 2.34

Theabovementionedvaluesaretheresultantoutputwegotfromimplementingtheformula Durbin Watson Statistic. So, now the output value is closer to 2 and so we may not have any significant correlation in the multivariate time series data we havecollected.

4.1 Forecasting Var Model:

Inordertoforecast,themodelexpectsthelagordernumberofobservationsfromthepastdata.Thisisbecauset hemodellagsvarioustimeseriesinthedataset.So,weneedtoprovidemanypreviousvaluesas indicated by the lag order used by themodel.The fitted lag order of the collected data is given below,

array([[-1388., -175., -1244.],
[2024., 128., 1232.],
[-1652., -34., -930.],
[-1008., -18., 631.]])

The above described data has been forecasted but it is on the scale of training data used by the model. To bring it back to its original form we need to de-differentiate it as many times we have

differentiatedtheoriginalinputdata.Herewedifferentiatedtheoriginalinputdata5times,sonowwe need to de-differentiate this resulted forecasted data 5times.

4.2 Inverting The Transformation To Get The Real Forecast:

Now, let us de-differentiate the forecasted data to get the original data. The de-differentiation of the forecasted data is given below,

Table 11. Forecasting Covid19 dataset

Date	Confirmedforecast	Deathsforecast	Recoveredforecast
2020-04-21	20246	635	3685
2020-04-22	21232	674	4249
2020-04-23	22873	708	4695
2020-04-24	24451	762	5303
2020-04-25	26299	799	5744

And finally the original input data is resulted by de-differentiation which is mentioned above.

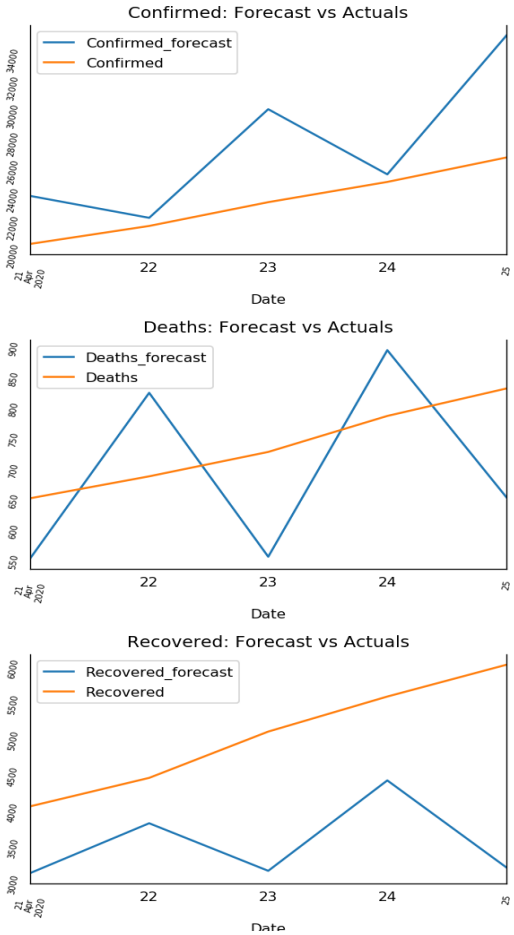

Figure 2. Plotting ForecastVs Actual

5. Evaluating forecast:

To evaluate the forecast we need to compute a comprehensive set of metrics. The metric we have usedheretoevaluatetheaccuracyisMAE(MeanAbsoluteError).MAEisthesimplesterrormetricin regression that is understandable. On evaluating the forecast in the collected data according to MAE we will get the followingoutput,

Forecast Accuracy of: Confirmed Cases mae : 120.5973
Forecast Accuracy of: Deaths mae : 14.2689
Forecast Accuracy of: Recovered mae : 223.0366

6.Conclusion:

In this article we have studied about the efficiency vector auto regression from its scratch to the end detail. Using the collected data of covid-19 pandemic in India we have evaluated the successfulmodel using Vector Auto Regression algorithm with good accuracy. So, now we can predict the future cases, deathsandrecoveryusingthismodelwhichhelpsgovernmenttotakenecessaryactionsin precautionary.

References

[1] Andrews, D. W. K. (1999), "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," Econometrica, 67, 543–564. DOI: 10.1111/1468-0262.00036. [Crossref], [Web of Science ®], [Google Scholar]

[2] Bekaert,G.,Engstrom,E.,andErmolov,A.(2017)"MacroRisksandtheTermStructureofInterestRates,"Finance and Economics Discussion Series2017-058.Washington,DC:BoardofGovernors of the Federal Reserve System. DOI: 10.17016/FEDS.2017.058. [Crossref], [Google Scholar]

[3] Bernanke, B. S., and Mihov, I. (1995) "Measuring Monetary Policy," NBER Working PaperNo.

[4] 5145. [Google Scholar] Brüggemann, R., Jentsch, C., and Trenkler, C. (2016), "Inference in VARs with Conditional Het-eroskedasticityofUnknownForm,"JournalofEconometrics,191,69–85.DOI:10.1016/j.jeconom.2015.10.004. [Crossref], [Web of Science ®], [GoogleScholar]

[5] Castelnuovo, E. (2016), "Monetary Policy Shocks and CholeskyVARs: An Assessment for the Euro Area," Emprical Economics, 50, 383–414. DOI: 10.1007/s00181-015-0930-2. [Crossref], [Web of Science ®], [GoogleScholar]

[6] Chaussé, P. (2010), "Computing Generalized Method of Moments and Generalized Empirical Likelihood with R," Journal of Statistical Software, 34, 1–35. DOI: 10.18637/jss.v034.i11. [Crossref], [Web of Science ®], [GoogleScholar]

[7] Donovon,P.,andHall,A.R.(2018),"TheAsymptoticPropertiesofGMMandIndirectInference underSecondOrderIdentification,"JournalofEconometrics,205,76–111.DOI:10.1016/j.jeconom.2018.03.006 [Crossref],[Web of Science ®], [Google Scholar]8. Gospodinov, N. (2010), "Inference in Nearly Non- stationary SVAR Models with Long –Run Identifying Restrictions,"Journal of Business &Economic Statistics, 28, 1–12. DOI: .1198/jbes.2009.08116. [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

[8] Gospodinov, N., Kan, R., and Robotti, C. (2014), "Spurious Inference in Unidentified Asset-Pricing Models," Federal Reserve Bank of Atlanta Working Paper Series, 2014-12. [GoogleScholar]

[9] Gospodinov,N.,Kan,R.,andRobotti,C.(2017),"SpuriousInferenceinReduced-RankAsset- Pricing Models," Econometrica, 85, 613–1628. DOI: 10.3982/ECTA13750. [Crossref], [Web of Science®], [GoogleScholar]

[10] Gospodinov, N., and Ng, S. (2015), "Minimum Distance Estimation of PossiblyNoninvertible MovingAverageModels,"JournalofBusiness&EconomicStatistics,33,403–

417.DOI:10.1080/07350015.201[Taylor & Francis Online], [Web of Science ®], [GoogleScholar]

[11] Gouriéroux, C., Monfort, A., and Renne, J.-P. (2017), "Statistical Inference for Independent Component Analysis: Appliction to SructuralVARModels," Journal of Econometrics, 196, 111–126. DOI: 10.1016/j.jeconom.2016.09.007. [Crossref], [Web of Science ®], [GoogleScholar]

[12]Guay, A., and Normandin, M. (2018), "Identification of Structural Vector Autoregresions Through Higher Unconditional Moments," Unpublished Working Paper. [GoogleScholar]

[13] Hall, A. R. (2005), Generalized Method of Moments, New York: Oxford University Press. [GoogleScholar]

[14]Hall,A. R., and Inoue, A. (2003), "The Large Sample Behavior of the Generalized Method ofMomentsEstimatorinMisspecifiedModels,"JournalofEconometrics, 114,361394. DOI:10.1016/S030476(03)00089 [Crossref], [Web of Science ®], [GoogleScholar]

[15] Hall,A.R.,Inoue.A.,Jana,K.,andSin,C.(2007),"InformationinGeneralizedMethodofMoments Estimation and Entropy-Based Moment Selection," Journal of Econometrics, 138, 488–512. DOI: 10.1016/j.jeconom. 2006.05.006. [Crossref], [Web of Science ®], [Google Scholar]

[16] Hall, A. R., and Peixe, F. P. M. (2003), "A Consistent Method for the Selection of Relevant Instruments," Econometric Reviews, 7, 269–288. DOI: 10.1081/ETC-120024752. [Taylor & Francis Online], [GoogleScholar]

[17] Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of MometnsEstima- tors," Econometrica, 50, 1029–1054 DOI: 10.2307/1912775. [Crossref], [Web of Science ®], [Google Scholar]

[18]Hansen,L.P.,Heaton,J.,andYaron, A.(1996),"Finite Sample Properties of Some Alternative GMMEstimatorsObtainedfromFinancialMarketData,"JournalofBusinessandEconomic Statistics,14,262–280.DOI:10.2307/1392442.[Taylor&FrancisOnline],[WebofScience®],[GoogleScholar]

[19] Herwartz, H. (2015), "Structural VAR Modelling with Independent Innovations—AnAnalysis of Macroeconomic Dynamics in the Euro Area Based on a Novel Identification Scheme," Working Paper, University of Göttingen. [GoogleScholar]