Biology Methods and Protocols, 2023, bpac035
https://doi.org/10.1093/biomethods/bpac035
Advance Access Publication Date: 15 December 2022
Methods Article

Improved prediction of new COVID-19 cases using a simple vector autoregressive model: evidence from seven New York state counties

Takayoshi Kitaoka[1] and Harutaka Takahashi [2,3,*]

[1]Meiji University, Tokyo, Japan
[2]Meiji Gakuin University, Tokyo, Japan
[3]Graduate School of Economics, Kobe University, Kobe, Japan

*Correspondence address. Graduate School of Economics, Kobe University, Kobe, Japan. Tel: +81-78-803-7245 FAX : +81-78-803-7289; E-mail: haru@eco.meijigakuin.ac.jp

Abstract

With the rapid spread of COVID-19, there is an urgent need for a framework to accurately predict COVID-19 transmission. Recent epidemiological studies have found that a prominent feature of COVID-19 is its ability to be transmitted before symptoms occur, which is generally not the case for seasonal influenza and severe acute respiratory syndrome. Several COVID-19 predictive epidemiological models have been proposed; however, they share a common drawback – they are unable to capture the unique asymptomatic nature of COVID-19 transmission. Here, we propose vector autoregression (VAR) as an epidemiological county-level prediction model that captures this unique aspect of COVID-19 transmission by introducing newly infected cases in other counties as lagged explanatory variables. Using the number of new COVID-19 cases in seven New York State counties, we predicted new COVID-19 cases in the counties over the next 4 weeks. We then compared our prediction results with those of 11 other state-of-the-art prediction models proposed by leading research institutes and academic groups. The results showed that VAR prediction is superior to other epidemiological prediction models in terms of the root mean square error of prediction. Thus, we strongly recommend the simple VAR model as a framework to accurately predict COVID-19 transmission.

Keywords: COVID-19 case forecast; vector autoregression; epidemiological model; root mean square error (RMSE); New York state counties

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), identified in 2019, has caused the coronavirus disease 2019 (COVID-19) pandemic. With the rapid spread of COVID-19, there is an urgent need for a framework to accurately forecast COVID-19 progression. To this end, a variety of COVID-19 epidemiological forecasting models have been proposed by major research institutes. Wang *et al.* [1] classified forecasting models into three categories: (i) mechanistic models; (ii) time series models, and (iii) models based on deep learning. Examples of mechanistic models are the susceptible–infected–recovered (SIR) model and the modified susceptible–exposed–infected–recovered (SEIR) population propagation model. The majority of deep learning models extend mechanistic models with deep learning methods. In this study, we compared the forecasting accuracy of our model to that of 11 state-of-the-art forecasting models proposed by major research institutes and academic groups (These models are cited by CDC).

To predict the number of new COVID-19 cases by county, Shang *et al.* [2] recently proposed a data-driven regression model called the vector autoregression (VAR) epidemiological model (They proposed the VAR model to predict the nationwide daily number of newly COVID-19 cases in the USA. However, they arbitrarily determine the maximum lag length of 3 (very short) and apply AIC to each equation to select the optimal lag within lags 1, 2, and 3. As a result, the individual equations of the VAR model have different lag structures. On the other hand, maximum lags length of our VAR model is 14 (long) and optimal lags are determined by VAR system information criteria (system AIC). Therefore, lag structure of each equation is the same). However, the VAR model of Shang *et al.* [2] does not obey the usual way of VAR model forecasting procedures developed by Sims [3]. Furthermore, Wang *et al.* [4] proposed another VAR model employing different variables than ours to predict the number of new COVID-19 patients nationwide.

VAR is a time series model and contrasts with mechanistic and deep learning models in two aspects: (i) VAR solely uses county-level new COVID-19 cases as the forecasting data and (ii) VAR captures COVID-19 cross-county transmission by introducing other counties' COVID-19 case data as lagged explanatory variables. The second point is important because to predict COVID-19 cases at the county level, it is necessary to consider cross-county infection as a transmission mechanism of SARS-CoV-2. This is different from that of other viral infections such as seasonal influenza and SARS. To characterize the transmission dynamics of COVID-19, two

Received: September 07, 2022. **Revised:** December 07, 2022. **Accepted:** December 08, 2022

important epidemiological terms were introduced: the incubation period (the time between infection and the onset of symptoms) and the serial interval (the time between the onset of disease in the primary infected person and the onset of disease in the secondary infected person). As estimated by Nishiura *et al.* [5], He *et al.* [6], and Alene *et al.* [7], the estimated mean serial interval and the incubation period of COVID-19 are 5.2 and 6.5 days, respectively. Notably, the estimated serial interval is shorter than the estimated incubation period (Note that the serial interval can be negative if a person becomes infected before symptoms appear in the individual who infected them, that is, if the infected person develops symptoms before the person that infected them does). For seasonal influenza and SARS, the serial interval is longer than the incubation period. This indicates the following important feature of COVID-19 – in contrast to seasonal influenza and SARS, a significant number of COVID-19 cases are caused by asymptomatic or pre-symptomatic infection. Owing to this feature of COVID-19, SARS-CoV-2 is not only transmitted among residents of the same county but also to residents of other counties through cross-county transmission, even before symptom onset. Shang *et al.* [2] note that the epidemiological models based on SIR or SEIR cannot capture this phenomenon (Some forecasting models attempt to incorporate the mobility behavior of individuals into the SRI-based model using a deep learning-based approach). In contrast, the VAR epidemiological model proposed herein does capture this feature by introducing new COVID-19 cases in other counties as a lagged explanatory variable.

As mentioned above, Shang *et al.* [2] proposed VAR as a promising COVID-19 forecasting model, but the authors did not demonstrate that the predictions made by VAR outperform those of other epidemiological models. Hence, it is not clear how good the VAR model forecasts are. The purpose of this study is to show that the county-level prediction of new COVID-19 cases by the simple VAR model is superior to that of other epidemiological models.

Methodology

In macroeconomics, economic forecasting is important for planning and evaluating government economic policy. In the 1970s, macroeconomists used large models consisting of hundreds of equations to make economic forecasts. However, since Sims [3] proposed VAR as a new macroeconomic forecasting method, no macroeconomists use such large models anymore.

VAR is a multi-equation system in which each variable is a linear function of the past lags of itself and the other variables. The popularity of VAR in economics is owing to its simple forecasting framework (forecasting by VAR model is said to be forecasting without theory. For a comprehensive introduction to VAR estimation, Stock and Watson [8] is recommended) while outperforming other forecasting frameworks. Here, we show that VAR performs similarly for predicting COVID-19 cases.

VAR

The regular VAR model with p lags, denoted by VAR_Lag p, can be written as follows:

$$\mathbf{y}_t = \mathbf{A}_0 + \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + \cdots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{C}\mathbf{x}_t + \mathbf{u}_t$$

where

$\mathbf{y}_t$: $n \times 1$ column vector of endogenous variables
$\mathbf{x}_t$: $m \times 1$ column vector of exogenous variables
$\mathbf{A}_0$: $n \times 1$ column vector of constant term
$\mathbf{A}_i$: $n \times n$ matrix of lag coefficients to be estimated ($i = 1, 2, \ldots, p$)
$\mathbf{C}$: $n \times m$ matrix of exogenous variable *coefficients to be estimated*
$\boldsymbol{u}_t$: $n \times 1$ column vector of disturbances.

In our model, the column vector is defined as follows:

$$\mathbf{y}_t = (y_{B,t}, y_{K,t}, y_{N,t}, y_{Q,t}, y_{R,t}, y_{W,t})'$$

where " ' " denotes transposition of a vector, $y_{i,t}$ indicates the number of newly confirmed COVID-19 cases in county i on day t, and B, K, N, NYC, Q, R, and W stand for the New York State counties Bronx, Kings, Nassau, New York City, Queens, Rockland, and Westchester, respectively.

Under the assumption that the time path $\mathbf{y}_t$ is stationary (see, in detail, Key Concept 14.5 in Stock and Watson [8]), $\boldsymbol{u}_t$ satisfies the following white noise disturbance process:

$$i)\ E(\boldsymbol{u}_t) = 0,\ ii)\ V(\boldsymbol{u}_t) = E(\boldsymbol{u}_t\boldsymbol{u}_t') = \boldsymbol{\Sigma},\ iii)E(\boldsymbol{u}_t\boldsymbol{u}_{t-s}')\ \mathbf{0} \text{ for s } > 0.$$

Assumptions *i*) through *iii*) imply that the vector of disturbances is contemporaneously correlated with full rank matrix $\boldsymbol{\Sigma}$, but uncorrelated with the leads and lags of the disturbances and uncorrelated with all of the right-hand side variables. Furthermore, each equation is estimated by the ordinary least squares method.

Again, VAR is a multi-equation system in which each variable is a linear function of the past lags of itself and the other variables. Such a framework allows VAR to adequately capture the nature of SARS-CoV-2 transmission at the county level and asymptomatic transmission between counties, which more accurately reflects the cross-county transmission that occurs through the cross-county movement of people.

Data

We analyzed daily new COVID-19 cases in the seven New York State counties assessed by Shang *et al.* [2] (Bronx, Kings, Nassau, New York City, Queens, Rockland, and Westchester). Bronx, Kings, and Queens are regarded as regions adjacent to New York City, and these areas are classified as a "large central metro" by the Centers for Disease Control and Prevention (CDC). In contrast, Nassau, Westchester, and Rockland have fewer direct connections to New York City; these counties are classified as a "large fringe metro" by the CDC. The data used in this study were downloaded from the following website: US COVID-19 cases and deaths by state | USAFacts (The URL of downloaded data file: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/new-york. County-level data was confirmed by referencing state and local agencies). The number of COVID-19 cases reflects the daily cumulative values for each county from 1 March 2020, through 8 August 2021. Based on the accumulated daily counts by county, the daily number of newly infected individuals was calculated by taking the difference, thus sample size of 519. We used newly infected individuals from the county-level daily data for our estimations. There were some days when the number of new cases was recorded as zero, such as February 6 and 26, and 12 March 2021. The reason why the number of new cases was marked as zero is probably due to a delay in recording. It is assumed that the actual number of new cases on these days was added into the new cases of the next day. Therefore, the number of new cases on days with zero new cases was assumed to be half of the number of new cases on the following day. For

Table 1. Unit root test (ADF test)

County	ADF test statistic	Optimal lag	P-value
Bronx	−2.015	9	0.2803
Kings	−1.865	9	0.3489
Nassau	−2.977	11	0.0378
NYC	−1.829	9	0.3663
Queens	−2.550	10	0.1043
Rockland	−3.435	9	0.0102
Westchester	−1.935	12	0.3160

Note: The equation for ADF test is:

$$\Delta y_t = \alpha + \gamma y_{t-1} + \sum_{k=1}^{L} \delta_k \Delta y_{t-k} + \varepsilon_t$$

where y_t is a time-series variable. α is constant and γ, $\delta_k (k = 1, \ldots, L)$ are the coefficients on the lag order of the autoregressive process. The null hypothesis for a unit root is $\gamma = 0$. Optimal lags of ADF test equation are determined by SIC. The maximum lag is 18. The *P*-values of ADF test results show that all variables cannot reject the null hypothesis at the 1% significant level. In cases of Nassau and Rockland the null hypothesis is rejected at the significant level of 5%. These variables might be stationary. However, even if VAR model contains both stationary variables and unit root variables,the Ordinary Least Squares (OLS) estimators of VAR model are consistent (Hamilton [11], Ch.18).

all days with zero new cases, we took half of the next day's value as the number of new cases.

To investigate the stationarity of the data, the augmented Dickey–Fuller (ADF) unit root test (see, in detail, Key Concept 14.8 in Stock and Watson [8].) was employed to the new case data of each county. We found that all of the level series had a unit root and were integrated into order one, denoted by I(1) (If y_{it} is nonstationary and the first difference of y_{it}, Δy_{it}, is stationary, then y_{it} is the integrated one process). Consequently, the path$\{\mathbf{y}_t\}$was concluded to be nonstationary as reported in Table 1.

Estimation

We used the popular econometric package EViews 12 from IHS Markit (In addition to EViews, Estima's RATS is a well-known econometric package specialized for time series analysis). First, we determined the lag order of the VAR model based on the VAR system information criteria, which were the Akaike information criterion (AIC), the Schwarz information criterion (SIC), and the Hanna–Quinn (HQ) information criterion. The formulae for calculating the AIC, SIC, and HQ are defined as (1) through (3) below:

$$\mathrm{AIC}\ (p) = -2\ (l/T) + \frac{2(n)^2 p}{T} \tag{1}$$

$$\mathrm{SIC}\ (p) = -2\ (l/T) + \frac{2(n)^2 p \log T}{T} \tag{2}$$

$$\mathrm{HQ}\ (p) = -2\ (l/T) + \frac{2(n)^2 p \log(\log T)}{T} \tag{3}$$

where n is the number of explanatory variables, p is the lag length, T is the sample size, and l is the value of the log of the system likelihood function with $(n)^2 p$ parameters estimated using T observations. The information criteria were calculated with a maximum lag length of 14. AIC is the most used criterion. However, because the sample size (T) was large (greater than 500), the AIC defined by equation (1) did not properly select the lag order. Thus, we applied the SIC or the HQ. The SIC recommended a lag length of 3, while HQ recommended a lag length of 8. The test results are reported in Table 2.

According to Alene *et al.* [7], the estimated average serial interval is 5.2 days (95% CI 4.9–5.5), which was estimated based on the data of individual infector–infectee pairs. However, the number of new COVID-19 cases was aggregated at the county level, and specific infector–infectee pairs were not able to be identified. Because the data were from online reports of confirmed cases, there was a confirmation lag between symptom onset and confirmation of a positive test result. Assuming that this average serial interval held at the county level, and that we could add the average confirmation lag of 3–4 days to the 95% CI of the above serial interval, we could thus regard the duration of infection (the infectious period) as 7.9–8.5 days. Based on this duration, a lag order of 8 was selected. Let denote it by VAR_Lag 8, henceforce. We established VAR_Lag 8 as the benchmark model for forecasting. The number of estimated coefficients was quite large which are not reported here.

Table 2. Lag-order test

Lag	AIC	SIC	HQ
0	86.349	86.408	86.372
1	81.484	81.953	81.668
2	80.941	81.821	81.286
3	80.392	81.682[a]	80.898
4	80.06	81.761	80.727
5	79.802	81.913	80.63
6	79.52	82.042	80.51
7	79.169	82.101	80.319
8	78.965	82.308	80.277[a]
9	78.845	82.6	80.318
10	78.683	82.847	80.316
11	78.543	83.117	80.337
12	78.377	83.362	80.333
13	78.236	83.632	80.353
14	78.132[a]	83.938	80.409

[a] Optimal lag order selected by each criterion.

All of the data had to be stationary for the VAR estimator to work. As we discussed in Data section, the new COVID-19 case data were nonstationary. Therefore, the VAR estimator did not meet consistency and would be biased. The standard way to solve this problem is to take the difference. However, Sims *et al.* [9] and Watson [10] proved the following useful proposition for large samples: regardless of whether the VAR contains an integrated component, the VAR has consistent ordinary least squares estimators in large samples. Because our sample size was large (greater than 500), the above proposition was held for our estimation. In other words, the standard VAR model could be directly applied to estimate the number of new COVID-19 cases by county. Therefore, there was no need to transform the model to a stationary form by differencing.

Forecasting

VAR forecasting

As described in "Estimation" section, the VAR estimators are consistent in the large sample. Therefore, we conducted VAR estimation to predict the number of new COVID-19 cases in each county. To make comparisons with the other forecasting models, we performed 4-week-ahead forecasting for three scenarios. The VAR_Lag 8 model was estimated based on a daily sample, and dynamic forecasting was performed for an out-of-sample period starting on the first forecast day.

A. 6_28 forecast: Estimate VAR_Lag 8 from 1 March 2020, through 27 June 2021, then conduct the 4-week-ahead forecast for 28 June 2021, through 24 July 2021.

B. 7_05 forecast: Estimate VAR_Lag 8 from 1 March 2020, through 4 July 2021, then conduct the 4-week-ahead forecast for July 5 through July 31.

C. 7_12 forecast: Estimate VAR_Lag 8 from 1 March 2020, through 11 July 2021, then conduct the 4-week-ahead forecast for July 12 through August 8.

Results

The root mean square error (RMSE) and the mean absolute percentage error (MAPE) for each of the above three scenarios are reported in Panels (a) and (b) of Table 3.

The RMSE and the MAPE are defined as (4) and (5) below:

$$\text{RMSE}: \sqrt{\sum\nolimits_{t=T+1}^{T+h} \left(\hat{y}_t - y_t\right)^2 / h} \tag{4}$$

$$\text{MAPE}: 100 \times \sum\nolimits_{t=T+1}^{T+h} \left|\frac{\hat{y}_t - y_t}{y_t}\right| / h \tag{5}$$

where $\hat{y}_t$ is a predicted value and y_t is the real value at time t.

Notably, the MAPE values for Rockland and Westchester were larger than the MAPE values for the other counties (Bronx, Kings, Nassau, NYC, and Queens) in all of the scenarios. The latter counties are classified as large central metro communities in the National Center for Health Statistics urban/rural CDC classification, while Rockland and Westchester are classified as large fringe metro communities. The number of new infections was lower in the fringe metro counties of Rockland and Westchester than in the central metro counties. Thus, a shock in the number of new infections is amplified in the fringe metro counties; because the VAR model is linear, it failed to capture such nonlinear shocks. In fact, regressions of the VAR using log-transformed data yielded better predictions for Rockland and Westchester as well. However, performance was poor for the central metro counties. Therefore, regressions of the VAR were performed only on level series.

Comparisons

As an example, for the Scenario (A) 6_28 forecast, the point forecasts at 4 specific days, July 3, July 10, July 17, and July 24, were compared with those of 11 other recently proposed forecast models listed in Table 4.

The point predictions of Scenarios (B) and (C) for these 4 days were also compared with the same four-point predictions of the

Table 3. 4-week-ahead-forcast errors

Panel (a): 4-weeks-ahead-forecast RMSE				Panel (b): 4-weeks-ahead-forecast MAPE			
County	**Scenario A**	**Scenario B**	**Scenario C**	**County**	**Scenario A**	**Scenario B**	**Scenario C**
Bronx	18	31	60	Bronx	26	23	29
Kings	42	92	159	Kings	31	29	34
Nassau	31	15	56	Nassau	34	13	29
NYC	43	72	117	NYC	38	34	39
Queens	34	34	74	Queens	24	17	22
Rockland	28	25	17	Rockland	230	145	69
Westchester	62	52	30	Westchester	200	102	35

Note: Scenario A: Estimate VAR (8) from 1 March 2020, through 27 June 2021, then conduct the 4-week-ahead forecast for 28 June 2021, through 24 July 2021.
Scenario B: Estimate VAR (8) from 1 March 2020, through 4 July 2021, then conduct the 4-week-ahead forecast for July through July 31.
Scenario C: Estimate VAR (8) from 1 March 2020, through 11 July 2021, then conduct the 4-week-ahead forecast for July 12 through August 8.

Table 4. Model descriptions

Model name		Methods	Classification
Var_Lag8 Model	Kitaoka and Takahashi	Vector Autoregression with eight lags	Time_series
CMU	Carnegie Mellon University	Autoregressive time series model	Time_series
Columbia	Columbia University	Meta-population SEIR model	Mechanistic method
Ensemble	University of Massachusetts, Amherst	Combination of 4 to 20 models depending on the availability of forecasts for each location	Mechanistic method
Facebook	Facebook AI research	A machine learning model with an autoregressive model	Deep_learning based
FRBSF-Wilson	Federal Reserve Bank of San Francisco/ Wilson	A SIR-derived econometric county panel data model with transmission rate assumed to be function of weather and mobility	Mechanistic method
JHU-APL	Johns Hopkins University, Applied Physics Lab	Meta-population SEIR model	Mechanistic method
JHU-UNC-Google	Johns Hopkins University, University of North Carolina, and Google	An ensemble of two different models: A multiplicative growth model and a curve-fitting model	Deep_learning based
LANL	Los Alamos National Laboratory	Statistical dynamical growth model accounting for population susceptibility	Mechanistic method
PandemicCentral	Pandemic Central	Random forest machine learning model	Deep_learning based
UGA-CEID	University of Georgia, Center for the Ecology of Infectious Disease	Statistical random walk model	Time_series
UVA	University of Virginia	An ensemble of three different models: An auto-regressive model, a machine learning (long short- memory) model, and a SEIR model	Deep_learning based

Note: The above models are cited by CDC. The details of model descriptions are: https://github.com/cdcepi/COVID-19-Forecasts/blob/master/COVID- 19_Forecast_Model_Descriptions.md.

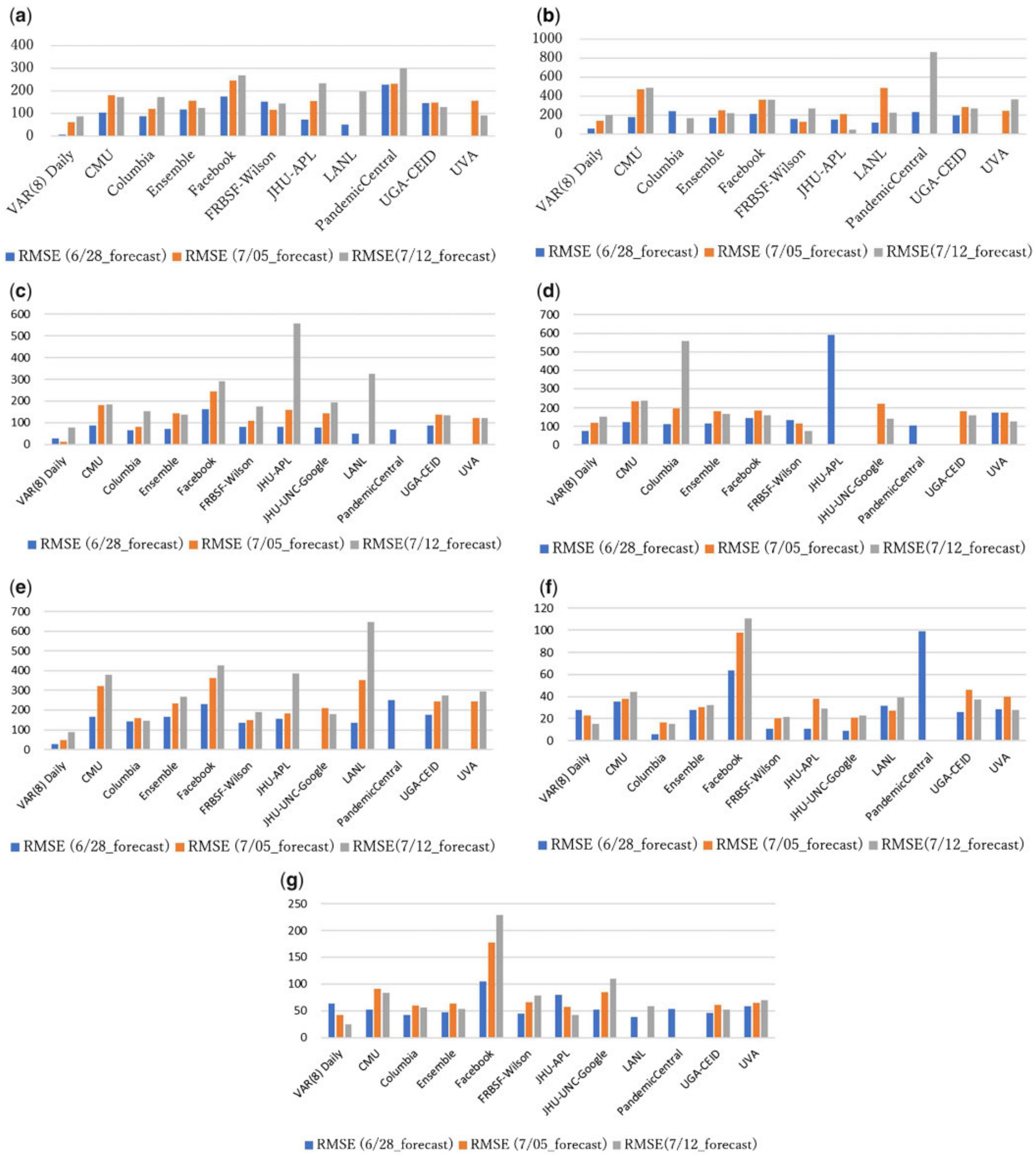

Figure 1. Forecast errors for seven New York counties. The horizontal axis is forecasting models and the vertical axis is RMSE. The forecast method of VAR_LAG 8 daily is dynamic forecast. The forecast values of other forecast models are extracted from "Previous COVID-19 Forcasts:Cases-2021 CDC": https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasting-us-cases-previous-2021.html.

other models. We used these four reported point estimates to make comparisons among models. The forecast for July 3 represents a 1-week-ahead forecast based on the data obtained up to June 28. Similarly, the forecast for July 10 represents a 2-week-ahead forecast based on data obtained up to June 28. The same interpretation applies to July 17 (a 3-week-ahead forecast) and July 24 (a 4-week-ahead forecast). The county-level forecasts for the 11 models were extracted from the following CDC files: 2021-06-28-all-forecasted-cases-model-data.cvs, 2021-07-05-all-forecasted-cases-model-data.cvs, and 2021-07-12-all-forecasted-cases-model-data.cvs(Downloadable from: Previous COVID-19 Forecasts: Cases | CDC). To compare the forecast accuracy between models, the RMSEs of the four-point forecasts are reported in Panel (a) through Panel (g) in Figure 1 (There are several measures for forecast accuracy such as RMSE, Mean Absolute Errors, MAPEs and Thile Inequality Coefficient. These evaluation measures are basically same type of measures. In fact, none of them changes our prediction results. Therefore, we used RMSE as a representative of predictive evaluation measures).

The results indicated that, compared with the other models, the VAR_Lag 8 model exhibited a much better forecasting performance for the 6_28, 7_05, and 7_12 forecasts for Bronx, Kings, Nassau, New York City, and Queens, but not for Rockland and Westchester. For the latter two counties, the forecasting results were still comparable with those of the other models. Although not reported here, the mean absolute error and the MAPE, which are other forecast error measures, also indicated similar results.

Concluding remarks

As we have discussed, the VAR prediction outperformed the predictions of other state-of-the-art models. The reason for this is that the VAR prediction adequately captures the pre-symptomatic and asymptomatic transmissibility of COVID-19 by introducing data from other counties as lagged explanatory variables.

In addition, we would like to point out three important simplifications of the VAR model.

1) PCR test bias, that is variation in the number of tests in days of week, was not considered here.
2) The number of cases infected with different COVID-19 variants was not taken into account. The number of infected cases was the sum of the mixed number of cases infected with different COVID-19 variants.
3) The handling of missing data was very simple, not the treatment based on any theories.

Despite these simplifications, the VAR model still provided good predictions. Based on the above, we strongly recommend the simple VAR model as a framework to accurately predict the regional transmission of COVID-19.

Supplementary data

Supplementary data is available at *Biology Methods and Protocols* online.

Data availability

The data that support the findings of this research are available in the Excel format in the supplementary files for this article.

Acknowledgments

We thank two anonymous reviewers for their helpful comments and Katherine Thieltges of Edanz (https://jp.edanz.com/ac) for editing the manuscript. H.T. would like to thank IMERA-Institute of Advanced Studies of Aix Marseille University, and Aix Marseille School of Economics for providing the best research environment for revising the paper.

Author contributions

Takayoshi Kitaoka (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing – original draft [Supporting], Writing – review and editing [equal]), and Harutaka Takahashi (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing – original draft [lead], Writing – review and editing [equal]).

Conflict of interest statement

None declared.

References

1. Wang L, Adiga A, Venkatramanan S *et al.* Examining deep learning models with multiple data sources for COVID-19 forecasting. 2020; 2010.14491 (arXiv.org).
2. Shang AC, Galow KE, Galow GG. Regional forecasting of COVID-19 caseload by non-parametric regression: a VAR epidemiological model. *AIMS Public Health* 2021;**8**:124–36.
3. Sims CA. Macroeconomics and Reality. *Econometrica* 1980;**48**:1–48.
4. Wang Q, Zhou Y, Chen X. A vector autoregression prediction model for COVID-19 outbreak. 2021; 2120.04843 (arXiv.org).
5. Nishiura H, Linton NM, Akhmetzhanov A. Serial interval of novel coronavirus (COVID-19) infections. *Int J Infect Dis* 2020;**93**:284–6.
6. He X, Lau EHY, Wu P *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* 2020;**26**:672–5.
7. Alene M, Yismaw L, Assemie MA *et al.* Serial interval and incubation period of COVID-19: a systematic review and meta-analysis. *BMS Infect Dis* 2021;**21**:257.
8. Stock JH, Watson MW. *Introduction to Econometrics*. 2nd edn. Boston MA: Pearson Education, 2007, 795.
9. Sims CA, Stock JH, Watson MW. Inference in linear time series model with some Unit Roots. *Econometrica* 1990;**58**:113–44.
10. Watson MW. Vector autoregressions and cointegration. In: Engle RF, McFadden DL (eds), *Handbook of Econometrics Vol.IV*, Vol. **47**. Amsterdam, Netherlands: Elsevier, 1944, 2743–841.
11. Hamilton JD. *Time Series Analysis*. Princeton, NJ: Princeton University Press, 1994, 799.