


Brief Report

Cite this article: Cruz-Cano R, Ma T, Yu Y, Lee M, Liu H. Forecasting COVID-19 cases based on social distancing in Maryland, USA: A time – series approach. *Disaster Med Public Health Prep*. doi: <https://doi.org/10.1017/dmp.2021.153>.

Keywords:
environmental exposure; public health; vital statistics

Corresponding author:
Raul Cruz-Cano,
Email raulcruz@umd.edu.

Forecasting COVID-19 Cases Based on Social Distancing in Maryland, USA: A Time–Series Approach

Raul Cruz-Cano PhD¹ , Tianzhou Ma PhD¹, Yifan Yu MS¹, Minha Lee MS² and Hongjie Liu PhD¹

¹Department of Epidemiology and Biostatistics, University of Maryland, College Park, Maryland USA and ²Maryland Institute of Transportation, University of Maryland, College Park, Maryland USA

Abstract

Objective: Our objective is to forecast the number of coronavirus disease 2019 (COVID-19) cases in the state of Maryland, United States, using transfer function time series (TS) models based on a Social Distancing Index (SDI) and determine how their parameters relate to the pandemic mechanics.

Methods: A moving window of 2 mo was used to train the transfer function TS model that was then tested on the next week data. After accounting for a secular trend and weekly cycle of the SDI, a high correlation was documented between it and the daily caseload 9 days later. Similar patterns were also observed on the daily COVID-19 cases and incorporated in our models.

Results: In most cases, the proposed models provide a reasonable performance that was, on average, moderately better than that delivered by TS models based only on previous observations. The model coefficients associated with the SDI were statistically significant for most of the training/test sets.

Conclusions: Our proposed models that incorporate SDI can forecast the number of COVID-19 cases in a region. Their parameters have real-life interpretations and, hence, can help understand the inner workings of the epidemic. The methods detailed here can help local health governments and other agencies adjust their response to the epidemic.

The rapid, global spread of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused hundreds of thousands of deaths. Although social distancing is considered a key measure to reduce the spread of the virus,¹ the exact impact of day-to-day social distancing on viral spread remains unclear.

Since the report of the first confirmed case of coronavirus disease 2019 (COVID-19) in Maryland on March 5, 2020,² more than 334,000 cases and 6,700 deaths have been reported.³ On March 16, 2020, the state government implemented restrictions on gatherings and closure of educational facilities and on March 30, 2020, a *stay at home* order was imposed. The Maryland Transportation Institute (MTI) implemented the Social Distancing Index (SDI) to measure the extent residents and visitors are practicing social distancing.³ To date, studies have evaluated the efficacy of social distancing strategies to reduce the magnitude of the epidemic,⁴ but models that accurately forecast daily caseloads based on social mobility patterns are yet to be explored.

In this article, we proposed to use a sequence of time series (TS) models to forecast and further understand the relationship between social distancing and the COVID-19 daily caseload. Our analysis of transfer function TS models can accomplish this objective because they present a dependable way to analyze data in which the current value of variable, eg, daily COVID-19 cases in Maryland depends on its previous values and those of other predictors, such as SDI. Other TS models have been used to accurately analyze this and previous pandemics.⁵ However, to our knowledge, this is the first attempt to develop a TS model that includes a social distance measure to predict daily COVID-19 cases. The magnitude of the COVID-19 epidemic makes it worthwhile to keep exploring any possible way to improve the models that can predict its behavior.

Methods

Data used in this study covered the timeframe from March 5, the day the first COVID-19 case was reported in Maryland, to June 1. On May 26, demonstrations against police violence started in Minnesota and spread all over the country in the next days. There exists anecdotal evidence that, due to use of preventive measures, such as face masks, these massive gatherings changed the relationship between social distancing and COVID-19 cases⁶; hence, we decided not to include information on or beyond this date in the creation of the models.

The variables included in our study are:

- *Primary exposure variable of interest: Daily Social Distancing Index (SDI)*³: SDI is an integer from 0 (no social distancing) to 100 (all residents are staying at home and no visitors are entering Maryland). The social distancing index is computed from 6 mobility metrics by this equation: $SDI = 0.8 * [\% \text{ staying home} + 0.01 * (100 - \% \text{ staying home}) * (0.1 * \% \text{ reduction of all trips compared with pre-COVID-19 benchmark} + 0.2 * \% \text{ reduction of work trips} + 0.4 * \% \text{ reduction of nonwork trips} + 0.3 * \% \text{ reduction of travel distance})] + 0.2 * \% \text{ reduction of out-of-county trips}$.
- The weights are chosen based on share of residents and visitor trips; what trips are considered more essential, and the principle that higher SDI scores should correspond to fewer chances for close-distance human interactions and virus transmissions.
- *Outcome variable: Daily COVID-19 cases*: Number of new daily COVID-19 cases according to the New York Times.²

In TS, a general transfer function can be used to describe the relationship between an input and an output series.⁷ We propose to use a transfer function TS model with the following 3-step procedure to relate the input daily SDI to the output COVID-19 cases series⁷ while accounting for the secular trend and weekly cycles of the exposure and outcome variables:

- *Step 1: Fit an autoregressive integrated moving average (ARIMA) model to the independent variable SDI*. This step helps to find patterns that need to be removed from the independent variables before we study its relationship with the outcome. An ARIMA model is defined by its parameters (p, d, q) where p represents the order of autoregression, and q the order of the moving average. The parameter d is the degree of difference, eg, $d = 1$ means that, instead of the original SDI series, we would use $(SDI_t - SDI_{t-1})$. We used plots of autocorrelation function (ACF) and partial ACF (PACF) to determine the values of these parameters, which is well-known procedure applied in many published works.^{5,7}
- *Step 2: Remove the patterns from the input series SDI and compute the cross-correlation with the daily and imported COVID-19 cases*. This step helps to determine the pure delay in the system s after removing the SDI patterns discovered in the previous step (pre-whiten). We defined s as the largest cross-correlation, which occurred at delay ≥ 0 days.
- *Step 3: Compute transfer function and fit it with noise model*. The study of the cross-correlation graph of the pre-whiten input and outcome series can help to determine the terms that are need in the numerator and denominator of the transfer function model. This is a complex matter and beyond the scope of this work, but in general a simpler model is recommendable.⁷ The analysis of residuals produced by the transfer function TS model might indicate the need to include more terms to improve its fit.⁷

In our particular case of study, the pre-whiten SDI series/Daily COVID-19 cases cross-correlation graph presented nonzero cross-correlations with some of them decaying exponentially, indicating the need to add terms in the numerator and denominator of the transfer function model.⁷ A parsimonious model was initially chosen for the transfer function⁷:

$$Cases_t = \theta_0 + (C(1 - \theta_1 B) / (1 - \delta B)) \nabla SDI_{t-s} + Noise$$

where B is the backward shift operator and ∇ is the differencing operation.

We use a sliding window of 60 day to train the model and estimate the parameters C , θ_0 , θ_1 , and δ in the transfer function, and test the forecasting results in the week immediately following. For the test week, estimated values of SDI were used as the input in the transfer function to forecast the daily cases, ensuring that information unavailable in a real-life case scenario would not be used to produce the results for the test week. Given the time limit discussed previously in this section, the first training window goes from March 5 to May 4 with a test week that encompasses May 5 to May 12. The last training window goes from March 26 to May 25 with a test window that goes from May 26 to June 3. The performance of the procedure for the test week was assessed using the mean absolute percentage error (MAPE) defined as $MAPE = \frac{100\%}{7} \sum_{i=1}^7 \left| \frac{cases_i - \widehat{cases}_i}{cases_i} \right|$. All analysis is performed using SAS software 9.4.

Results

Based on the full data, we observed that the autocorrelations depicted in the ACF plot decayed slowly, indicating the need to differentiate this series. The ACF of the differentiated SDI series tailed off at lags $7k$ ($k = 1, 2, \dots$), while PACF cutoff at lag 7. These observations lead to a simple time-series model composed of a differencing term ($d = 1$) and autoregressive (AR) term $p = (1, 7)$ that was fitted for the SDI data. The ACF and PACF plots after removing this AR term from the differentiated series showed no significant auto-correlations, indicating that this ARIMA model adequately fitted the SDI data and, hence, could be used to pre-whiten it. The analysis of the cross-correlation of the pre-whiten SDI with the daily cases led to a delay $s = 9$ d (cross-correlation = -0.23). Examination of residuals from the initial transfer function model in forecasting COVID-19 case indicated that an AR term $p = (1, 7)$ of the daily COVID-19 cases was also needed in the final transfer function TS model.

After adding this autoregressive term to the transfer function model no more significant correlations appeared in the ACF and PACF, indicating that the resulting model adequately captured the relationship between SDI and the number of daily COVID-19 cases in Maryland in this time period. We then proceeded to estimate the coefficients of this model for each of the training windows described above. To benchmark the performance, we compared our proposed transfer function model with a simple ARIMA model for daily cases forecasting (ie, excluded the SDI) with $d = 1$ and autoregressive term $p = (1, 7)$.

Notice at least 1 of the parameter estimates associated with the SDI (C , θ_1 , or δ) are statistically significant with a P -value < 0.05 (Table 1), except in the first week and the week from March 23 to May 22. Hence, the results in Table 1 indicate that including the SDI as in input variable with the appropriate delay in the models can be an important predictor of daily COVID-19 cases. Focusing on the statistically significant results, the positive values for the autoregressive terms indicate that past values of cases lead to a larger number of infections, while the negative values for C paired with the positive estimates for the δ show that larger SDI leads to fewer cases with a 9-d delay.

The MAPE values for test weeks that start up until May 24 vary between 12.2% and 20.1% as seen in Table 2 hence fall or are around what would be considered *good forecasts*.⁸ The unevenness of the performance might be attributed to external factors that influence the number of reported cases in a given day during those earlier days of the pandemic, eg, local weather and scarce availability of tests. The MAPE of the simple ARIMA models that exclude the SDI and rely exclusively on the secular trend and weekly cycle

Table 1. T-statistic (*p*-values) for the TS transfer function parameters

Model building period	Autoregressive term Lag $p=1$	Autoregressive term Lag $p=7$	C Lag $p=0, s=9$	θ_1 Lag $p=1, s=9$	δ Lag $p=1, s=9$
5-March/4-May	4.3 ($<.0001$)	2.4 (0.0165)	-1.3 (0.1810)	-1.6 (0.1103)	-0.9 (0.3468)
6-March/5-May	-0.1 (0.9490)	1.1 (0.2581)	-2.3 (0.0231)	1.4 (0.1483)	54.5 ($<.0001$)
7-March/6-May	-0.2 (0.8570)	0.8 (0.4070)	-2.8 (0.0046)	0.9 (0.3806)	46.5 ($<.0001$)
8-March/7-May	4.3 ($<.0001$)	2.2 (0.0248)	-1.7 (0.0867)	-2.1 (0.0392)	-1.2 (0.2213)
9-March/8-May	3.7 (0.0002)	1.6 (0.1033)	-3.0 (0.0030)	-1.0 (0.3373)	8.5 ($<.0001$)
10-March/9-May	4.6 ($<.0001$)	2.2 (0.0249)	-1.8 (0.0755)	-2.6 (0.0094)	-1.7 (0.0989)
11-March/10-May	2.2 (0.0295)	2.0 (0.0471)	-3.4 (0.0006)	-1.4 (0.1654)	9.9 ($<.0001$)
12-March/11-May	0.7 (0.4576)	1.2 (0.2193)	-5.1 ($<.0001$)	-1.6 (0.1019)	22.3 ($<.0001$)
13-March/12-May	0.0 (0.9809)	1.2 (0.2195)	-2.3 (0.0194)	3.3 (0.0009)	48.9 ($<.0001$)
14-March/13-May	0.5 0.6499	0.9 (0.3606)	-2.4 (0.0151)	1.1 (0.2674)	33.8 ($<.0001$)
15-March/14-May	4.1 ($<.0001$)	2.1 (0.0360)	-2.2 (0.0303)	0.8 (0.4354)	0.3 (0.7522)
16-March/15-May	0.3 (0.7878)	0.8 (0.3966)	-3.0 (0.0028)	0.1 (0.9389)	31.3 ($<.0001$)
17-March/16-May	1.1 (0.2703)	0.6 (0.5204)	-4.2 ($<.0001$)	-1.4 (0.1623)	25.7 ($<.0001$)
18-March/17-May	0.8 (0.4401)	0.6 (0.5744)	-4.2 ($<.0001$)	-1.3 (0.1941)	24.9 ($<.0001$)
19-March/18-May	2.9 (0.0037)	1.2 (0.2119)	-2.7 (0.0074)	-1.0 (0.2963)	2.0 (0.04)
20-March/19-May	0.4 (0.6912)	0.46 (0.6450)	-1.57 (0.1163)	2.09 (0.0362)	33.79 ($<.0001$)
21-March/20-May	-0.21 (0.8367)	0.65 (0.5167)	-1.85 (0.0646)	0.53 (0.5962)	27.90 ($<.0001$)
22-March/21-May	2.54 (0.0112)	2.76 (0.0058)	-1.83 (0.0669)	-6.54 ($<.0001$)	-15.33 ($<.0001$)
23-March/22-May	1.56 (0.1190)	1.81 (0.0704)	-1.74 (0.0823)	-0.13 (0.8995v)	0.80 (0.4213)
24-March/23-May	-0.49 (0.6260)	0.39 (0.6941)	-2.09 (0.0364)	-0.00 (0.9977)	26.42 ($<.0001$)
25-March/24-May	1.45 (0.1461)	2.97 (0.0030)	-2.15 (0.0314)	-12.46 ($<.0001$)	-8.94 ($<.0001$)
26-March/25-May	0.47 (0.6392)	0.85 (0.3962)	-2.69 (0.0072)	-0.72 (0.4718)	3.28 (0.0011)

of number of COVID-19 cases in this period of time showed even more volatility varying between 9.5% and 37.8% and is on average 1.94% worse than those of the proposed transfer function models. The performance of both the transfer function and classic ARIMA models for the week composed completely of days starting on May 25 lead to MAPE of 41.8% and 42.2%, respectively, hinting that this point in time might mark a significant departure on how social distancing relates to COVID-19 cases.

Discussion

The required AR terms $p = (1,7)$ suggest that the SDI and daily count variables are influenced by their observed or estimated values on the previous day (secular trend) and a week prior (weekly

cycle) supporting what has been observed in previous studies,⁹ while the delay of 9 d for the COVID-19 cases is within the range of the number of days that symptoms take to appear,¹⁰ indicating not only that our proposed transfer function models provide an adequate prediction performance but also their characteristics of correspond to patterns seen in the pandemic. The degradation of the models performance after the start of the 2020 national protests attest to the limitation of the current version of the models and point toward the need to redo the process to obtain the optimal values for the (p,d,q) parameters and the delay s instead of just recalculating the coefficients of the models. The study of the P -values seen in Table 1 help to reinforce conclusions that have been drawn about the COVID-19 pandemic previously, namely that having a large number of cases in a community leads to even more

Table 2. Forecasting daily COVID-19 cases 1 wk in advance

Week	Model building period	Test week	MAPE transfer function TS with SDI	MAPE simple ARIMA
1	5-March/4-May	5-May/12-May	12.2	15.5
2	6-March/5-May	6-May/13-May	15.4	17.5
3	7-March/6-May	7-May/14-May	17.2	16.2
4	8-March/7-May	8-May/15-May	15.6	18.3
5	9-March/8-May	9-May/16-May	12.2	11.8
6	10-March/9-May	10-May/17-May	13.0	12.0
7	11-March/10-May	11-May/18-May	13.9	15.0
8	12-March/11-May	12-May/19-May	13.5	9.5
9	13-March/12-May	13-May/20-May	20.1	9.8
10	14-March/13-May	14-May/21-May	17.4	16.8
11	15-March/14-May	15-May/22-May	20.1	17.0
12	16-March/15-May	16-May/23-May	19.4	17.5
13	17-March/16-May	17-May/24-May	19.5	18.7
14	18-March/17-May	18-May/25-May	19.6	23.5
15	19-March/18-May	19-May/26-May	19.6	19.7
16	20-March/19-May	20-May/27-May	28.3	37.8
17	21-March/20-May	21-May/28-May	26.9	33.6
18	22-March/21-May	22-May/29-May	32.7	35.4
19	23-March/22-May	23-May/30-May	29.4	36.9
20	24-March/23-May	24-May/1-June	24.6	36.2
21	25-March/24-May	25-May/2-June	41.8	42.2
22	26-March/25-May	26-May/3-June	27.5	41.9

infections and that the decrease of social distancing behavior among the members of a group are associated with an increase in the number of positive cases few days later. A limitation of this study is its single focus on the SDI; future work might include evaluation if the conclusions reach in this manuscript hold true for other measure of social distancing, such as Unacast Social Distancing Scorecard.

Although the models described in this report were optimized for the epidemic in Maryland, the steps described here can be used to develop models to forecast the number of COVID-19 cases in a other regions several days in advance. Parameters used in this transfer function model will change according to region and time because of modifications to social distancing regulations and other factors (eg, contact tracing) but the transfer functions can include other independent variables in addition to SDI, hence providing useful information in the debate of economy resume vs pandemic control for this and future pandemics.

Funding statement. Research reported in this publication was partially supported by the Department Research Discretionary Funds.

References

1. Lewnard JA, Lo NC. Scientific and ethical basis for social-distancing interventions against COVID-19. *Lancet Infect Dis.* 2020;20(6):631-633. doi: [10.1016/S1473-3099\(20\)30190-0](https://doi.org/10.1016/S1473-3099(20)30190-0)
2. Smith M, Yourish K, Almkhatar S, *et al.* An ongoing repository of data on coronavirus cases and deaths in the U.S. New York Times. <https://github.com/nytimes/covid-19-data>. Accessed May 18, 2021.
3. Zhang L, Ghader S, Pack M, *et al.* An interactive COVID-19 mobility impact and social distancing analysis platform. *medRxiv.* 2020. doi: [10.1101/2020.04.29.20085472](https://doi.org/10.1101/2020.04.29.20085472)
4. Castillo R, Staguhn E, Weston-Farber E. The effect of state-level stay-at-home orders on COVID-19 infection rates. *Am J Infect Control.* 2020;48(8):958-960. doi: [10.1016/j.ajic.2020.05.017](https://doi.org/10.1016/j.ajic.2020.05.017)
5. Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. *Sci Total Environ.* 2020;729:138817. doi: [10.1016/j.scitotenv.2020.138817](https://doi.org/10.1016/j.scitotenv.2020.138817)
6. Janes C. Protests probably didn't lead to coronavirus spikes, but it's hard to know for sure. Washington Post. https://www.washingtonpost.com/health/protests-probably-didnt-lead-to-coronavirus-spikes-but-its-hard-to-know-for-sure/2020/06/30/d8179678-baf5-11ea-8cf5-9c1b8d7f84c6_story.html. Published June 30, 2020.
7. Brocklebank JC, Dickey DA, Choi BS. *SAS for Forecasting Time Series*. 3rd ed. Cary, NC: SAS Institute; 2018.
8. Lewis CD. *Industrial and Business Forecasting Methods*. London: Butterworth Scientific; 1982.
9. Ricon-Becker I, Tarrasch R, Blinder P, *et al.* A seven-day cycle in COVID-19 infection and mortality rates: are inter-generational social interactions on the weekends killing susceptible people? *medRxiv.* January 2020:2020.05.03.20089508. doi: [10.1101/2020.05.03.20089508](https://doi.org/10.1101/2020.05.03.20089508)
10. Li Q, Guan X, Wu P, *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N Engl J Med.* 2020; 382(13):1199-1207. doi: [10.1056/NEJMoa2001316](https://doi.org/10.1056/NEJMoa2001316)