ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2020-75402

A Comparison of Time-Series Models in Predicting COVID-19 Cases

COVID-19 Vakalarının Tahmin Edilmesinde Zaman-Serisi Modellerinin Bir Karşılaştırması

Mehmet KOÇAK[a,b]

[a]The University of Tennessee Health Science Center, Department of Preventive Medicine, Division of Biostatistics, Memphis, Tennessee, USA
[b]İstanbul Medipol University, Regenerative and Restorative Research Center, İstanbul, TURKEY

ABSTRACT Objective: As the world is striving to control the COVID-19 pandemic, one aspect of this strife is to project how many new cases will be experienced in the coming days so that health services can better be planned and real time health policies can be developed depending on the spread of the pandemic, its speed and direction. To address this, we compared time-series modeling approaches as to which one more accurately projects how many new cases to expect within 5 days. **Material and Methods:** In this research, we used the accumulating COVID-19 cases for all countries since the beginning of the pandemic in China in December 31, 2019, and aimed at identifying best time-series model to project COVID-19 cases and deaths. **Results:** We showed that Conditional Lest Square modeling with AR(1) auto-correlation structure should be the model to be chosen for case projections. For death projections, Conditional Lest Square modeling with AR(2) auto-correlation structure showed slightly better performance than its compatibles, and should be the model of choice. We also observed that the observed level of confidence interval is lower than its expected level. **Conclusion:** Future cases and dates due to COVID-19 can be projected successfully with time-series models with Conditional Lest Square modeling using AR(1) auto-correlation structure for cases and AR(2) auto-correlation structure deaths.

Keywords: COVID-19; future case prediction; pandemic rate; time series modeling

ÖZET Amaç: Dünya COVID-19 pandemisini kontrol almayla mücadele ederken, bu mücadelenin bir boyutu da, pandeminin yayılması, hızı ve yönüne göre sağlık hizmetlerini daha iyi planlamak ve gerçek zamanlı sağlık politikaları geliştirebilmek için, gelecek günlerde kaç yeni vakanın tecrübe edileceğini tahmin etmektir. Buna bir cevap olarak, zaman-serisi modelleme yaklaşımlarını, hangisinin 5 gün içinde kaç tane yeni vaka olacağını en doğru olarak tahmin edeceği açısından karşılaştırdık. **Gereç ve Yöntemler:** Bu araştırmada, 31 Aralık 2019 tarihinde Çin'de başlayan pandemiden bu yana, tüm ülkelerin biriken COVID-19 vakalarını kullandık, ve COVID-19 vakalarını ve ölümlerini en iyi tahmin edecek zaman-serisi modelini tanımlamayı hedefledik. **Bulgular:** Vaka tahminlerinde AR(1) oto-regresyon yapısını kullanan Şartlı En-Küçük Kareler modelinin en tercih edilen model olması gerektiğini gösterdik. Ölüm projeksiyonu için de, AR(1) oto-regresyon yapısını kullanan Şartlı En-Küçük Kareler modeli, rakiplerinden biraz daha iyi performans gösterdik ve bu yüzden tercih edilecek model olmalı. Aynı zamanda, gözlenen güven aralığı seviyesinin, beklenenden daha düşük olduğunu gözledik. **Sonuç:** Gelecekteki COVID-19'a bağlı vaka ve ölümler, sırasıyla AR(1) ve AR(2) oto-regresyon yapısını kullanan Şartlı En-Küçük Kareler zaman-serisi modelleriyle başarılı bir şekilde tahmin edilebilir.

Anahtar kelimeler: COVID-19; gelecek vaka tahmini; salgın hızı; zaman serileri modelleme

The world is going through historical times since December 31, 2019, when the first cases of new Coronavirus (COVID-19) are officially reported from China. Human Coronavirus (HCoV) is not new to the epidemiology world as it was first reported in 1960s.[1] Later strains involving serious respiratory tract infections were reported with various names as SARS-CoV in 2003, as HCoV NL63 in 2004, as HKU1 in 2005, and MERS-CoV in 2012.[2]

Correspondence: Mehmet KOÇAK
The University of Tennessee Health Science Center, Department of Preventive Medicine, Division of Biostatistics, Memphis, Tennessee, USA
E-mail: mkocak1@uthsc.edu

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 09 Apr 2020 ***Accepted:*** 10 Apr 2020 ***Available online:*** 30 Apr 2020

While the international community was aware of this particular virus and its serious potential epidemic potential, its latest version SARSCoV-2 (2019) was portrayed as a typical seasonal influenza and its potential to be a full-blown epidemic than to be a fast spreading pandemic was initially downplayed. The world soon realized its differences as first an epidemic mainly in Wuhan, China and its neighboring regions in South Asia;[3] it then was considered to be a pandemic by the World Health Organization (WHO) on March 11, 2020. Although the epicenter of the COVID-19 pandemic was Wuhan, China, it quickly moved to South Korea, Iran, then Italy, and through Italy, to the rest of the Europe. At the time of the writing of this manuscript, the main epicenters of the pandemic included the United States, Italy, Spain, France, Germany, United Kingdom, Iran, and Turkey.

In this paper, we compare multiple time-series modelling approaches to determine which one provides more accurate projections of new COVID-19 cases and in the coming days. This is a significant need for the governments to increase the healthcare readiness for the new patients and develop strategies to contain and eliminate the pandemic.

MATERIAL AND METHODS

We obtained the COVID-19 data on April 07, 2020, from one of the main COVID-19 data repositories, https://ourworldindata.org/coronavirus-source-data, which is updated every day.

We first illustrate a typical time-series projections. We modeled the COVID-19 cases until April 07, 2020 and projected the expected new cases for the following 10 days. We present the results in Table 1 and Figure 1 below. With these projections, we expect that the number of new COVID-19 cases in Turkey will be anywhere from 2494 to 3492, which may go up as high as beyond 5000 or may go down to lower 1600.

We have the following analysis design:

- **Models:** Maximum Likelihood (ML), Conditional Least Squares (CLS), and Unconditional Least Squares (ULS)[4,5,6]
- **Autoregression structure:** AR(1), AR(2), AR(3)[7]
- **Outcomes:** COVID-19 Cases, COVID-19 Deaths
- **The lengths of time series:** 10 to 30 days of time series data with increment of 2 days
- **Total COVID-19 case restrictions:**
  - To project the new COVID-19 cases, we utilized all accumulated data from all countries which had at least a total of 100 cases as of April 7, 2020;

TABLE 1: Projections for Covid-10 cases for Turkey for 10-days from April 7, 2020.

Projection Date	Projected No. of Cases	%95 Confidence Interval Lower Bound	%95 Confidence Interval Upper Bound
2020-04-08	4031	3463	4598
2020-04-09	4180	3385	4975
2020-04-10	4329	3358	5301
2020-04-11	4479	3359	5599
2020-04-12	4628	3377	5879
2020-04-13	4778	3408	6147
2020-04-14	4927	3448	6405
2020-04-15	5076	3496	6656
2020-04-16	5226	3550	6901
2020-04-17	5375	3609	7141

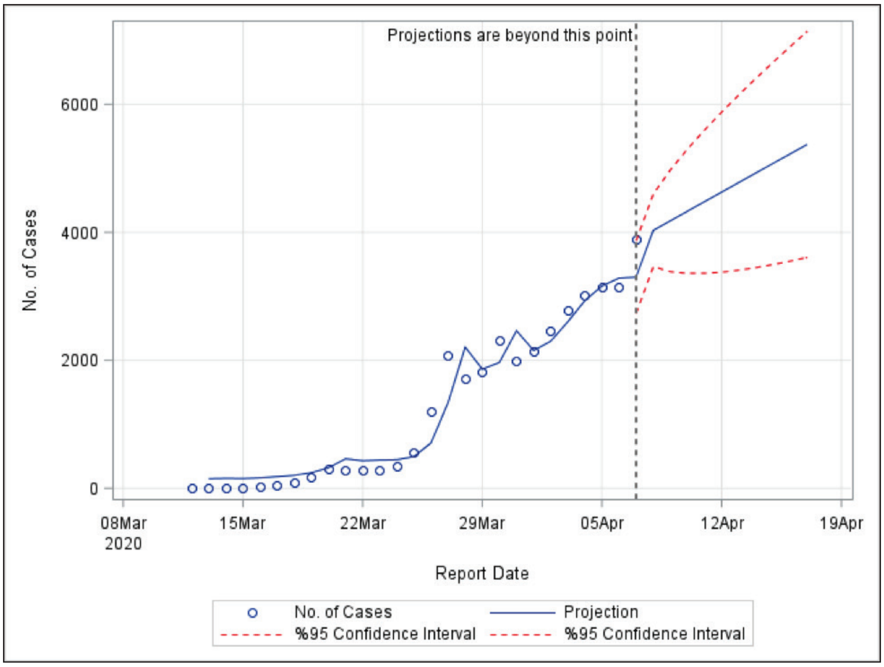

FIGURE 1: Projections for Covid-10 cases for Turkey for 10-days from April 7, 2020.

- To project the new COVID-19 deaths, we utilized all accumulated data from all countries which had at least a total of 100 deaths as of April 7, 2020.

Under each design scenario, we obtained the projected COVID-19 cases or deaths, respectively, and calculated the following diagnostic measures:

1- An indicator variable to show whether or not the actual COVID-19 cases on the coming days fall within the estimated confidence bound from the corresponding model

2- Absolute difference between the projected COVID-19 cases and actual COVID-19 cases

We then compared different modelling approaches in terms of the above comparative measures with respect to time-series length for both COVID-19 cases and deaths.

All computations in this research were conducted on SAS® Version 9.4.[8]

As we are utilizing publicly available COVID-19 summary data which does not include any human subject data, no Institutional Review Board review is needed for our research. We have conducted this research according to the principles of Helsinki Declaration.

RESULTS

We present the comparative results by the length of time-series data, autocorrelation and estimation approaches in Table 2 and Figure 2 below for the obtained level of confidence and absolute mean residuals.

We conclude from these results that AR(1) model performs more favorably in terms of actual confidence coverage for the future projections. We also present an overall comparison across the projection days (Table 3), which suggests that although AR(3) model seems to have slightly higher confidence coverage, it loses its advantage with higher residuals; therefore, AR(1) still seems to be an modeling approach of choice for such projections.

As seen in Table 2, Table 3, and Figure 3, we conclude that overall, AR(1) model seems to be performing more favorably compared to AR(2) and AR(3) although they all perform compatibly for Future Day-1.

Among the AR(1) models, CLS and ML approaches perform similarly and slightly better than ULS (Figure 3, Table 2, Table 3). As the overall mean absolute residual is smaller for the CLS model, we suggest that it will be the model of choice in projecting the future COVID-19 cases.

TABLE 2: Observed level of confidence and mean absolute residuals in time-series projections for future day-1. (The best result under each combination is bolded and highlighted with gray)

		AR(1)			AR(2)			AR(3)		
		CLS	**ML**	**ULS**	**CLS**	**ML**	**ULS**	**CLS**	**ML**	**ULS**
% of Projections within 95% Confidence	10 Days of Data	**72.0%**	67.7%	68.7%	70.5%	**72.8%**	65.3%	**72.0%**	71.7%	69.1%
	12 Days of Data	**70.2%**	68.2%	67.3%	67.0%	**67.6%**	63.8%	70.6%	**73.8%**	68.7%
	14 Days of Data	**66.4%**	65.1%	65.1%	**69.6%**	68.8%	64.8%	**67.0%**	64.8%	62.1%
	16 Days of Data	**71.4%**	71.2%	67.6%	**70.5%**	69.4%	69.9%	**76.6%**	75.5%	72.8%
	18 Days of Data	74.5%	72.1%	73.8%	74.5%	**75.0%**	74.5%	**71.8%**	71.6%	69.3%
	20 Days of Data	74.3%	**75.2%**	72.7%	**71.4%**	**71.4%**	67.3%	71.8%	**73.5%**	69.7%
	22 Days of Data	74.5%	**74.5%**	74.2%	**71.1%**	70.7%	69.9%	70.1%	**72.6%**	68.8%
	24 Days of Data	60.9%	**62.4%**	60.9%	**68.5%**	65.6%	67.8%	**76.1%**	75.8%	75.6%
	26 Days of Data	73.3%	**74.4%**	73.3%	72.1%	**73.5%**	73.2%	**70.6%**	69.8%	69.5%
	28 Days of Data	69.3%	69.3%	**70.3%**	69.3%	**69.4%**	68.5%	69.3%	**71.6%**	68.5%
	30 Days of Data	**61.2%**	**61.2%**	**61.2%**	62.7%	**65.7%**	62.7%	68.7%	**69.7%**	67.2%
Mean Absolute Residual	10 Days of Data	**17.1**	19.0	18.5	33.7	**31.4**	36.4	**29.2**	30.0	31.8
	12 Days of Data	**16.8**	17.6	17.9	27.3	**26.2**	30.5	29.1	**23.8**	31.1
	14 Days of Data	**29.7**	31.3	31.6	**27.6**	28.0	29.8	34.4	**34.3**	36.6
	16 Days of Data	**24.1**	24.5	25.8	32.2	28.5	**26.2**	**21.0**	**21.1**	23.1
	18 Days of Data	**30.1**	31.6	31.6	**29.1**	29.6	30.2	**35.8**	36.2	37.0
	20 Days of Data	29.9	**29.1**	31.7	36.8	**35.4**	39.5	**34.0**	35.0	36.0
	22 Days of Data	**25.2**	**25.2**	26.5	**33.0**	34.2	34.1	**39.0**	**39.0**	42.1
	24 Days of Data	**39.0**	48.2	50.6	42.3	43.3	**41.0**	**36.3**	37.0	37.7
	26 Days of Data	49.1	**48.3**	49.1	**48.9**	50.4	51.1	61.0	**57.9**	64.4
	28 Days of Data	48.7	48.8	**48.4**	59.4	**56.8**	61.3	**63.7**	65.7	65.0
	30 Days of Data	**75.0**	75.1	75.1	88.0	**87.8**	**87.8**	89.0	92.3	**88.7**

TABLE 3: Overall observed level of confidence and mean absolute residuals in time-series projections across all future day projections from day-1 through day-5.

		AR(1)			AR(2)			AR(3)		
		CLS	**ML**	**ULS**	**CLS**	**ML**	**ULS**	**CLS**	**ML**	**ULS**
Future Day-1	% of Projections within 95% CI	**69.8%**	69.2%	68.6%	69.8%	**70.0%**	68.0%	71.3%	**71.9%**	69.2%
	Mean Absolute Residual	4.6	3.8	**3.6**	**5.4**	5.5	5.8	**9.9**	10.3	10.1
Across Future Day-1 to Day-5	% of Projections within 95% CI	**70.4%**	70.3%	69.7%	66.9%	**67.4%**	65.5%	65.7%	**66.0%**	63.5%
	Mean Absolute Residual	**43.3**	44.6	45.9	47.9	**47.0**	48.5	47.1	**46.2**	49.1

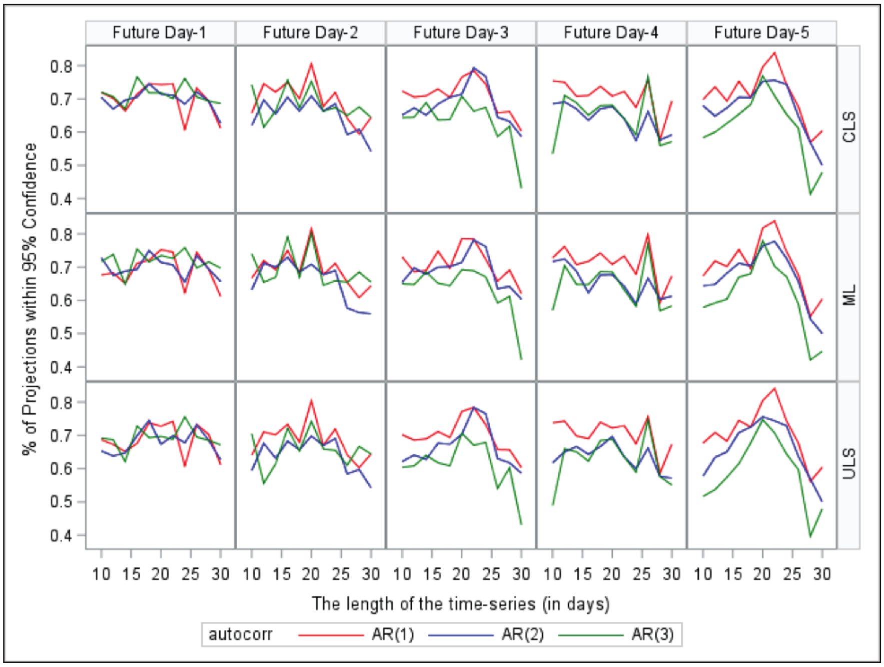

FIGURE 2: Observed level of confidence in time-series case projections by auto-regression structure.

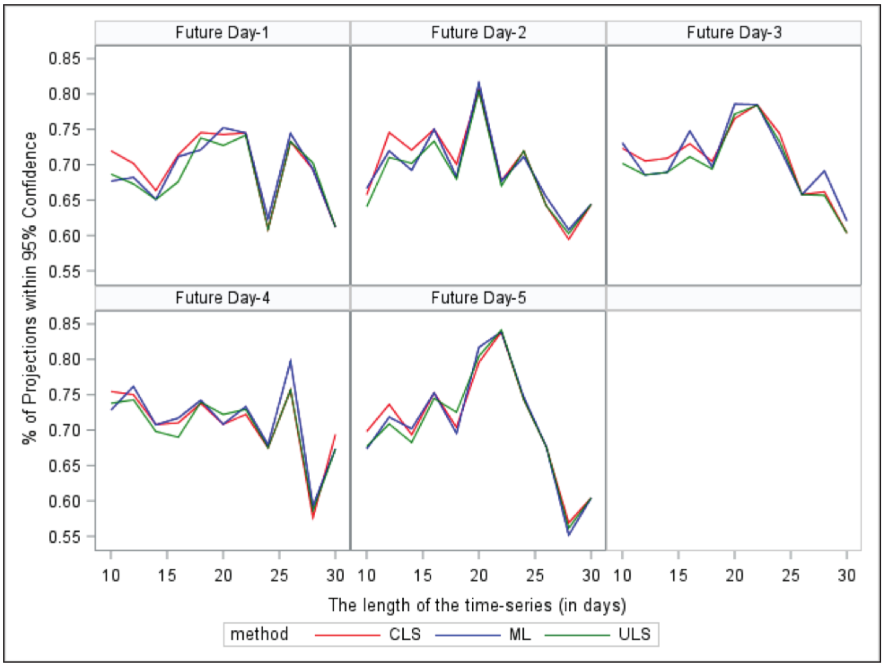

FIGURE 3: Observed level of confidence in time-series projections by estimation method under the AR(1) autocorrelation approach.

For death projections, AR(2) model provides a more confident modeling strategy as shown in Table 4 and Figure 4 and Figure 5 below. Within AR(2), CLS is a better estimation approach (Figure 5). Therefore, we suggest that time-series models with CLS and AR(2) be the model choice in predicting the future COVID-19 deaths.

DISCUSSIONS

In this research, we compared three auto-correlation structure and three estimation approaches in time-series modeling to predict the future cases and deaths due to COVID-19 using all accumulated data from all countries with at least total cases as of April 07, 2020. Obtaining future projections is critical for the central and local governments in health-care resource management and planning. Therefore, models that accurately estimate the future events need to be identified.

Among the six model combinations we compared, for COVID-19 cases, the model comparison measures we used suggested that a time-series model with Conditional Least-Squares estimation using AR(1) auto-corre-

TABLE 4: Overall observed level of confidence and mean absolute residuals in time-series projections across all future day projections from day-1 through day-5.

		AR(1)			AR(2)			AR(3)		
		CLS	**ML**	**ULS**	**CLS**	**ML**	**ULS**	**CLS**	**ML**	**ULS**
Future Day-1	% of Projections within 95% CI	56.5%	56.4%	**56.7%**	**64.3%**	61.7%	63.1%	58.1%	**59.1%**	58.1%
	Mean Absolute Residual	3.7	**3.6**	**3.7**	**3.7**	**3.7**	**3.7**	**4.4**	**4.4**	**4.4**
Across Future Day-1 to Day-5	% of Projections within 95% CI	55.5%	**56.0%**	55.2%	**53.2%**	52.4%	52.6%	52.5%	**53.5%**	52.2%
	Mean Absolute Residual	6.2	**6.1**	6.2	**6.2**	**6.2**	6.3	6.4	**6.3**	**6.3**

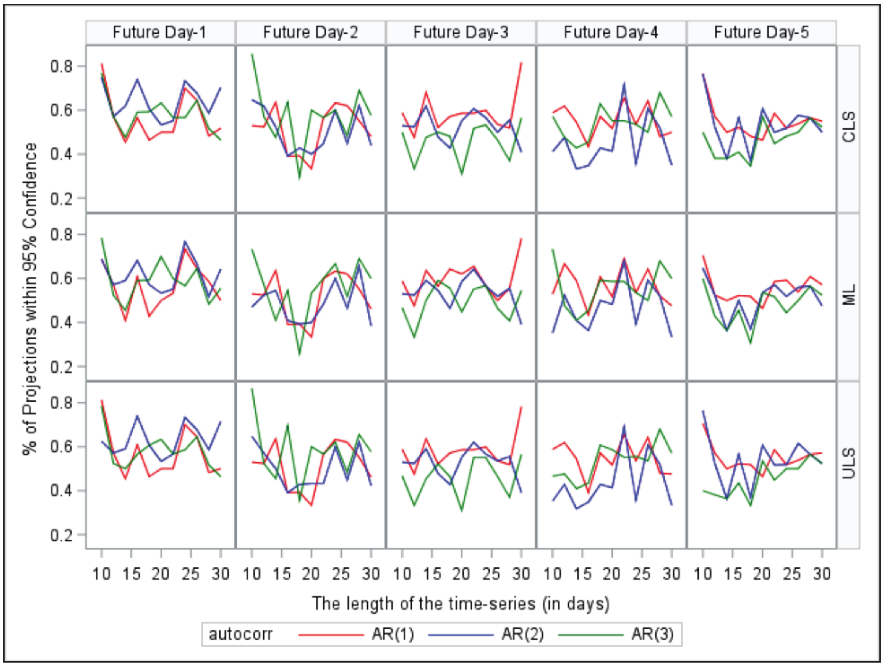

FIGURE 4: Observed level of confidence in time-series case projections by auto-regression structure.

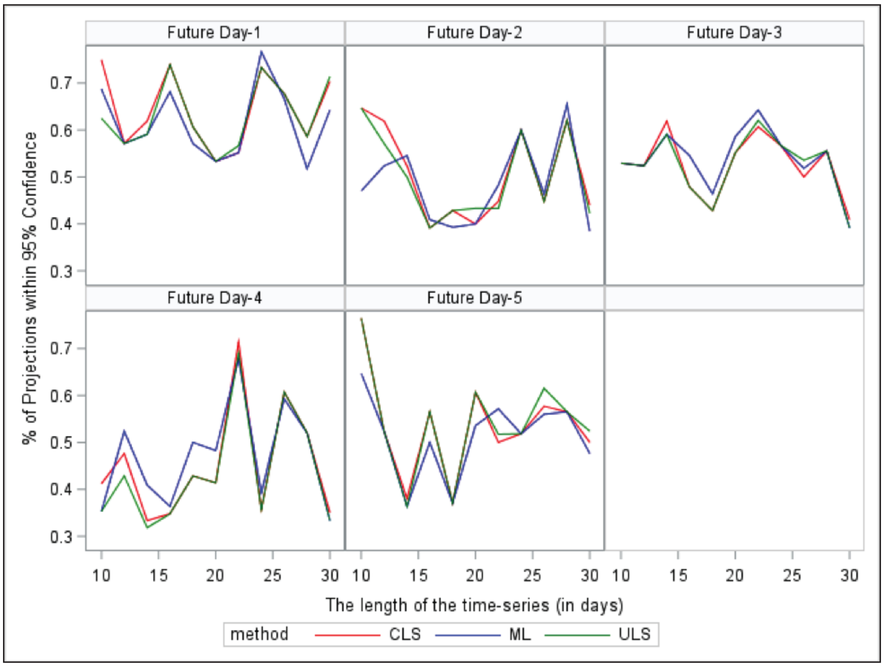

FIGURE 5: Observed level of confidence in time-series death projections by auto-regression structure AR(2).

lation structure perform slightly better although using a Maximum Likelihood estimation approach would also provide similar confidence with a slightly increased error. In projecting for future COVID-19 deaths, the model comparison measures we used suggested that a time-series model with Conditional Least-Squares estimation using AR(2) auto-correlation structure perform slightly better than its competitors.

In all these models, a main similarity was the fact that each model achieved an observed confidence level much smaller than the targeted level (Table 2). For example, with a 95% projection band, the models for COVID-19 case projections achieved an observed confidence less than 75% regardless of the time-series data size. For COVID-19 deaths, it was below 70%. This may be due to not timely reporting of the cases and deaths as most countries put their efforts in pandemic control efforts than timely reporting of the data. It is easy to find examples that some countries do not provide data on some days, and provide the cumulative data on another day, which creates an undue variation in the accumulating time-series data and breaks the auto-correlation structure inherently existed in the profile. Some of these data flowing issues may be manually corrected; however, we chose not to do it in this manuscript so that the data we used remains untouched.

One of the major difficulties in projecting pandemic events is that the pandemic field is a very dynamic field, which has a high potential to break any auto-correlation structure in the data. For example, any country-wide quarantine efforts or population shifts such as bringing a country's citizens back to their home country from epicenters of pandemic in other parts of the world, or a health-care system of a country being overwhelmed by the influx of new cases, may easily shift the magnitude and the direction of the changes in the time-series profile, which makes the estimation process much harder.

Despite all such challenges, health administrators and governments need to see what near future holds in terms of what the expected additional burdens on the hospitals and societies are, projections have to be made even with reduced confidence as we illustrated, so that better near future planning regarding hospital beds, intensive care unit rooms, quarantine facility needs, even burial house preparations and cemetary allocations can better be planned.

CONCLUSION

We conclude that for COVID-19 cases, CLS time-series models with AR(1) autocorrelation structure and for COVID-19 death, CLS time-series models with AR(2) autocorrelation structure can provide a reasonable future projection strategy, which can easily be implemented in all statistical packages, including but not limited to, SAS, R, Stata, SPSS, Minitab, etc.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

This study is entirely author's own work and no other author contribution.

REFERENCES

1. Tyrrell DA, Bynoe ML. Cultivation of viruses from a high proportion of patients with colds. Lancet. 1966;1(7428):76-7. PMID: 4158999. [Crossref]
2. Lim YX, Ng YL, Tam JP, Liu DX. Human coronaviruses: a review of virus-host interactions. Diseases. 2016;4(3):26. PMID: 28933406. [Crossref] [PubMed] [PMC]
3. Liu W, Wang F, Li G, Wei Y, Li X, He L, et al. Analysis of 2019 novel coronavirus infection and clinical characteristics of outpatients: an epidemiological study from the fever clinic in Wuhan, China. Lancet. 2/14/2020. Available at SSRN: https://ssrn.com/abstract=3539646 [Crossref]
4. Jones RH. Maximum likelihood fitting of ARMA models to time series with missing observations. Technometrics. 1980;22(3):389-95. [Crossref]
5. Kohn R, Ansley CF. Efficient estimation and prediction in time series regression models. Biometrika. 1985;72(3):694-7. [Crossref]
6. Ljung GM, Box GE. On a measure of lack of fit in time series models. Biometrika. 1978;65(2):297-303. [Crossref]
7. Tsay RS, Tiao GC. Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models. J Am Stat Assoc. 1984;79(385):84-96. [Crossref]
8. SAS Institute. SAS/ETS 9.1 User’s Guide. SAS Institute; 2004. p.2436.