Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine Update

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine-update

Forecasting COVID-19 pandemic in Alberta, Canada using modified ARIMA models

Jian Sun, PhD [a,b,*]

[a] *School of Public Health, University of Alberta, Edmonton, Alberta, Canada*
[b] *Department of Medicine, University of Calgary, Calgary, Alberta, Canada*

ARTICLE INFO

Keywords:
ARIMA
COVID-19
Infectious disease
SAS
Time series analysis

ABSTRACT

Background and objectives: Auto regressive integrated moving average (ARIMA) model is a popular model to forecast future values of a time series using the past values of the same series. However, if the variance of the time series varies with time, the 95% confidence interval estimated by the ARIMA will not be accurate. This study proposes a method to revise the ARIMA model to suit time series with heteroscedasticity.
Methods: Multiple historical ARIMA models were constructed with publicly available COVID-19 data in Alberta, Canada. The time series between different time periods were applied for these models. The means and their 95% confidence intervals of the differences between the forecasted values and the corresponding actual values were computed. The forecasted values of the general ARIMA models were modified by adding these differences.
Results: The average incident cases forecasted with the proposed method are lower than those with a general ARIMA model during the forecasted period. The 95% confidence intervals of the forecasted incidence with the proposed method are narrower. During the forecasted period (13 weeks) the average incidence was predicted to increase first and then decrease exponentially.
Conclusion: The proposed method can be used to automatically specify the best ARIMA model, to fit time series with heteroscedasticity and to forecast longer period of the trends in the future. In the next 13 weeks, the Covid-19 incidence may decrease but not eliminate. To stop the transmission of infections eventually, persistent effects complying with accurate forecasts are necessary.

1. Introduction

In December 2019, a series of pneumonia cases with no identified cause appeared in Wuhan, China, with clinic symptoms similar to viral pneumonia [1, 2]. The novel coronavirus was referred to as SARS-CoV-2, which belongs to the same family of virus that caused severe acute respiratory syndrome (SARS) in 2003 [3]. On March 11, 2020 World Health Organization (WHO) named the disease as coronavirus disease 2019 (COVID-19) [3].

Since the original outbreak, COVID-19 has spread to 221 countries and territories. As of September 1, 2021, a total of 218 930 223 confirmed cases and 4 539 725 deaths were reported [4]. In Alberta, Canada, the first COVID-19 case was reported on March 6, 2020, and through August 26, 2021, a total of 248 954 cases were reported. Among them, the number of deaths reached 2 364 [5].

To control the pandemic spread, Alberta Government implemented a series of strategies including travel restriction, isolation requirement and social distancing mandate. Vaccinations started on December 14, 2020. As of August 26, 2021, 77.7% of 12+ population has received at least one dose (66.1% of the total population in the province), and 69.3% were fully vaccinated (59% of the total population) [5]. With the progress of the vaccinations, Alberta Government planned reopening by stage. To effectively control the pandemic and resume socioeconomic activities, accurately predicting the COVID-19 dynamics using appropriate statistical models are necessary.

Various statistical models were used to predict the upcoming number of cases and forecast future spread of infectious diseases [6]. Ceylan [7] reviewed different statistical methods such as multivariate linear regression [8], grey forecasting models [9], backpropagation neural networks [10], and simulation models [11]. ArunKumar et al. [6] reviewed machine learning models for short-term forecasts of COVID-19 pandemic, such as support vector regression (SVR) [12], cubic regression [13], random forest [13], ridge regression [13], deep learning long short-term memory (LSTM) network [14] and so on.

[*] Corresponding author at: 1933 119A Street SW, Edmonton AB, T6W 0E2, Canada.
E-mail address: sunjian@hotmail.com.

https://doi.org/10.1016/j.cmpbup.2021.100029
Received 8 April 2021; Received in revised form 1 September 2021; Accepted 24 September 2021
Available online 26 September 2021

The ARIMA models which were first popularized by Box and Jenkins [15], have been successfully applied in the past to estimate the incidence and prevalence of various infectious diseases [7,9,10]. Recently, ARIMA models had also be used for forecasting COVID-19 pandemic [6,7,13]. However, because ARIMA is suitable for time series with consistent means and variances, it could not be used directly if the series do not satisfy the conditions. Differencing is a preferred approach to stabilize the mean values of the time series, while log transformation can usually reduce but not eliminate the nonconsistency of the variance (i.e., heteroscedasticity). Heteroscedasticity where the spread is proportional to the mean of the time series will tend to be improved by taking logarithm, but if it's not increasing with the mean, then the heteroscedasticity may be made worse by that transformation. In addition to log transformation, other types of transformations such as cube, square, square root and so on are possible [16]. If the approximate form of the heteroscedasticity were known, a type of transformations might approximately make the variance constant. However, in most of the cases, heteroscedasticity cannot eliminate by transformation, because the relationship between the variance and mean of the time series are unknown. If the variance of the time series varies with time, the 95% confidence interval (95% CI) estimated by the ARIMA will not be accurate. In addition, the current values of the time series usually have stronger correlations with recent values of the series, but not with earlier ones. With the traditional ARIMA modeling technique, information contained in the earlier series is ignored or attenuated.

This paper proposes a method to modify the traditional ARIMA model. Information computed from historical ARIMA models and the real data reported in the past were applied to adjust the forecasted values derived from the general ARIMA model. As an example, weekly incident cases of the Alberta COVID-19 were analyzed and forecasted.

2. Statistical methods

Basic statistical methods for model building and goodness test are described briefly in this section. To apply the proposed method, limited statistical knowledges are needed.

2.1. ARIMA model

A time series can be expressed as a set of data points ordered in time. ARIMA is a class of models that explains a given time series based on its own past values. The equation derived by ARIMA can be used to forecast future values. Theoretically, any nonseasonal time series that exhibits patterns and is not random white noise can be modeled with ARIMA models. A general ARIMA model can be classified as an ARIMA (*p, d, q*) model, where *p* is the number of autoregressive (AR) terms, *q* is the number of moving average (MA) terms in the prediction equation, and *d* is the order of differences needed for stationarity. The first-order and second-order of differences can be represented as Eqs. (1) and (2), respectively.

$$y'_t = y_t - y_{t-1} \tag{1}$$

$$y''_t = y_t - 2y_{t-1} + y_{t-2} \tag{2}$$

Higher orders of differences were seldom used for ARIMA models.

The ARIMA model may be simplified to ARMA model if $d = 0$. ARMA model may be further simplified to AR model if $q = 0$, or MA model if $p = 0$. AR (*p*) model refers to the current value of the time series y_t linearly in terms of its previous values $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ and the current residuals ε_t. MA (*q*) model refers to the current value of the time series y_t linearly in terms of its current and previous residual series $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$ [7]. The general formula of AR (*p*) and MA (*q*) models can be expressed in Eqs. (3) and (4), respectively.

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t \tag{3}$$

$$y_t = C + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \tag{4}$$

where *C* is a constant, $\phi_{1}, \phi_{2}, \ldots, \phi_p$ are autoregressive model parameters and $\theta_{1}, \theta_{2}, \ldots, \theta_q$ are moving average model parameters.

An ARMA model can be represented as

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \tag{5}$$

where *p* represents a gap between *t* and *t-p*, and *q* represents a gap between *t* and *t-q*.

ARIMA models can also be expressed as in Eq. (5) but $y_{t,}\ y_{t\text{-}1,}\ y_{t\text{-}2,} \ldots, y_{t\text{-}p}$ in the equation represent the differences of the time series instead of the original values of the time series.

2.2. Subset ARIMA models

A model that includes parameters for only some lags is called a subset model [17]. For example, an ARIMA(2,0,3) can be represented as in Eq. (6). Eq. (7) is a subset model of ARIMA(2,0,3) which may be expressed as ARIMA((2),0,(1,3)).

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_q \varepsilon_{t-3} \tag{6}$$

$$y_t = C + \phi_2 y_{t-2} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_q \varepsilon_{t-3} \tag{7}$$

For ARIMA(2,0,3), AR and MA terms can be expressed as $p = 2$ and $q = 3$, respectively. Similarly, for subset ARIMA((2),0,(1,3)), AR and MA terms can be expressed as $p = (2)$ and $q = (1\ 3)$, respectively.

2.3. Schwarz Bayesian criterion (SBC)

The Schwarz Bayesian criterion (SBC) [18] or Bayesian information criterion (BIC) is a popular criterion for model selection among a finite set of models. In this study, from many possible models with different lags and types to fit the same time series, the one with the lowest SBC was selected as the best model. SBC is defined as in Eq. (8).

$$SBC = -2logL + KlogN \tag{8}$$

where *L* is the likelihood, *K* is the number of model parameters, and *N* is the number of observations of the time series used to build the model.

3. Calculation

3.1. Data source

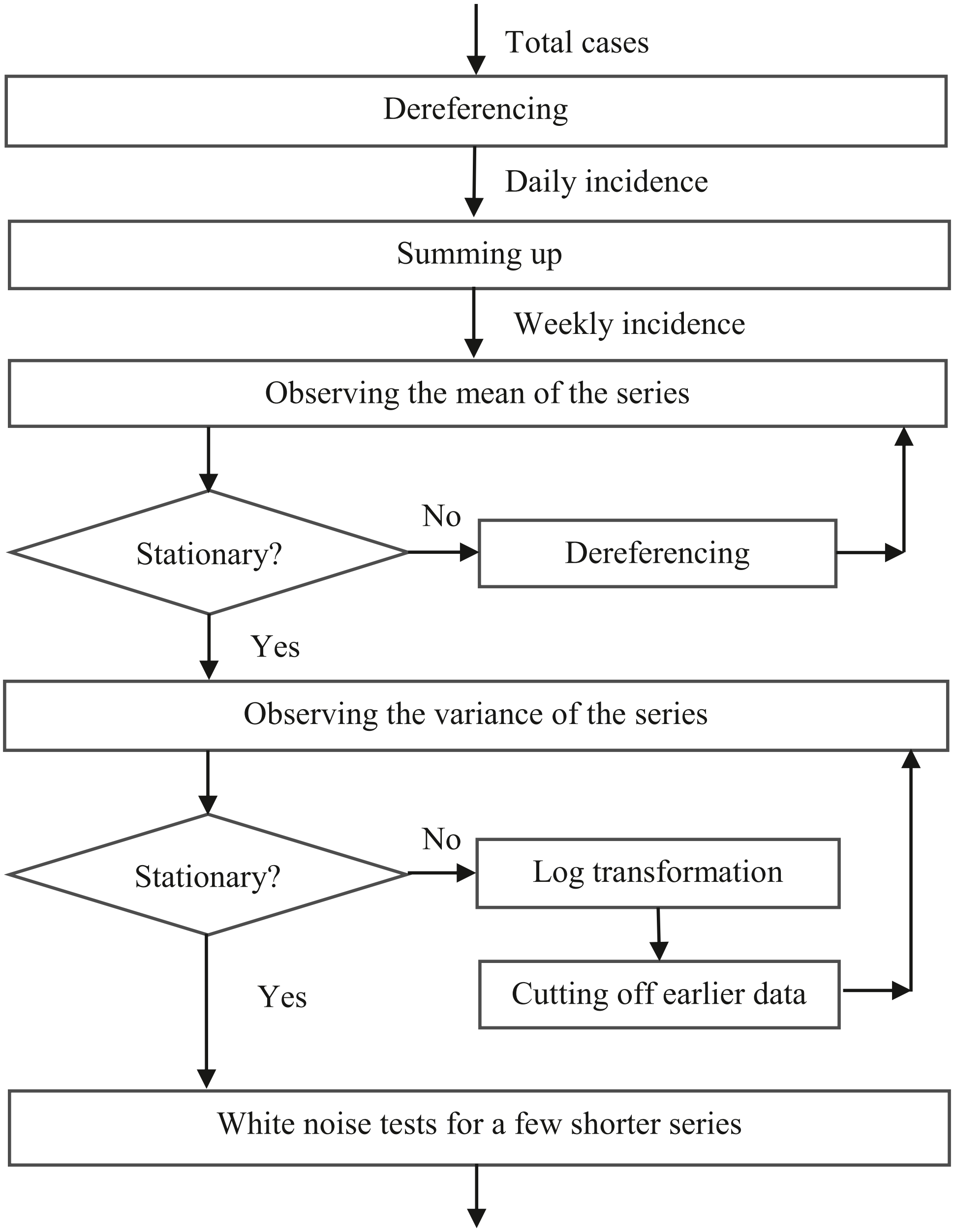
The Government of Alberta posts cumulative numbers of COVID-19 cases on its website, which contains the numbers of confirmed and possible cases [5]. Both confirmed and possible cases were included in the calculations, considering that diagnosis dates might be later than actual infected dates and omissions might occurred.

3.2. Data preprocessing

The algorithm for data preprocessing is shown in Fig. 1. The number of total cases was collected from March 6, 2020 to August 26, 2021. Daily incidences derived by first-order differencing were summed up to 77 weekly incidences.

The weekly numbers were differenced and log transformed for stationarity. The time series used for analysis started on May 1, 2020 (the start date of the 9th week from the date while the first case occurred in the province). Because the variances of the earlier values of the incidence after log transformation were still larger, they were excluded from the analysis.

In Eqs. (3)–(5), subscript *t* represents the "current" time. It is usually the last available time point of the series. This approach needs to construct



file a, start date point: *start* = 9, the earliest end date point: *end1* = 48

Fig. 1. Algorithm for data preprocessing.

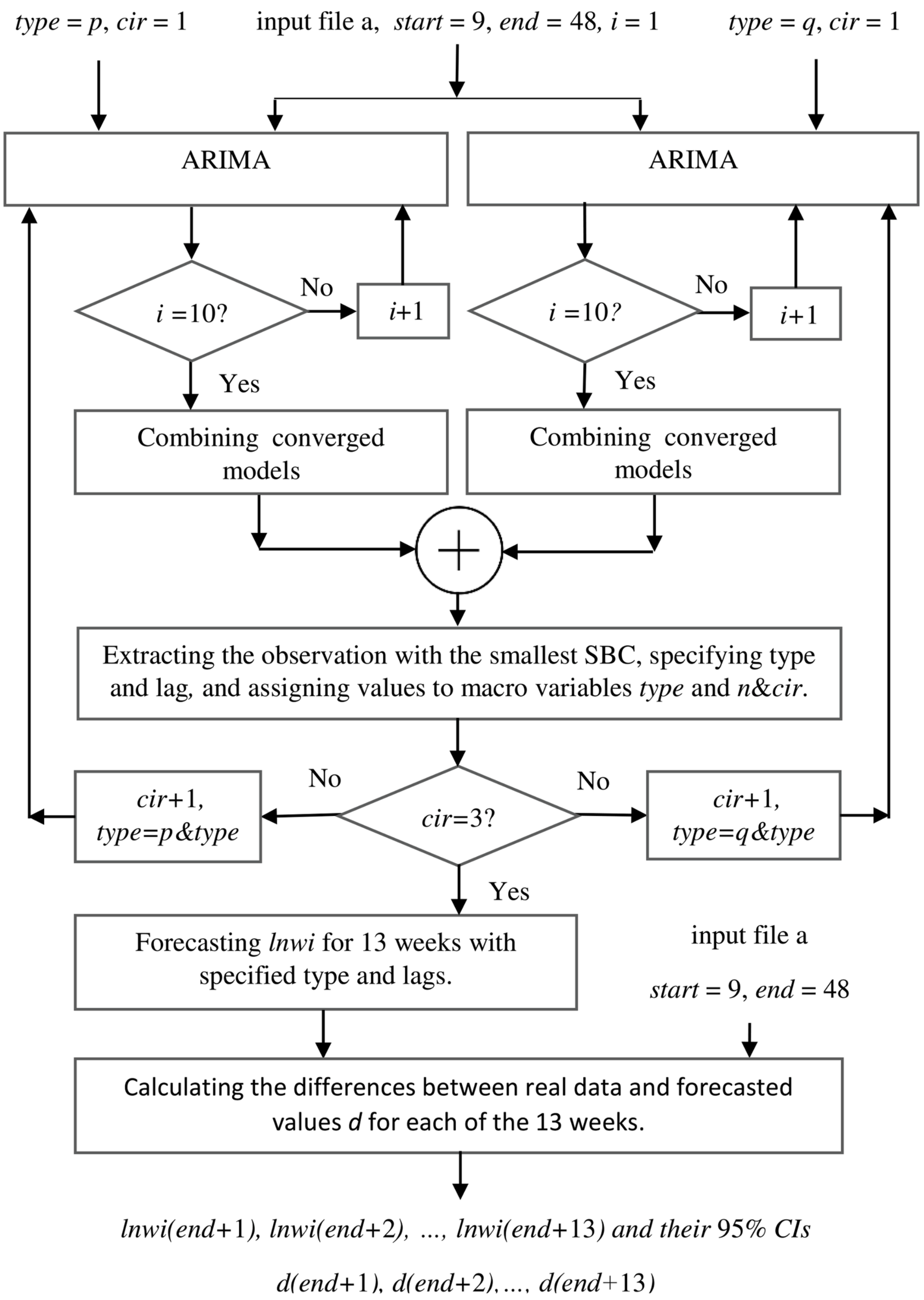

type = p, cir = 1
input file a, start = 9, end = 48, i = 1
type = q, cir = 1
ARIMA
ARIMA
i =10?
No
i+1
Yes
i =10?
No
i+1
Yes
Combining converged models
Combining converged models
+
Extracting the observation with the smallest SBC, specifying type and lag, and assigning values to macro variables type and n&cir.
cir+1, type=p&type
No
cir=3?
No
cir+1, type=q&type
Yes
Forecasting lnwi for 13 weeks with specified type and lags.
input file a
start = 9, end = 48
Calculating the differences between real data and forecasted values d for each of the 13 weeks.
lnwi(end+1), lnwi(end+2), …, lnwi(end+13) and their 95% CIs
d(end+1), d(end+2),…, d(end+13)

Fig. 2. Algorithm for specifying type and lags, building the best model to fit the series ended in week 48, and calculating the differences between real data and forecasted values.

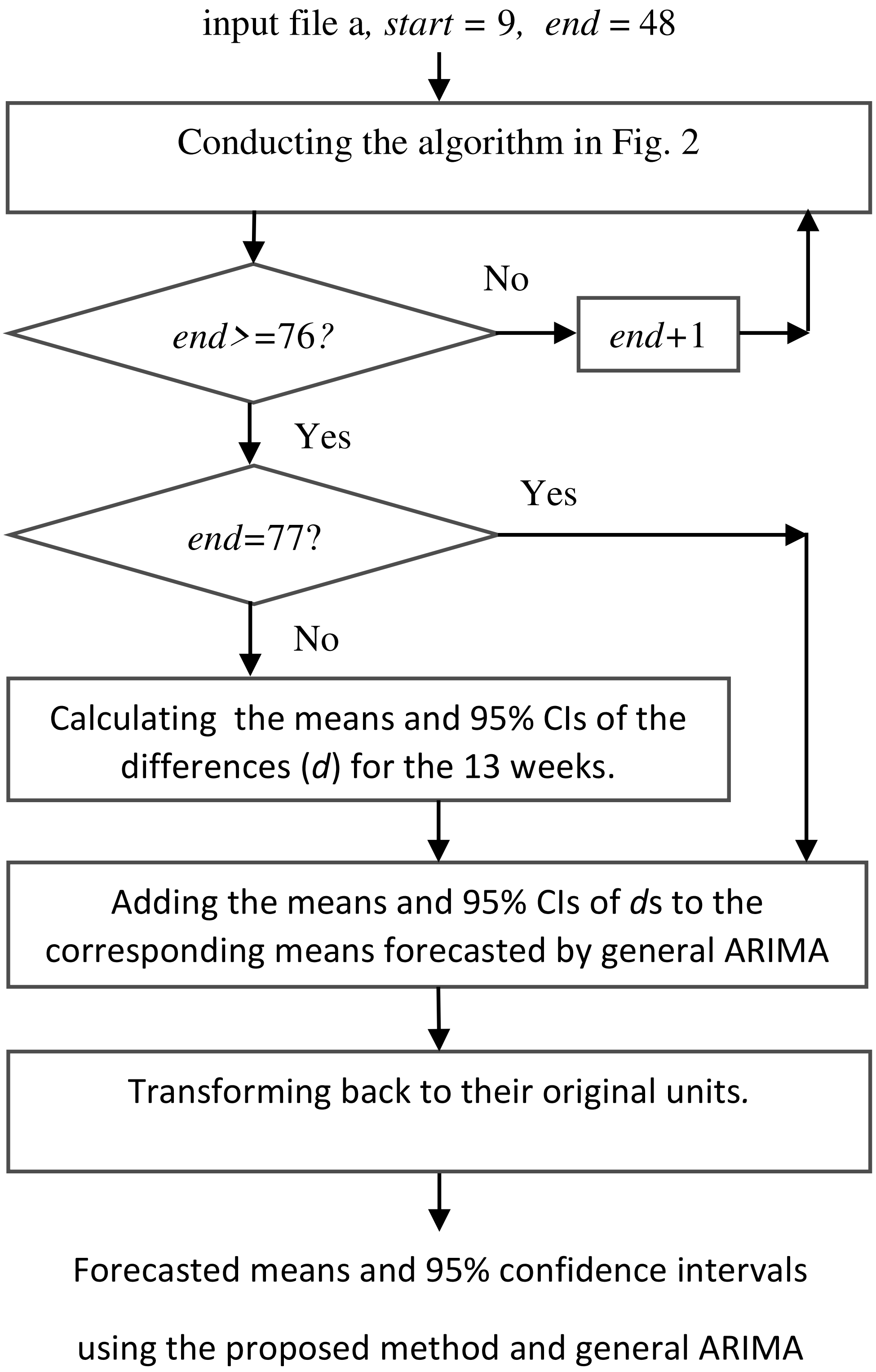

input file a, *start* = 9, *end* = 48
Conducting the algorithm in Fig. 2
end>=76?
No
end+1
Yes
end=77?
Yes
No
Calculating the means and 95% CIs of the differences (*d*) for the 13 weeks.
Adding the means and 95% CIs of *d*s to the corresponding means forecasted by general ARIMA
Transforming back to their original units.
Forecasted means and 95% confidence intervals
using the proposed method and general ARIMA

Fig. 3. Algorithm for forecasting means and 95% confidence intervals.

multiple ARIMA models for the same series with different time periods. In this paper, the model used to fit the time series ended at time *t* is called current ARIMA model. The others for the same series but ended earlier (before *t*) are called historical ARIMA models.

SAS ARIMA procedure [17] was used for the analyses. To fully use historical information, the end date points (weeks for this study) for the first and subsequent historical ARIMA models should be as early as possible. The same series with different end weeks were pretested by IDENTIFY statement in the ARIMA procedure. The earliest end week (week 48) was determined based on approximate white noise tests of the hypothesis that none of the autocorrelations of the series up to lag equals 6 are significantly different from 0. According to the tests with a few shorter series, any series ended earlier than week 48 was a white noise series. Further analyses with those series were not necessary.

The description of data preprocessing appears complicated. With a general SAS ARIMA procedure, however, the number of differences, number of log transformations, start and the earliest end weeks were determined in minutes.

3.3. Model constructing

The algorithm in Fig. 2 illustrated the process to specify the model type and 3-lag combination for the best model building. To specify the type and lags, six circles of estimation processes were conducted for a series started in week 9 and ended in week 48.

At the first two circles, all possible subset AR and MA models with a single lag from 1 to 10 were created. Each model outputted one observation which consisted of a lag, model type, SBC and *COVN*. Variable *COVN* = 0 indicates that the estimation process has converged [17]. All observations with *COVN* = 0 were combined. The first type and lag number contained in the observation with the smallest SBC were specified.

At the next two circles, the lag number derived from the first two circles were assigned to "*p*" or "*q*" terms based on the derived type. All possible subset ARIMA models with this fixed lag and another one varied from 1 to 10 were created for the same series. With the same processes, the second type and lag number contained in the observation were specified. For example, if at the first two circles, *type* = "*p*" and lag equals 1 was specified, the *type* would be updated to "*pp*" and "*qp*". Then all ARIMA models with *p* = 1 and any other lag up to 10 for *p* or *q* would be conducted and the best model with two lags were determined.

Similarly, the last two circles specified the third lag and updated the type again. The best ARIMA model to fit the series ended in week 48 was built using the lags of correlations specified at this circle pair. Finally, the differences between real data and forecasted values *d* (.) for each of the 13 weeks were calculated.

3.4. Modifying general ARIMA and forecasting

Fig. 3 illustrates the algorithm for forecasting means and 95% CIs using the proposed method and general ARIMA. The algorithm in Fig. 2 was iterated from *end* = 48 to 77 to fit the series ended from weeks 48 to 77. With these best models, except the last one for the series ended at week 77, 13-week forecasts were performed by FORECAST statements in SAS ARIMA procedure. The results were compared with the actual values which had been reported. The means and their 95% CIs of the differences between actual values and the corresponding forecasted values for each of the 13 weeks were computed. The forecasted mean values derived from the general ARIMA model fitted to the whole time series were modified by adding the differences (means and 95% CIs).

Finally the forecasted means and the 95% CIs of the log transformation of the weekly incidences were transformed back to their

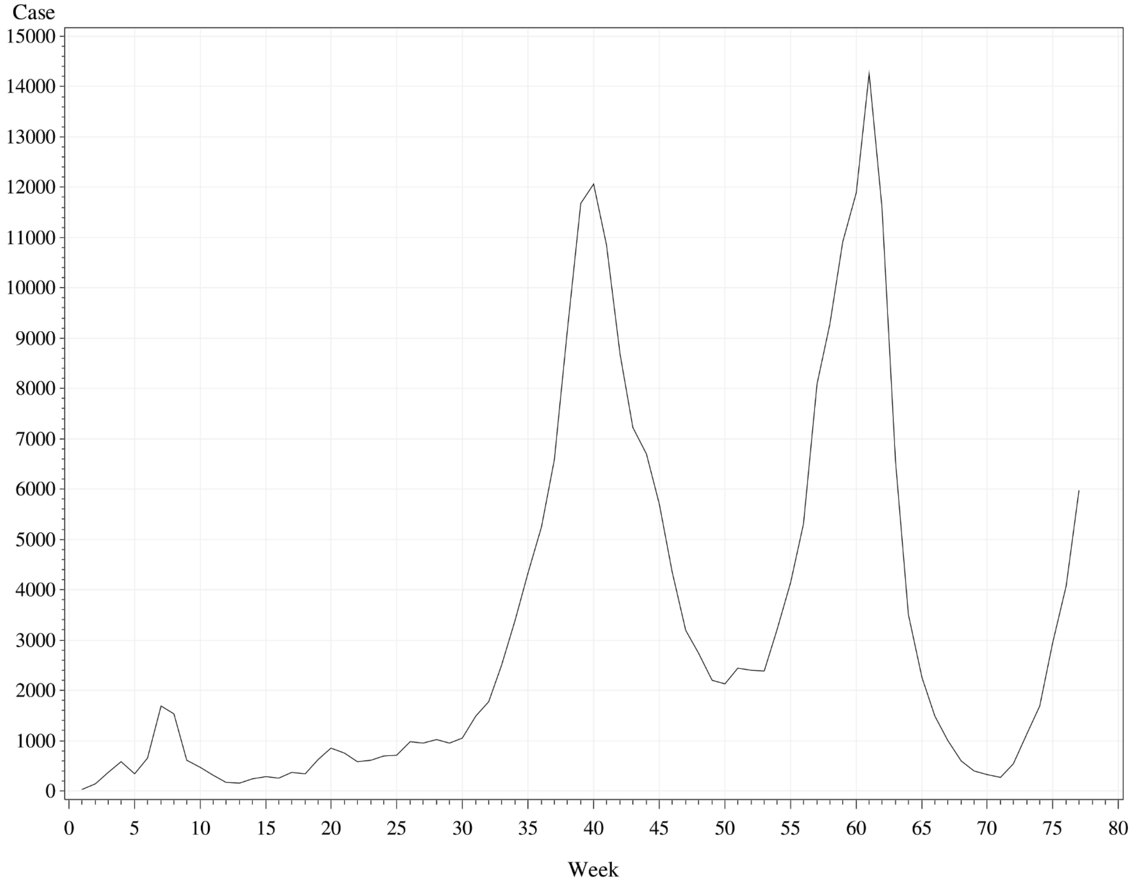


Fig. 4. Weekly number of incidences (confirmed and possible).

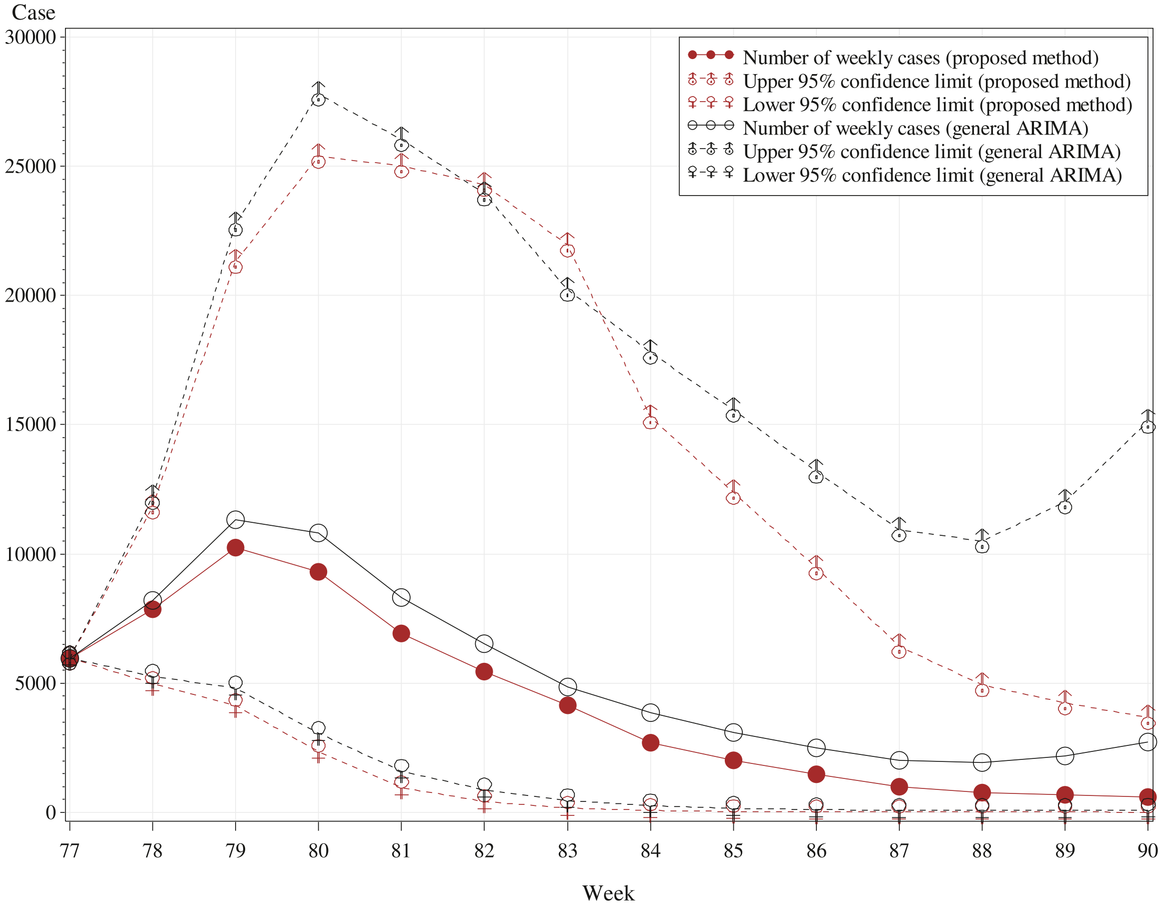


Fig. 5. 13-week forecast of the weekly incidence.

original units by exponentiation. Before conducting the transformation, a square of the standard error of the forecast divided by 2 (*STD*STD/2*) was added to the logarithms of the means [17].

In Appendix, Table A.1 lists weekly incidence (*wi*) data in Alberta. SAS LOG function in data step A creates *lnwi*, the log transformation of *wi*. Table A.2 is a SAS macro for model building, modifying and forecasting. Table A.3 provides a SAS code for plotting the forecasted results for readers to verify the SAS macro.

4. Results

The entire course of the weekly number (confirmed and possible) is illustrated in Fig. 4. During the 77 weeks, the curve varied violently. In weeks 7, 40 and 61, the fluctuated curve reached its three peaks (1 682, 12 065 and 14 262), respectively. Then it dropped down to 266 in week 71 (July 9 – 15, 2021). From week 71, the curve was rising again quickly.

Using the SAS macro in Table A.2, "$p = (1\ 8)$ $q = (8)$" was specified for each ARIMA model to fit time series ended from weeks 48 to 60. "$p = 1$ $q = (8)$" and "$p = 2$ $q = (8)$" were specified for series ended in week 61 and 62, respectively. while "$p = (1\ 8)$" was specified for other models to fit the series ended at week 63 and later. With the specified lag numbers and types, a total of 30 ARIMA models were constructed.

Fig. 5 illustrated the 13-week forecasts for the weekly cases by using the proposed method and general ARIMA model. The mean was predicted to increase from the reported number 5 972 in week 77 (August 20 - 26, 2021) to 10 248 (95% CI: 4 135 – 21 318) in week 79 (September 3 - 9, 2021) by using the proposed method, lower than that by general ARIMA which increased to 11 319 (95% CI: 4 810 - 22 761). From week 79 the number begin to decline slowly. In week 88, the curve derived from general ARIMA turns to increase, while the curve derived from the proposed method continues to decline to 610 (95% CI: 12 – 3 671) in week 90 (November 18 - 24, 2021). As shown in Fig. 5, using the proposed method, most of the 95% CIs were narrower than those using the general ARIMA model.

5. Discussion

5.1. Why new method?

COVID-19, a newly emerged infectious disease, has spread to almost all the world including Canada, with severe impact on health and public health systems, human lives and world economics. Applying statistical modeling to forecast the trend, the optimal design of effective public health policy for prevention and control of COVID-19 is important. However, it is not easy for healthcare researchers or policy makers to construct appropriate models, if they do not have a strong background in statistics.

This paper proposed a novel method to forecast future values of COVID-19 pandemic in Alberta, Canada. To use this method, instead of complicated mathematical calculations, basic expertise on ARIMA modeling and SAS coding techniques are sufficient.

5.2. What is new?

Instead of building one ARIMA model to fit the whole available time series, multiple ARIMA models were built to fit the same series but ended on different dates. Forecasts were performed with these historical models. The predicted values were compared with the real values. The differences between the real and the forecasted values were used to modify the results of the general ARIMA model.

A SAS macro was created to specify the lags of correlations (Table A.2

in Appendix). Iterative calculations were completed by the macro. Because current values of the time series usually have stronger correlations with recent values of the series, but not with earlier ones, the iteration number for lag specification was assigned to 10 (I initially assigned it to 16). A total of 1 800 (60x30) possible three-level lag and type combinations were created for the models to fit the same series but ended in 30 different weeks. The model type and lag combination with the smallest SBC was selected for each model. The best ARIMA model with the specified type and lags was constructed automatically by the macro.

5.3. 95% confidence interval

The proposed method solved the intrinsic problem of the general ARIMA models, which is not suitable for the series with heteroscedasticity such as the Alberta COVID-19 data. The 95% CIs derived by using the proposed method may be more reasonable, since they came from the real forecast errors in the past and all the forecasts were performed by using the best models.

Even if the prevention and control measures for the pandemic were consistent, due to some unexpected impacts, the forecasted values would still vary between the predicted 95% CIs with 95 percent possibility. The longer the forecasting period, the broader the 95% CI would be. To accurately forecast future values of the time series, narrower 95% CIs are expected. Because the 95% CIs of the forecasted incidences derived by using the proposed method were narrower than those derived from a general ARIMA model, long-term forecasts were possible.

5.4. The best model

Identifying an appropriate ARIMA model to fit the time series based on autocorrelation and partial autocorrelation function plots is not easy. A larger autocorrelation function in the output of an IDENTIFY statement in SAS ARIMA procedure may not necessarily be a significant parameter when it is selected for the modeling. In contrast, a weak autocorrelation displayed in the plot graph may be a stronger parameter but cannot be identified. In some circumstances, similar as the effects of covariates on multivariable regression models [19], a weaker autocorrelation becomes stronger while it is combined with another autocorrelation with different lag. Using the proposed method, Analysts don't need to observe the plots of autocorrelations or partial autocorrelations for model identifications. Specifying the lags and model type, and constructing the best model can be performed automatically by running the macro in Table A.2.

If accurate forecasts are not required in public health practice, a general ARIMA can be used. However, the proposed method is still necessary for specifying the best ARIMA models. To correctly forecast values in the future, correct models must be specified first to fit the series. The so-called best model is a relatively optimum model because we cannot compare all possible models. Since the comparations included subset models, the "best" models specified by iterations had higher possibility to be the real best models. To my knowledge, at least for COVID-19 studies, very few researchers [20] considered subset models (sparse coefficient ARIMA) when they specified their best ARIMA models. With the traditional ARIMA modeling technique, manually specifying lags and model type for the best subset model is difficult, or sometimes even impossible.

SBC was used to select the best models. Using SBC, candidate models with fewer lags of correlations are easier to be selected than using Akaike information criterion (AIC) [21]. In addition, the estimation processes for some models might not converge. Those models were automatically excluded by running the macro in Table A.2.

The best ARIMA model was selected from candidate models with one, two and three lags of correlations. With the same processes, I investigated models with four lags, All of them had larger SBC. For different series, if more lags were selected, the effects on forecasting might be ignorable.

5.5. Data smoothing

The proposed method can be used to specify a "best" model but needs more calculations. To simplify the calculations, daily data were summed up to weekly data. Grouping data is a useful smoothing method to remove noise from a dataset, allowing important patterns to stand out. To predict the long-term trend of the pandemic, smoothed data are more convenient than daily data with severe fluctuation.

6. Conclusions

This paper described a novel method to forecast COVID-19 pandemic in Alberta, Canada. The proposed method can be used to analyze time series with heteroscedasticity, to specify the best model, and to forecast the series into relatively far future.

The COVID-19 pandemic in Alberta, Canada was predicted to increase in the coming weeks and decrease later. To stop the transmission of infection eventually, persistent efforts are necessary for reducing the incidence.

Declaration of Competing Interest

The author declares that there is no conflict of interest. The author does not have financial and personal relationships with other people or organizations that could inappropriately influence (bias) this work. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgment

The author thanks Professor Devidas Menon, School of Public Health, University of Alberta, for his encouragement and initial manuscript revision.

Appendix. COVID-19 incidence data and SAS codes for model building and forecasting

Table A.1
Log transformation of weekly incidence (*lnwi*) in Alberta.

	Code	Comment
1	**data** a;	*wi* is weekly incidence.
2	input wi @@;	
3	week+**1;**	
4	lnwi = log(wi);	
5	cards;	
6	26 137 366 575 338 656 1682 1537 616 473 311 175 161 246 288 257 367 336 620 854	
7	758 586 608 689 716 982 944 1018 946 1054 1484 1773 2496 3385 4333 5246 6589 9129 11679 12065	
8	10852 8685 7227 6693 5705 4351 3198 2740 2203 2129 2439 2403 2387 3214 4145 5307 8093 9274 10910 11893	
9	14262 11592 6546 3508 2261 1497 1001 591 392 332 266 541 1135 1690 2941 4078 5972	
10	;	

Table A.2
SAS macro for model building and forecasting.

	Code	Comment
1	**%macro** arima(in = a, var = lnwi, start = **9**, end1 = **48**, end2 = **77**);	**Line 1:** Invoke macro ARIMA with variable name, start, the first and last end dates.
2	**%macro** a(est, type);	**Lines 2-18 (Macro A):** Estimate ARIMA with any possible lag up to 10. Exclude nonconvergent models. &var(1) = *lnwi* (1) on line 6 is the first-order difference of *lnwi*. *lnwi* is log transformation of weekly incidence.
3	%do i = **1** %to **10**;	
4	proc arima data = &in plots = none;	
5	where &start <= week <= &end;	
6	identify var = &var(**1**) noprint;	
7	estimate &est outstat = b&i(drop = _type_ where = (_stat_ in ("SBC" "CONV"))) noprint;	
8	run;	
9	proc transpose data = b&i out = c&i(keep = col1 col2);	
10	run;	
11	%end;	
12	data c&type(drop = col2);	
13	set c1-c10;	
14	if col2 = **0**;	
15	type = "&type";	
16	n1 = _n_;	
17	run;	
18	**%mend a**;	
19	**%macro** b(in1, in2, cir);	**Lines 19-31 (Macro B):** Specify the lags and type with the smallest SBC. Create macro variables for the lags and type.
20	data d;	
21	set c&in1 c&in2;	
22	run;	
23	proc sort;	
24	by col1;	
25	run;	
26	data e;	
27	set d;	
28	if _n_ = **1**;	
29	call symput("n&cir",put(n1, **2.**));	
30	call symput("type", type);	
31	**%mend b**;	
32	%do end = &end1 %to &end2;	**Line 32:** Start the iteration for each ARIMA model to fit the series ended in different week.
33	%***a***(%str(p = (&i)),p);	**Line 33:** Invoke macro A to estimate AR with one lag.
34	%***a***(%str(q = (&i)),q);	**Line 34:** Invoke macro A to estimate MA with one lag.
35	%***b***(p, q, **1**);	**Lines 35-36:** Invoke macro B to specify the 1st lag and type.
36	run;	
37	%if &type = p %then %do;	**Lines 37-44:** Invoke macro A with different 2-lag combinations based on the result of macro B.
38	%***a***(%str(p = (&i &n1)), pp);	
39	%***a***(%str(p = (&n1) q = (&i)), qp);	
40	%end;	
41	%else %do;	
42	%***a***(%str(p = (&i) q = (&n1)), pq);	
43	%***a***(%str(q =(&i &n1)), qq);	
44	%end;	
45	%***b***(p&type, q&type, **2**);	**Lines 45-46:** Invoke macro B to specify the 2nd lag and type.
46	run;	
47	%if &type =pp %then %do;	**Lines 47-62:** Invoke macro A with different 3-lag combinations based on the result of macro B.
48	%***a***(%str(p = (&i &n1 &n2)), ppp);	
49	%***a***(%str(p = (&n1 &n2) q = (&i)), qpp);	
50	%end;	
51	%else %if &type = pq %then %do;	
52	%***a***(%str(p = (&i &n1) q = (&n2)), ppq);	
53	%***a***(%str(p = (&n1) q = (&i &n2)), qpq);	
54	%end;	
55	%else %if &type = qp %then %do;	
56	%***a***(%str(p = (&i &n2) q = (&n1)), pqp);	
57	%***a***(%str(p = (&n2) q = (&i &n1)), qqp);	
58	%end;	
59	%else %do;	
60	%***a***(%str(p = (&i) q = (&n1 &n2)), pqq);	
61	%***a***(%str(q = (&i &n1 &n2)), qqq);	
62	%end;	
63	%***b***(p&type, q&type, **3**);	**Lines 63-64:** Invoke macro B to specify the 3rd lag and type.
64	run;	
65	data f&end;	**Lines 65-69:** Rename lags.
66	set e;	
67	n2 =&n2;	
68	n3 =&n1;	
69	run;	
70	%if &type = ppp %then %let lag = %str(p = (&n1 &n2 &n3));	**Lines 70-77:** Create macro variable *lag* based on the result of macro B.
71	%else %if &type=qpp %then %let lag=%str(p=(&n2 &n3) q=(&n1));	
72	%else %if &type=ppq %then %let lag=%str(p=(&n1 &n2) q=(&n3));	
73	%else %if &type=qpq %then %let lag=%str(p=(&n2) q=(&n1 &n3));	

(*continued on next page*)

Table A.2 (*continued*)

	Code	Comment
74	%else %if &type=pqp %then %let lag=%str(p=(&n1 &n3) q=(&n2));	
75	%else %if &type=qqp %then %let lag=%str(p=(&n3) q=(&n1 &n2));	
76	%else %if &type=pqq %then %let lag=%str(p=(&n1) q=(&n2 &n3));	
77	%else %let lag = %str(q = (&n1 &n2 &n3));	
78	proc arima data = &in plots = none;	**Lines 78-83:** Build an ARIMA to fit *lnwi*(1) with specified lags and type. Forecast incidences for 13 weeks.
79	where &start <= week <= &end;	
80	identify var = &var(**1**) noprint;	
81	estimate &lag noprint;	
82	forecast lead = **13** id = week out = result&end(drop = &var residual where = (week > &end)) noprint;	
83	run;	
84	%if &end < &end2 %then %do;	**Lines 84-92:** Calculate the difference between forecasted and actual data for each of the historical ARIMA. The difference for the current ARIMA is not available.
85	data arima&end;	
86	merge result&end(in = a) &in(in = b keep = week &var);	
87	by week;	
88	if a and b;	
89	d = lnwi-forecast;	
90	j = &end;	
91	run;	
92	%end;	
93	%end;	**Line 93:** End the iteration started on line 32.
94	data lag;	**Lines 94-96:** List lags and types for all ARIMA models.
95	set f&end1-f&end2;	
96	run;	
97	data arima(keep =n week j d);	**Lines 97-100:** Combine the differences.
98	set arima&end1-arima%eval(&end2-**1**);	
99	n =week-j;	
100	run;	
101	proc sql;	**Lines 101-106:** Calculate the mean and the 95% CI of the differences for each of the 13 weeks.
102	create table dif1 as	
103	select n+&end2 as week, mean(d) as meand, std(d) as stdd, mean(d)-**1.96***std(d) as l95d, mean(d)+**1.96***std(d) as u95d	
104	from arima	
105	group by n;	
106	quit;	
107	data dif2;	**Lines 107-117:** transform the means and 95% CIs derived from proposed method and a general ARIMA.
108	merge result&end2 dif1(in = a);	
109	by week;	
110	if a;	
111	forecastd = exp(forecast+meand+stdd*stdd/**2**);	
112	l95d =exp(forecast+l95d);	
113	u95d =exp(forecast+u95d);	
114	forecast =exp(forecast+std*std/**2**);	
115	l95 =exp(l95);	
116	u95 =exp(u95);	
117	run;	
118	proc sql;	**Lines 118-126:** Add the real data of the last date point (week 77) for plotting.
119	create table out as	
120	select week, wi as forecast, wi as l95, wi as u95, wi as forecastd, wi as l95d, wi as u95d	
121	from &in	
122	where week =&end2	
123	union all	
124	select week, forecast, l95, u95, forecastd, l95d, u95d	
125	from dif2;	
126	quit;	
127	**%mend arima**;	
128	%***arima***;	**Line 128:** invoke the macro.

Table A.3
SAS code for plotting the forecasted results.

	Code	Comment
1	axis1 label =("Week") minor =none;	Revisions may need to suit different data.
2	axis2 label =("Case") order =(**0** to **30000** by **5000**);	
3	legend1 label=none value=("Number of weekly cases (proposed method)" "Upper 95% confidence limit (proposed method)" "Lower 95% confidence limit (proposed method)" "Number of weekly cases (general ARIMA)" "Upper 95% confidence limit (general ARIMA)" "Lower 95% confidence limit (general ARIMA)") position=(top right inside) across=**1** frame;	*Forecastd, u95d* and *l95d* are the mean, upper and lower limits of the forecasted number of weekly cases by proposed method.
4	symbol1 color =brown interpol =line line =**1** value =dot height =**2**;	
5	symbol2 color =brown interpol =line line =**2** value ="," height =**2**;	

(*continued on next page*)

Table A.3 (*continued*)

	Code	Comment
6	symbol3 color =brown interpol =line line =**2** value ="*" height =**2**;	*Forecast, u95* and *l95* are the mean, upper and lower limits of the forecasted number of weekly cases by a general ARIMA.
7	symbol4 color =black interpol =line line =**1** value =circle height =**2**;	
8	symbol5 color =black interpol =line line =**2** value ="," height =**2**;	
9	symbol6 color =black interpol =line line =**2** value ="*" height =**2**;	
10	ods rtf file ="E:\mylab\figure5.doc";	
11	**proc gplot** data =out;	
12	plot (forecastd u95d l95d forecast u95 l95)*week/grid overlay haxis =axis1 vaxis =axis2 legend =legend1;	
13	**run**;	
14	ods rtf close;	

References

[1] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, et al., A Novel Coronavirus from Patients with Pneumonia in China, N. Engl. J. Med. 382 (2020) (2019) 727–733.

[2] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet 395 (2020) 497–506.

[3] D. Cucinotta, M. Vanelli, WHO declares COVID-19 a pandemic, Acta Biomed. 91 (1) (2020) 157–160, https://doi.org/10.23750/abm.v91i1.9397.

[4] Countries where COVID-19 has spread. https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/ (accessed September 1, 2021).

[5] Government of Alberta, COVID-19 Alberta statistics, Interactive aggregate data on COVID- 19 cases in Alberta. https://www.alberta.ca/stats/covid-19-alberta-statistics.htm (accessed August 27, 2021).

[6] K.E. ArunKumar, D.V. Kalaga, C.M.S. Kumar, G. Chilkoor, M. Kawaji, T.M. Brenza, Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA), Appl. Soft. Comput. (2021) 103, https://doi.org/10.1016/j.asoc.2021.107161.

[7] Z. Ceylan, Estimation of COVID-19 prevalence in Italy, Spain, and France, Sci. Total Environ. 2020 (2020) 729, https://doi.org/10.1016/j.scitotenv.2020.138817.

[8] M.C. Thomson, A.M. Molesworth, M.H. Djingarey, K.R. Yameogo, F. Belanger, L. E. Cuevas, Potential of environmental models to predict meningitis epidemics in Africa, Trop. Med. Int. Health 11 (6) (2006) 781–788, https://doi.org/10.1111/j.1365-3156.2006.01630.x.

[9] Y. Wang, Z. Shen, Y. Jiang, Comparison of ARIMA and GM(1,1) models for prediction of hepatitis B in China, PLoS One 13 (9) (2018), https://doi.org/10.1371/journal.pone.0201987.

[10] Q. Liu, Z. Li, Y. Ji, L. Martinez, U.H. Zia, A. Javaid, W. Lu, J. Wang, Forecasting the seasonality and trend of pulmonary tuberculosis in Jiangsu Province of China using advanced statistical time-series analyses, Infect Drug Resist. 12 (2019) 2311–2322, https://doi.org/10.2147/IDR.S207809.

[11] E.O. Nsoesie, R.J. Beckman, S. Shashaani, K.S. Nagaraj, M.V. Marathe, A simulation optimization approach to epidemic forecasting, PLoS One 8 (6) (2013), https://doi.org/10.1371/journal.pone.0067164.

[12] D. Parbat, M. Chakraborty, A python based support vector regression model for prediction of COVID19 cases in India, Chaos Solitons Fractals 138 (2020), https://doi.org/10.1016/j.chaos.2020.109942.

[13] M.H.D.M. Ribeiro, R.G. Silva, V.C. Mariani, L.S. Coelhoa, Short-term forecasting COVID- 19 cumulative confirmed cases: Perspectives for Brazil, Chaos Solitons Fractals 135 (2020), https://doi.org/10.1016/j.chaos.2020.109853.

[14] V.K.R. Chimmula, L. Zhang, Time series forecasting of COVID-19 transmission in Canada using LSTM networks, Chaos Solitons Fractals 135 (2020), https://doi.org/10.1016/j.chaos.2020.109864.

[15] G.E.P. Box, G.M. Jenkens, Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, CA, 1976.

[16] R.A. Yaffee, M. McGee, Time Series Analysis and Forecasting with Applications of SAS and SPSS, Academic Press Inc., 2000.

[17] SAS Institute Inc, SAS/ETS® 13.2 User's Guide, SAS Institute Inc., Cary, NC, 2014.

[18] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461. -46.

[19] J. Sun, Improving stepwise logistic regression using a SAS macro, J. Math. Stat. Sci. 7 (2021) 68–78.

[20] Y. Wang, C. Xu, S. Yao, Y. Zhao, Y. Li, L. Wang, X. Zhao, Estimating the Prevalence and Mortality of Coronavirus Disease 2019 (COVID-19) in the USA, the UK, Russia, and India, Infect Drug Resist. 13 (2020) 3335–3350, https://doi.org/10.2147/IDR.S265292.

[21] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723, https://doi.org/10.1109/TAC.1974.1100705.