MEDS Public Health and Preventive Medicine (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/phpm.2025.050213
ISSN 2616-1915 Vol. 5 Num. 2

Research on the Application of Social Media and Search Engine Big Data in Forecasting Major Infectious Diseases

Zongjing Liang[1,a], Zhijie Li[1,b], Yun Kuang[2,c,*]

[1]*School of Economics and Management, Guangxi Normal University, Guilin, Guangxi, China*
[2]*Library, Guilin Normal University, Guilin, Guangxi, China*
[a]*zjliang@mailbox.gxnu.edu.cn*, [b]*812811205@qq.com*, [c]*kyun@mail.glnc.edu.cn*
[*]*Corresponding author*

Keywords: COVID-19; ARDL Model; Social Media Big Data; Search Engine Big Data; Predictive Analysis

Abstract: This paper constructs a major infectious disease epidemic prediction model based on social media and search engine big data. The research object is the daily new infections of COVID-19 in China during the secondary infection peak in 2020 and 2022. The data time range is January 20, 2020-April 30, 2020 and January 2, 2022-December 25, 2022. The constructed model is the autoregressive distributed lag model (ARDL). The model dependent variable is the daily new infections, and the independent variables are Baidu Index, Weibo, news and video releases. Empirical results: The short-term effect equation shows that the information dissemination platform (such as video and Baidu) has a significant short-term impact on the prediction of the number of infections, reflecting that the public's behavior of obtaining and disseminating epidemic information through these channels has a more direct impact on the number of epidemics. The long-term cointegration equation shows that the long-term impact of video releases in 2022 has significantly increased, indicating that video platforms have played an increasingly important role in the long-term dissemination of epidemic information predictions and may become an important source of information for the public to understand and respond to the epidemic in the long term. The empirical results show that Baidu Index, Weibo, news, and video releases all play a positive role in predicting the number of new infections, but their effects vary. The conclusions of this study can provide a new research paradigm for the prediction of major infectious diseases that may occur again in the future.

1. Introduction

The novel coronavirus outbreak that began in late 2019 has spread rapidly around the world due to its unknown transmission mechanism and high infectivity. Due to its severity and urgency, the World Health Organization (WHO) declared the COVID-19 outbreak a "public health emergency of international concern" on January 30, 2020[1]. Through the joint efforts of countries around the world, the COVID-19 outbreak has been effectively suppressed. On May 5, 2023, the World Health Organization announced that COVID-19 no longer constitutes a "public health emergency of

international concern", marking the end of the global state of emergency. In its announcement, WHO emphasized that although COVID-19 is no longer considered a global emergency, the disease remains a global health threat, the virus is still spreading and mutating, and there is a risk that new variants will trigger a new peak in the number of cases and deaths [2]. It is precisely because of the serious impact of COVID-19 on society, economy, culture, etc. In order to effectively monitor and prevent the spread of similar major epidemics, it is very important to propose effective prediction methods.

Currently, there are three main models for studying major infectious diseases such as COVID-19: (1) system dynamics model, (2) econometric model, and (3) machine learning model [3]. (1) System dynamics model. The simple SIR model is more reliable than the complex SEIR model in predicting the epidemic, especially in modelling the progress of the epidemic after lockdown and quarantine measures and the secondary outbreak after resumption of work [4]. The SIR model was used to predict the daily number of COVID-19 infections in Algeria [5]. (2) Econometric model. The LSTM Markov model was proposed to improve the prediction accuracy [6]. (3) Machine learning model. Four deep learning models (LSTM, GRU, CNN, MCNN) were used to predict the epidemic trends in Brazil, Russia, and the United Kingdom [7]. The number of confirmed cases and deaths of COVID-19 in Brazil, Portugal, and the United States was predicted using artificial neural networks (ANN) [8]. By comparing the existing research results, it can be seen that although many methods have been proposed for epidemic prediction from different perspectives, the following problems still need to be further studied: (1) The method for determining the lag order is subjective. When predicting the epidemic based on big data, there is a lag characteristic between variables. However, current research generally uses a manually set method to determine the lag order of variables. This method is highly subjective and will lead to the disadvantage of excessive prediction errors. (2) There is a lack of comparative research on multiple online media prediction methods. Social media data and online search data have significant predictive value for the epidemic, but current research lacks combined research results. Based on this, this paper proposes to use the ARDL model to construct an epidemic prediction method for social media and search engine data to achieve automatic setting of the lag order, thereby achieving the purpose of improving the accuracy of epidemic prediction. The research steps of this paper are to first introduce the research data source and research methods, then solve the model, and finally discuss the calculation results.

2. Empirical research

2.1. Research method

The ARDL model is a deformation model of the standard least squares regression model, and it is also a time series analysis technique used to study the dynamic relationship between variables. This model combines the characteristics of the autoregressive model (AR) and the distributed lag model (DL), aiming to solve the problems of autocorrelation and lag effects in time series data. It contains both the explained variable (dependent variable) and its lag period, as well as the explanatory variable (independent variable) and its lag period as regression terms. The model assumes that there is a dynamic relationship between the dependent variable (Y) and the independent variable (X), which can be described by the lagged dependent variable and the lagged independent variable, so as to further analyze the short-term and long-term equilibrium relationship between the variables.

2.2. Original data

According to the statistics of the World Health Organization website, from January 11, 2020 to December 30, 2022, the spread of the number of new Covid-19 cases in China can be divided into

two stages. The first stage was from January to April 2020. After that, the epidemic was stable for a long time, and in 2022, the epidemic peak appeared again. According to the time period analysis of the actual number of epidemic cases, it is determined that the spread of the epidemic is divided into two time periods, namely January 20, 2020-April 30, 2020, and January 2, 2022-December 25, 2022. According to the availability of data, the data used in this study are daily data in 2020 and weekly data in 2022. The five variables related to COVID19 are: daily new cases (CASES), Baidu Index (BAIDU), number of Weibo (MICROBLOG), video releases (VIDEO) and news releases (NEWS). The calculation tool uses the software Eviews10.

2.3. Research conclusions

According to the solution steps of the ARDL model, the original data is first tested for stationarity, and the test method is generally to use the ADF test. After empirical calculation, it is found that all data are 0-order single integral or 1-order single integral, so the data meets the modelling requirements. Then ARDL solution is performed. After calculation, the model form in 2020 is ARDL (3, 0, 3, 0, 0), and the model form in 2022 is ARDL (3, 4, 0, 1, 4). Finally, the model stationarity test is performed. The test results show that the trend lines of the CUSUM test and CUSUMQ test of the two models are basically within the error limit at the 5% significance level, indicating that they have passed the CUSUM test and CUSUMQ test. Therefore, it can be considered that the two models have good stability and are suitable for predictive analysis. From the ARDL calculation results, conclusions can be drawn about the short-term effect relationship and long-term cointegration relationship between variables. The short-term effect and long-term cointegration results are analyzed below.

2.3.1. Short-term effect of ARDL model

Table 1: Short-term estimation results of ARDL model.

Time	2020		2022	
Variable	Coefficient	Prob.	Coefficient	Prob.
C	2.000423	0.522	2.44546	0.2125
LNCASES (1)			0.417966	0
LNBAIDU	0.243453	0.2588		
LNBAIDU (1)			0.075946	0.6981
LNNEWS			0.175986	0.4313
LNNEWS (1)	0.328999	0.1566		
LNMICROBLOG	0.130429	0.3932		
LNMICROBLOG (1)			0.222499	0.2435
LNVIDEO	0.124138	0.7453		
LNVEDIO (1)			0.642557	0.0089
D (LNCASES (1))	0.440903	0.0001	0.615145	0.0001
D (LNCASES (2))	0.201378	0.0708	0.366674	0.0355
D(LNNEWS)	0.144822	0.5606		
D (LNNEWS (1))	0.283806	0.1802		
D (LNNEWS (2))	0.376707	0.0688		
D(LNBAIDU)			0.026717	0.9266
D (LNBAIDU (1))			0.248839	0.5665
D (LNBAIDU (2))			0.012789	0.9753
D (LNBAIDU (3))			1.598586	0.002
D(LNMICROBLOG)			0.03094	0.8431
D(LNVEDIO)			0.459698	0.0731
D (LNVEDIO (1))			0.177279	0.5433
D (LNVEDIO (2))			0.068746	0.7994
D (LNVEDIO (3))			1.225467	0.0003

Note: In the table, LN means taking the logarithm of each variable, and D means the difference term.

The short-term effect of the ARDL model shows the short-term estimation results of the ARDL model in 2020 and 2022, describing the short-term impact of the number of cases (CASES), Baidu Index (BAIDU), news releases (NEWS), Weibo releases (MICROBLOG) and video releases (VEDIO) on the number of epidemics. The following is a detailed comparison and analysis of the short-term estimation results in 2020 and 2022. The short-term estimation results of the ARDL model are shown in Table 1.

From Table 1, it can be found that in the short-term estimation equation of ARDL in 2020, the coefficients of all explanatory variables are positive, which proves that Baidu Index, news release volume, Weibo release volume and video release volume have a positive effect on the number of infections. That is, the more attention the search index and social media pay, the higher the number of infections, indicating that the search index and social media play a predictive role in the spread of the epidemic. However, the impact of each variable on the epidemic is different. Only the coefficient of variable D (LNCASES (1)) in the equation is statistically significant. D (LNCASES (1)) is the first difference term of LNCASES (1), which indicates the impact of short-term changes in the number of cases on the dependent variable. The corresponding p-value is very small (<0.05), indicating that the variable has a significant impact on the dependent variable in the short term. The positive value of the coefficient means that the increase in the number of cases will have a positive impact on the dependent variable.

For the estimated equation for 2022, the variable LNCASES (1) (logarithm of the number of cases) has a significant effect on the dependent variable, with a p-value of 0, indicating that the variable has a very significant long-term effect on the dependent variable. Each unit increase in LNCASES (1) will increase the dependent variable by 0.417966. LNVIDEO (1) (logarithm of the video index) has a significant effect on the dependent variable, with a corresponding coefficient of 0.642557, indicating that each unit increase in LNVIDEO (1) increases the dependent variable by 0.642557, and a p-value of less than 0.05, indicating that the variable has a significant positive effect on the dependent variable. The p-value of D (LNCASES (1)) is also very small, indicating that short-term changes in LNCASES (1) have a significant effect on the dependent variable. This is consistent with the results in 2020, indicating that changes in LNCASES (1) continue to have an impact on the dependent variable in the short term. D (LNCASES (2)) also has a significant impact on the dependent variable in 2022, with a p-value less than 0.05, which means that the second-period change of LNCASES (2) has a significant short-term impact on the dependent variable in 2022. In addition, the coefficient values of variables D (LNBAIDU (3)) and D (LNVEDIO (3)) are positive and statistically significant, proving that the search index and the number of video releases play a positive role in predicting the spread of the epidemic in the short term.

In summary, from 2020 to 2022, the long-term impact of LNCASES (1) increased significantly, indicating that the impact of changes in the number of cases on the dependent variable has changed significantly between these two years. The video release variable shows an increase in its influence in 2022, which may be related to the growth of the epidemic in 2022 due to the spread of video-related content.

2.3.2. Long-term cointegration relationship of the ARDL model

The results of the long-term cointegration relationship of the ARDL model are shown in Table 2.

Table 2: Long-term estimation results of the ARDL model.

	2020		2022	
Variable	Coefficient	Prob.	Coefficient	Prob.
LNBAIDU	2.362871	0.1287	0.181704	0.7013
LNNEWS	3.19315	0.3634	0.421052	0.4332
LNMICROBLOG	1.265906	0.4734	0.53234	0.2731
LNVIDEO	1.204845	0.751	1.537342	0.0006
C	19.4155	0.4086	5.85085	0.1476

The long-term cointegration relationship can show the long-term estimation results of the ARDL model in 2020 and 2022, describing the long-term impact of variables such as Baidu Index (LNBAIDU), news release volume (LNNEWS), Weibo release volume (LNMICROBLOG), and video release volume (LNVEDIO) on the number of epidemics. The following is a detailed analysis and comparison of the two-year long-term estimation results.

(1) LNBAIDU (Baidu Index). The coefficient in 2020 is 2.362871, the P value is 0.1287, and the coefficient is not significant, indicating that the long-term impact of Baidu Index on the number of epidemics is not significant. The coefficient in 2022 is 0.181704, the P value is 0.7013, and it does not reach the significance level, indicating that the long-term impact of Baidu Index on the number of epidemics in 2022 is not significant.

(2) LNNEWS (news release volume). The coefficient in 2020 is 3.19315, and the P value is 0.3634, which does not reach the significance level, indicating that the long-term impact of news releases on the number of epidemics is not significant. The coefficient in 2022 is 0.421052, and the P value is 0.4332, which still does not reach the significance level, indicating that the long-term impact of news releases on the number of epidemics is not significant.

(3) LNMICROBLOG (microblogging). The coefficient in 2020 is 1.265906, and the P value is 0.4734, which does not reach the significance level, indicating that the long-term impact of microblogging releases on the number of epidemics in 2020 is not significant. The coefficient in 2022 is 0.53234, and the P value is 0.2731, indicating that the long-term impact of microblogging releases on the number of epidemics in 2022 is not significant.

(4) LNVEDIO (video releases). The coefficient in 2020 is 1.204845, and the P value is 0.751, which does not reach the significance level, indicating that the long-term impact of video releases on the number of epidemics in 2020 is not significant. The coefficient in 2022 is 1.537342, and the P value is 0.0006, which is a significant positive correlation, indicating that the long-term impact of video releases on the number of epidemics in 2022 is significant, and the increase in video releases is related to the long-term increase in the number of epidemics.

(5) C (constant term). 2020: The constant term is 19.4155, and the P value is 0.4086, which does not reach the significance level. 2022: The constant term is 5.85085, and the P value is 0.1476, which still does not reach the significance level.

Overall summary: The long-term impact of Baidu Index and news releases on the number of epidemics in the two years is not significant, indicating that these two information channels have no significant direct effect on the prediction of the number of epidemics in the long term. Weibo releases also have no significant long-term impact in the two years. The long-term impact of video release volume in 2022 has significantly increased, indicating that video platforms have occupied an increasingly important position in the long-term dissemination of epidemic information, and may become an important source of information for the public to understand and respond to the epidemic in the long term. The main difference between 2022 and 2020 is the number of video releases, which has a significant and strong long-term impact on the number of epidemics in 2022, while the long-term impact of other variables has not changed much.

3. Conclusions

In order to improve the accuracy of epidemic prediction, this paper constructs an epidemic prediction method based on the ARDL model (i.e., autoregressive distributed lag model) by integrating multi-source big data (such as Baidu, Weibo, news, and videos), thereby revealing the short-term fluctuations and long-term cointegration relationship between epidemic spread and multi-source big data. The research time period of this paper is divided into two periods, namely 2020 and 2022, corresponding to two peak periods of epidemic infection. The epidemic data is the number of new cases per day, the search engine big data uses the Baidu Index, and the social media

big data uses the number of Weibo, news, and video releases. The research results show that social media and search engine big data have short-term effects and long-term cointegration relationships on epidemic prediction. The short-term effect equation shows that the lagged term of the number of cases in 2020 has a significant positive callback effect in the short term, while the short-term effects of other variables (Baidu Index, news releases, Weibo and video releases) are not significant. The lagged term of the number of cases in 2022 still shows a short-term positive callback effect, and the short-term positive effect of video releases on the number of epidemics becomes significant, indicating that the importance of video platforms in the dissemination of epidemic information has increased. By comparison, it is found that the short-term impact of information dissemination platforms (such as videos and Baidu) has become more significant in 2022 compared with 2020, reflecting the impact of the public's behavior of obtaining and disseminating epidemic information through these channels on the prediction of the number of epidemics. The long-term cointegration equation shows that the long-term impact of Weibo and news releases on the prediction of the number of epidemics in two years is not significant, indicating that these two information channels have no significant effect on the prediction of the number of epidemics in the long term. Weibo releases also have no significant long-term impact in two years. The long-term impact of video releases in 2022 has significantly increased, indicating that video platforms have played an increasingly important role in the long-term dissemination of epidemic information forecasts, and may become an important source of information for the public to understand and respond to the epidemic in the long term. The main difference between 2022 and 2020 is the number of videos released, which has a significant and strong long-term impact on the number of epidemics in 2022, while the long-term impact of other variables has not changed much. The conclusions of this study can provide a new research paradigm for the prediction of major infectious diseases that may occur again in the future.

Acknowledgements

The work was supported by National Social Science Foundation of China" Spatio-temporal Model, Monitoring, Prevention and Control of the Spread of Major Infectious Diseases Based on Big Data" (Project No.: 20XTJ004).

References

[1] World Health Organization. WHO Director-General's statement on IHR Emergency Committee on Novel Coronavirus (2019-nCoV) [EB/OL]. https://www.who.int/director-general/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-(2019-ncov), 2020-01-30 [2024-05-05].

[2] World Health Organization. Statement on the 15th meeting of the IHR Emergency Committee on the COVID-19 pandemic [EB/OL]. https://www.who.int/news/item/05-05-2023-statement-on-the-15th-meeting-of-the-ihr-emergency-committee-on-the-covid-19-pandemic, 2023-05-05 [2024-05-05].

[3] RAHIMI I, CHEN F, GANDOMI A H. A review on COVID-19 forecasting models [J]. Neural Computing and Applications, 2023, 35(33): 23671-23681.

[4] RODA W C, VARUGHESE M B, HAN D, et al. Why is it difficult to accurately predict the COVID-19 epidemic? [J]. Infectious disease modelling, 2020, 5(1): 271-281.

[5] Boudrioua M S, Boudrioua A. Predicting the COVID-19 epidemic in Algeria using the SIR model [J]. Medrxiv, 2020, 2020.2004. 2025.20079467.

[6] MA R, ZHENG X, WANG P, et al. The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method [J]. Scientific Reports, 2021, 11(1): 1-14.

[7] NABI K N, TAHMID M T, RAFI A, et al. Forecasting COVID-19 cases: A comparative analysis between Recurrent and Convolutional Neural Networks [J]. Results in Physics, 2021, 24(1): 104137.

[8] De Oliveira L S, Gruetzmacher S B, Teixeira J P. COVID-19 Time Series Prediction [J]. Procedia Computer Science, 2021, 181(1): 973-980.