

Influencing Factors and Mortality Prediction of Covid-19 Contact Infection Rate in China

Xinping Yang¹, Wei Zheng¹, Yunyuan Yang², Jie Zhang³

1. School of mathematics and computer science, Chuxiong Normal University, Chuxiong, China

2. School of Geography and Tourist Management, Chuxiong Normal University, Chuxiong, China

3. Chuxiong Medical College, Chuxiong, China

[†]Email: zw@cxtc.edu.cn

Abstract

Taking the COVID-19 data from 2020-1-23 to 3-21 days released by the China Health Protection Committee as the sample, the hospital remaining rate, mortality rate and cure rate are selected as covariates, and the contact infection rate is used as response variable to establish a high precision ADL model, results of return substitution show that the predicted value of contact infection rate almost coincides with the sample value, and shows three stages of change characteristics. After March 1, 2020, the overall trend is downward, stable below 12%. Main influencing factors of contact infection rate are analyzed quantitatively. Based on this, the ARIMA (1, 2, 2) model is established to analyze and predict the mortality change trend. The results showed that the domestic COVID-19 mortality rate is stable near 4 % after 2020-3-27.

Keywords: COVID-19; ADL Model; ARIMA Model; Contact, Infection Rate; Mortality Rate

1 INTRODUCTION

As of March 25, 2020, 81960 cases were diagnosed in COVID-19 and 354160 cases overseas. The outbreak has exploded in Italy, the United States, Spain, Germany, France, Iran and other places. Quantitative study on the contact infection rate and crude mortality rate of covid-19 in China can provide reference for blocking the import of covid-19.

Public reports: as of March 21, 2020, the infection rate of close relatives in Beijing reached 17%, and that of family reached 86% [1]. Researchs related to infection and mortality are divided into three categories: (1) Descriptive research of epidemic data. Li[2] made statistics on 425 confirmed cases, and pointed out that the incubation period was 5.2 days, the doubling time was 7.4 days, and the regeneration number was about 2.2 days. China's Center for Disease Control and prevention, COVID-19 emergency response mechanism Epidemiology Group [3] analyzed the age, gender and residence history of 72314 cases of four cases up to 2020-2-11, revealing that the crude mortality rate was 2.3% (SARS 9.6%). In addition, the infection treatment of children [4], pregnant women [5] and other special groups of patients has also been concerned. For example, in literature 4366 children diagnosed with covid-19 in Wuhan from January 7, 2020 to January 15, 2020 were analysed, and it was found that in the early stage of the epidemic, children had been infected with new coronavirus. (2) Statistical inference of epidemic indicators. Mainly around the scale of infection, the number of regenerations, doubling time and so on. More SEIR or modified SEIR model were often used in research [6-8]. Such as, based on SEIR optimization model and AI, the epidemic outbreak forecast under public health intervention was carried out in document 8. It was predicted that the epidemic peak would reach the end of February in China, and gradually decline at the end of April, and the final infected people would reach 120000, the prediction results have a certain degree of recognition. (3) Other methods, the epidemic trend and incidence rate were predicted by ARIMA model [9]. Joseph T Wu et al [10] used MCMC method to estimate that the regeneration number of Wuhan covid-19 was 2.68, and the epidemic doubling time was 6.4 days. Adam J Kucharski et al [11] used random transmission model to estimate that the average number of basic

regenerations in a week before and after Wuhan City was reduced from 2.35 to 1.05, and the infection may decrease in late January. In addition, there are maximum likelihood method [12], bootstrap method [13], etc. In the above study, literature 2 was comprehensive based on case data analysis, literature 6-7 used the same method for the same research area (Hubei and Wuhan) and obtained different parameter settings; Document 11 was very different from the actual situation.

On the basis of the epidemic data, the distribution lag model (ADL) was established to analyze the main influencing factors of the contact infection rate and predict the contact infection rate; In order to estimate the future mortality rate, the autoregressive moving average model (ARMA) was established based on the data of mortality rate, so as to provide reference for the prevention of epidemic situation in China and overseas.

2 DATA AND METHODS

Collect the epidemic data released by China Health Commission from January 20, 2020 to March 21, These include: cumulative number of close contacts tracked (recorded as cc_t). cumulative number of confirmed cases reported (recorded as nc_t). Number of close contacts under medical observation (recorded as mo_t). Cumulative number of deaths (recorded as nd_t). Cumulative number of cured and discharged (recorded as cu_t). Because the number of new cures from January 20, 2020 to January 22 bulletin is 0, so 59 samples of the bulletin data from January 23, 2020 to March 21 are used for analysis. In order to find the quantitative relationship between the above indicators, the structural relative numbers which can reflect the changes of epidemic situation were defined respectively (The ratio of partial to total values in a certain range, which reflects the internal composition of the range). (1) contact infection rate (inr_t) = nc_t/cc_t , (2) Hospital observation rate (mos_t) = mo_t/cc_t , (3) Mortality (fr_t) = nd_t/nc_t , (4) Cure rate (cr_t) = cu_t/nc_t . (1) (3) respectively reflect the epidemic situation and the severity of death, (2) (4) reflect the hospital observation and confirmed treatment. These indexes are the frequency indexes of binary variables which are statistically to subordinate binomial distribution. The Bernoulli law of large numbers shows that each frequency is stable at the corresponding probability [14].

Through the correlation analysis, finding the internal or mutual influence relationship of the above four indicators. Taking the contact infection rate as the response variable and the other three indicators and their lag terms as the explanatory variables, a distributed lag model is established [15], main factors affecting the contact infection rate are analysed, and the autoregressive moving average model is established to estimate the trend of mortality. Using software is Eviews and R language.

3 MAIN INFLUENCING FACTORS OF CONTACT INFECTION RATE

3.1 Correlation Analysis of Indicators

In general, it is an effective method to select explanatory variables combined with qualitative analysis results and correlation coefficients. The four variables inr_t , fr_t , mos_t and cr_t all have lag term, so the workload of model selection and test is large. In order to avoid the complexity of variable selection, the correlation coefficient between them and the autocorrelation coefficient of each variable are calculated respectively (see Table 1).

TABLE 1 INR_T , FR_T , MOS_T AND CR_T CORRELATION COEFFICIENTS (LEFT) AND RESPECTIVE LAG CORRELATION COEFFICIENTS (RIGHT)

	inr_t	fr_t	mos_t	cr_t	1	2	3	4	5	6	7	8	9
inr_t	1.00000	0.51722	-0.87616	-0.85050	0.94	0.86	0.79	0.72	0.66	0.60	0.53	0.46	0.38
fr_t	0.51722	1.00000	-0.79485	0.93692	0.97	0.94	0.90	0.85	0.81	0.76	0.70	0.65	0.59
mos	-0.87616	-0.79485	1.00000	0.84836	0.97	0.90	0.88	0.83	0.77	0.71	0.65	0.59	0.53
cr_t	-0.85050	0.93692	0.84836	1.00000	0.97	0.93	0.89	0.85	0.81	0.76	0.71	0.66	0.60

According to table 1, the correlation between contact infection rate and mortality rate is weak (correlation coefficient is only 0.51722), indicating that the contact infection rate may only be affected by the same period and the smaller

lag term of mortality. There is a strong negative correlation between contact infection rate and hospital observation rate (correlation coefficient was - 0.87616), the results show that stay in hospital waiting for observers are not all infected, it takes time to exclude the cases of non-crown pneumonia, to pass through the fluorescence quantitative PCR of virus nucleic acid sequence and Ig M dual detection. So there is a negative correlation. The patients who were cured and discharged from hospital still need to be observed in isolation, which also shows a negative correlation (correlation coefficient was-0.87616). Autocorrelation coefficients of contact infection rate, mortality rate, hospital observation rate and cure rate gradually decrease with the increase of lag time. The autocorrelation coefficients of the first seven stages of the first three indicators are greater than 0.6, and the autocorrelation coefficients of the first nine stages of cure rate are greater than 0.6. There is a high degree of autocorrelation. In modelling, the lag term with autocorrelation coefficient greater than 0.60 is used as an alternative explanatory variable for variable selection.

3.2 Establishment and Solution of Distributed Lag Model (ADL model)

After determining the model form, the forward selection variable method is usually adopted to choose the appropriate explanatory variable, in order to prevent the over fitting of the model. That is to say, it starts from zero model (only including intercept term). Entering a variable at a time, the parameter estimation is completed by least squares, the criterion leaving and staying variable determines whether the variable remains in the model. These criteria are as follows: (1) Significance criterion. After entering the variable, t-test is significant, F-test is significant. (2) Model goodness criteria. Akaike and Schwarze information, Han nan Quinn criterion. (3) Criterion of determinable coefficient. Combining the forward selection variable method with the criterion of leaving and staying, the following model is established to describe the quantitative relationship among the four indexes. The model parameter estimation and various statistical evaluation indexes are shown in Table 2. corresponding model expression is as follows:

$$inr_t = 27.49760 - 5.388708fr_t + 1.886486fr_{t-1} - 0.016304cr_t - 0.027068mos_t - 0.09173mos_{t-7} + \varepsilon_t \quad (1)$$

Where ε_t is white noise with 0 mean and σ^2 variance.

TABLE 2 SUMMARY OF PARAMETER ESTIMATION AND STATISTICS OF DISTRIBUTION LAG MODEL

Variable	Coefficient	Std. Error	t-Statistic	Prob
C	27.497600	0.557989	49.279820	0.000000
fr_t	-5.388708	0.352442	-15.289620	0.000000
cr_t	-0.016304	0.002728	-5.975827	0.000000
fr_{t-1}	1.886486	0.357888	5.271165	0.000000
mos_t	-0.027068	0.003274	-8.268027	0.000000
mos_{t-7}	-0.09173	0.004496	-21.613790	0.000000
R-squared	0.988688	S.D dependent		1.246935
Adjusted R-squared	0.987459	Akaike infonation criterion(AIC)		-0.991293
F-statistic	804.1016	Schwarze criterion(SC)		-0.76615
Prob(F-statistic)	0.000000	Huannan Quinn criterion(HQC)		-0.904979
S.E.of regression	0.139643	Duibin-Watson(DW)		2.373393

According to table 2: R^2 is 0.988688, the concomitant probability of t-test and F-test are all 0 after the fourth decimal point, all significant, and values of AIC, SC, HQC and SE are close to zero, SD = 1.246935 closes to 1. Under 59 sample data, Du bin Watson (DW) table shows that the critical values d_U of 2.5% and 1% were 1.69 and 1.60 respectively, and $d_U < DW = 2.373393 < 4 - d_U$, indicating that the residual is white noise and the model performs well. The model (1) has a larger F value and the largest decidability coefficient R^2 compared with the other given up models, and the superiority of the six indexes is obviously improved.

The predicted value (inrf_t) of contact infection rate are obtained from the model (1), and the curve is drawn (see Figure 1 (left)). standard deviation of ε_t is estimated to be $\hat{\sigma} = 0.135398$, the dynamic confidence interval of 95% for contact infection rate is: $\text{inrf}_t \pm 2\hat{\sigma}$ (see Figure 2, figure 1 (right)).

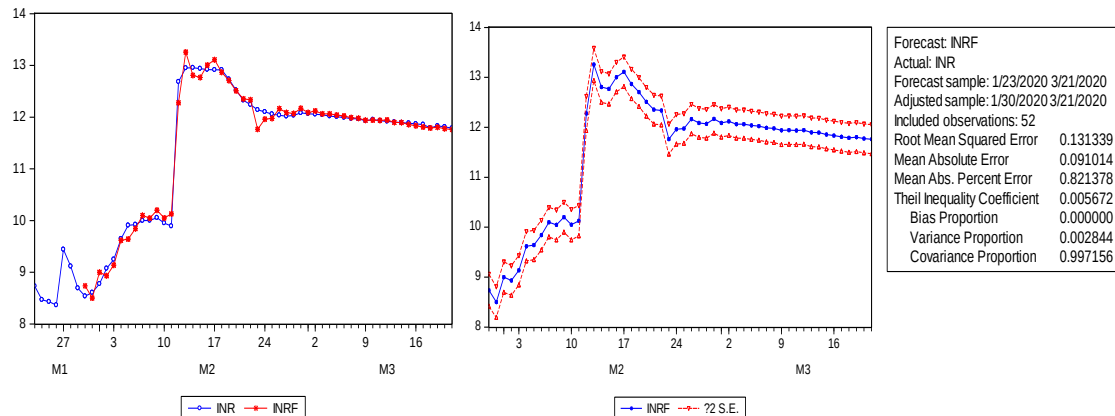


FIG. 1 THE CURVE OF CONTACT INFECTION RATE(INR_T) A AND PREDICTIVE VALUE (INRF_T) (LEFT) AND ITS 95% CONFIDENCE INTERVAL (RIGHT)

The mean absolute error (MAPE), Hill's inequality coefficient (The il), deviation rate (BP), variance rate (VP) and covariance rate (CP = 1-BP-VP) of the model are 0.131339%, 0.005672, 0.000000, 0.002844 and 0.997156, respectively, which indicated that the model had a high prediction accuracy. It was also confirmed that the curve of aggregated infection rate was highly coincident with its predicted value.

3.3 Analysis of Influencing Factors

The model (1) shows that the coefficients of fr_t , cr_t , mos_t and mos_{t-7} are negative, which can promote the control of contact infection. If other conditions remain unchanged, the current mortality rate will increase by 1%, which will reduce the possibility of infection by 5.3887085%, This indicates that the dead host significantly reduces the possibility of infection to the contact. There is 1% increase in the mortality of the lag 1 period, which would increase the possibility of infection by 1.886486%, It shows that the critical patients have strong infectivity and need strict isolation treatment. The cure rate increased by 1% and the possibility of infection decreased by 0.016304%. There is 1% increase in the observation rate of the current period and the observation rate of the lag 7 periods, which reduced the possibility of infection to the contact by 0.027068% and 0.09173% respectively, it shows that screening and observation in hospital is an important means to effectively prevent contact infection. If the observation time reaches 7 days, the possibility of contact infection will be reduced by nearly 0.1%. Therefore, blocking the source of disease (dead cases and critical cases) and strict observation in hospital are effective means to control the expansion of the epidemic.

The mortality (fr_t) and the lag 1 period of mortality(fr_{t-1}) in model (1) have a great influence on the contact infection rate, which indicates that the critical cases require higher infection protection than the mild cases, and the COVID-9 death cases should be dealt with immediately and properly. At the same time, further discussion of mortality rates is particularly important.

4 MORTALITY PREDICTION

4.1 ARIMA Model

For the purpose of establishing a mortality prediction model, the ADF stability test of the mortality sequence is carried out, and the test results are shown in the table (Table 3). The results show that: after the second-order difference, the mortality sequence is stable. According to the autocorrelation and partial correlation coefficients of the second-order difference mortality sequence (d^2fr), ARIMA (1, 2, 2) model is used to describe the variation characteristics of the mortality. The summary table of parameter estimation and statistics is shown in table 4, The corresponding expressions are:

$$(1 + 0.479044B)(1 - B)^2 fr_t = (1 - 0.287791B + 0.697280B^2)\varepsilon_t \quad (2)$$

Where: $B(x_t) = x_{t-1}$, $B^k(x_t) = x_{t-k}$, ε_t is white noise with 0 mean and σ^2 variance. There is no correlation between different periods ($E\varepsilon_s\varepsilon_t = 0$, $s \neq t$), $\hat{\sigma} = 0.0723964$.

The equivalent model (2) is obtained after deformation:

$$fr_t = 1.520956fr_{t-1} + 0.4790432fr_{t-2} - 0.479044fr_{t-3} - 0.287791\varepsilon_{t-1} + 0.697280\varepsilon_{t-2} + \varepsilon_t \quad (3)$$

TABLE 3 STATIONARITY TEST STATISTICS AND RESULTS SUMMARY OF MORTALITY SERIES UNDER THREE STANDARDS

series	ADF Test Statistic	1%Critical	5%Critical	10%Critical	result
fr_t	-0.078654	-3.550396	-2.913549	-2.594521	Not significant
dfr_t	-3.254251	-3.555023	-2.915522	-2.595565	1% standard not significant
d^2fr_t	-6.851645	-3.555023	-2.915522	-2.595565	significant

TABLE 4 PARAMETER ESTIMATION AND STATISTICS OF ARIMA (1,2,2) MODEL

Variable	Coefficient	Std. Error	t-Statistic	Prob
ar(1)	-0.479044	0.077642	-6.169899	0.000000
ma(1)	0.287791	0.102347	2.811907	0.006900
ma(2)	-0.697280	0.101872	-6.844639	0.000000
R-squared	0.716958	Mean dependent var		0.006346
Adjusted R-squared	0.706277	S.D dependent var		0.136875
S.E. of regression	0.074181	Akaike info criterion(AIC)		-2.312527
Log likelihood	67.75077	Schwarz criterion(SC)		-2.204026
Durbin-Watson(DW)	1.881105	Hannan Quinn criterion(HQC)		-2.270462

The reciprocal roots of the coefficient polynomials of ARIMA model (3) are -0.48, 0.70 and -0.99 respectively, which fall in the unit circle. R^2 is 0.716958, t- test concomitant probability is far less than 0.05, all significant, AIC, SC and HQC are around - 2.2, $\hat{\sigma}^2 = 0.005241239$ is close to 0, indicating that the model fitting is good.

4.2 Mortality Prediction

ARMA model is usually only suitable for short-term prediction [16]. Based on the model (3) to complete the mortality returned prediction, the prediction of the next week (March 21, 2020 to March 28) is carried out. The results are shown in figure 2. The curve (frf) calculated by model (3) almost coincides with the curve (fr) calculated by the sample. Contact infection rate calculated by February 12, 2020 sample (see Figure 1 (left)) is a jumping breakpoint, resulting in a slightly larger prediction error (0.43%) of February 12. Mortality is changes in four stages. From January 27 to February 5 in 2020 is lower than 2.997501%, and the lowest is 2.013268% of February 5 in 2020. In this stage, the cumulative number of confirmed cases (denominator) increases rapidly, leading to the gradual decrease of the mortality rate in this stage. After February 5, 2020, the mortality increases gradually, and the clinical cases on February 12 are incorporated into the confirmed cases, which lowers the mortality again; The growth rate of mortality from February 6 to February 24 in 2020 is faster than that from February 25 to March 9. After March 9, 2020, the mortality increases slowly. It shows that with a large number of mild patients discharges from hospital, medical resources are concentrated on severe and critical patients, and the treatment effect is obvious. After March 28, 2020, the mortality of domestic covid-19 is stable at around 4%(See Figure 2).

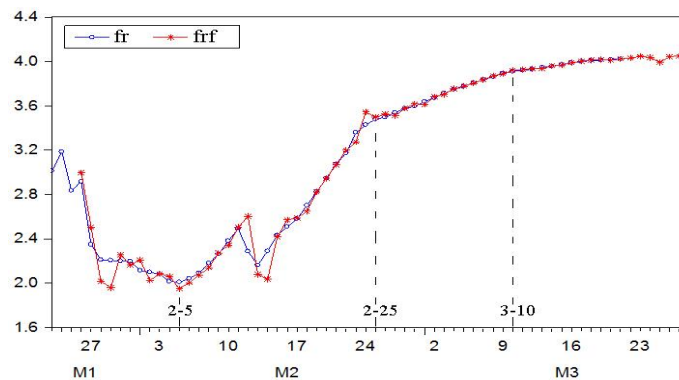


FIG. 2 LINE DIAGRAM OF MORTALITY (FR) AND MORTALITY PREDICTION (FRF)

According to the commune data, as of March 25, 2020, the average mortality in China was 4.02%, and the average overseas death rate is 4.62%. Over the same period, the death rate in Italy is more than 10%, follows by Spain and Iran.

The ADL model and the ARIMA model are used to analyze the contact infection rate and mortality quantitatively, and the preliminary conclusion is obtained. The purpose is to provide a way of thinking for the analysis of COVID-19 indicators, we hope colleagues can put forward better solutions.

ACKNOWLEDGMENT

This research was funded by "Analysis of the Influence Mechanism of Modern Service Industry in Yunnan Province Based on Bayes Method" on the Project of Yunnan University Joint Fund. (2017FH001-068).

REFERENCES

- [1] Pang xinghuo. Beijing center for disease prevention and control [EB/OL]. https://www.sohu.com/a/382449846_108794. 2020-3-22.
- [2] Qun Li, M.Med., Xuhua Guan.Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia[J]. The New England Journal of Medicine.2020. doi: 10.1056/NEJMoa2001316.
- [3] Chinese center for disease control and prevention, Covid-19 emergency response mechanism Epidemiology Group.Epidemiological characteristics of new coronavirus pneumonia [J]. Chinese journal of Epidemiology, 2020, 41(2): 145-151.
- [4] Lan Ke,Liu YingLe, Yu Hong Jie, et al. Detection of Covid-19 in Children in Early January 2020 in Wuhan, China[J]. The New England Journal of Medicine, 2020. DOI: 10.1056/NEJM189002061220607.
- [5] Chen xu, LI yang,Wang jinxi, Cai hong liu, Cao hong cui, Sheng ji fang. Pregnant women complicated with COVID-19: a clinical analysis of 3 cases[J]. Journal of Zhejiang university(medical sciences), 2020. DOI: 10.3785/j. issn. 1008-9292.2020.02.05.
- [6] Tang biao, Wang xia, Li qian. Estimation of the Transmission Risk of the 2019-nCov and Its Implication for Public Health Interventions[J]. Journal of Clinical Medicine, 2020. DOI: 10.3390/jcm9020462.
- [7] Cao sheng li, Feng pei hua, Shi peng peng. Study on the epidemic development of COVID-19 in Hubeiprovince by a modified SEIR model[J]. Journal of Zhejiang university(medical sciences), 2020. DOI:10.3785/j.issn.1008-9292.2020.02.05.
- [8] Yang zi feng, Zeng zhi qi, Wang ke, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions[J]. Journal of Thoracic Disease, 020. doi: 10.21037/jtd.2020.02.64.
- [9] Domenico Benvenuto, Marta Giovanetti, Lazzaro Vassallo, et al. Application of the ARIMA model on the COVID-2019 epidemic dataset[J]. Data in brief, 2020. DOI: 10.1016/j.dib.2020.105340.
- [10] Joseph T Wu, Kathy Leung, Gabriel M Leung, Nowcasting and forecasting the potential domestic and international spread of the COVID-19 outbreak originating in Wuhan, China: a modelling study[J]. The Lancet,2020, 395(10225): 689-697.
- [11] Kurland, Philip B., and Ralph Lerner, eds. The Founders' Constitution. Chicago: University of Chicago Press, Lattimore, Richmond, trans. The Iliad of Homer. Chicago: University of Chicago Press, 1951
- [12] Zhao shi, Salihu S. Musa, Lin qian ying,et al. Estimating the Unreported Number of Novel Coronavirus (COVID-19) Cases in China in the First Half of January 2020: AData-Driven Modelling Analysis of the Early Outbreak[J]. Journal of Clinical Medicine,2020, 9, 38:1-6.

- [13] Yang zheng, Yuan zi xia, Jia zu yao. Estimating the number of people infected with New Coronavirus in Wuhan based on migration data[J]. Journal of University of Electronic Science and Technology of China,2020. DOI: 10.12178/1001-0548.2020030.
- [14] Zhang yan, Cheng shizhen. Probability theory and mathematical statistics (2nd Edition) [M]. Beijing:T singhua university press,2017,8:107-11.
- [15] Yi dan hui. Data analysis and Eviews application[M]. Beijing: China renmin university press, 2008, 10:32-45,135-143.
- [16] Wang Yan. Time series analysis based on R[M]. Beijing: China renmin university press, 2015, 9: 98-104.