Applying The Spatial Transmission Network to the Prediction of Infectious Diseases Across Multiple Regions

Huimin Wang

Sichuan University

Jianqing Qiu

Sichuan University

Cheng Li

Sichuan University

Hongli Wan

Sichuan University

Changhong Yang

Sichuan Center for Disease Control and Prevention

Tao Zhang (✉ statzhangtao@scu.edu.cn)

Sichuan University

Research Article

Keywords: spatial transmission network, infectious disease, VARMA, time series analysis

Posted Date: April 5th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-369262/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Applying the spatial transmission network to the prediction
of infectious diseases across multiple regions**

Huimin Wang¹, Jianqing Qiu¹, Cheng Li¹, Hongli Wan¹,
Changhong Yang², Tao Zhang^{1*}

¹ Department of Epidemiology and Health Statistics, West China School of Public
Health and West China fourth Hospital, Sichuan University, Sichuan, China

² Sichuan Center for Disease Control and Prevention, Chengdu, Sichuan, China

*Corresponding author

Tao Zhang

Renmin South Road 3rd Section NO.16, Chengdu 610041, Sichuan

Department of Epidemiology and Health Statistics, West China School of Public

Health and West China fourth Hospital, Sichuan University

Email: statzhangtao@scu.edu.cn (TZ)

Tel number: +8615908141271

Abstract

Background: Timely and accurately forecasting of the infectious diseases is essentially important for achieving precise prevention and control. A good forecasting method of infectious diseases should have the advantages of *interpretability*, *feasibility* and *forecasting performance*. Since our previous research had illustrated that the spatial transmission network showed good *interpretability* and *feasibility*, this study further explored its *forecasting performance* for the infectious diseases across multiple regions.

Methods: Under the topological framework of spatial transmission network, the vector autoregressive moving average (VARMA) model was built in a systematic way for parameter learning. Moreover, we utilized the prediction function of the VARMA model to further explore the *forecasting performance* of the spatial transmission network. The fitting and forecasting performance of the spatial transmission network were subsequently evaluated by comparing the accuracy and precision with the classical autoregressive moving average (ARMA) model. The influenza-like illness (ILI) data in Chengdu, Deyang and Mianyang of Sichuan Province from 2010 to 2017 were used as an example for illustration.

Results: ① The estimated spatial transmission network revealed that the influenza may probably spread from Chengdu to Deyang during the study period. ② For fitting accuracy, the spatial transmission network had different fitting performance for each city. The spatial transmission network performed slightly worse than the ARMA model in Deyang, but had better fitting performance in the other two cities. ③ For forecasting accuracy, the spatial transmission network outperformed the ARMA model

by at least 1% for both mean absolute error (MAE) and mean absolute percentage error (MAPE). ④ The forecasting standard errors of the spatial transmission network were smaller than those of the ARMA model.

Conclusions: This study applied the spatial transmission network to the prediction of infectious diseases across multiple regions. The results illustrated that the spatial transmission network not only had good accuracy and precision in *forecasting performance*, but also could indicate the spreading directions of infectious diseases among multiple regions to a certain extent. Therefore, the spatial transmission network is a promising candidate to improve the surveillance work.

Keywords: spatial transmission network; infectious disease; VARMA; time series analysis

1 Background

The outbreak of infectious diseases poses a serious threat to public health and imposes a heavy burden on the world economy. According to a latest research, there are around 291,000 to 646,000 deaths caused by seasonal influenza-related respiratory illnesses worldwide every year ^[1-2]. The US Centers for Disease Control and Prevention (CDC) estimated that from October 1st, 2019 to April 4th, 2020, there were 39,000,000 to 56,000,000 flu infections and 24,000 to 62,000 flu deaths in the US ^[3]. In addition, the World Health Organization (WHO) assessed that the first influenza pandemic in this century damaged the global economy by around \$800 billion ^[4].

In view of the severe threats of infectious diseases, improving the infectious diseases surveillance work consistently is on the top priority for health promotion ^[5]. How to utilize the surveillance data to achieve accurate and reliable forecasting and early warning in the early stage of infectious diseases is not only an important basis for formulating corresponding prevention and control strategies, but also a main research focus at present. Furthermore, among the frontiers of current researches, the core and key lie in how to accurately grasp the spatio-temporal transmission trends and rules of infectious diseases, and reveal the spatial dynamic transmission routes of diseases over time in order to achieve precise prevention and control of infectious diseases and invest limited resources in the most needed places. This naturally puts forward high demands for statistical methods. For a good forecasting method of infectious diseases, it should have at least the following three advantages: ① *Interpretability*, that is, the model reflects the epidemic features of infectious diseases in order to guarantee the forecasted results to make sense in practice; ② *Feasibility*, for example, supported statistical software (such as SPSS, SAS and R), required data and applicable diseases types, etc; ③ *Forecasting performance*, the inherent ability of the model to infer the future epidemics of infectious diseases.

In our previous research ^[6], the spatial transmission network was proposed to show good *interpretability* and *feasibility* for the transmission mechanism of diseases. Its *interpretability* lied in the sufficient consideration of the spatio-temporal effects of epidemic and actual process of the prevention and control of infectious diseases, as well as the potential ability to reveal the direction and time-lagged infectious diseases

transmission. As for the *feasibility*, the spatial transmission network only needed data of administrative infectious diseases surveillance system, so it was convenient for practice. The spatial transmission network was essentially a set of the time-lagged association among infectious diseases surveillance data across different places, and its construction consisted of structure learning and parameter learning. The vector autoregressive moving average (VARMA) model was adopted for parameter learning. The VARMA model has been proved to possess good mathematical properties in theory (such as identifiability, stationarity and inevitability) ^[7], and has been successfully applied in financial, economic and other fields ^[8]. However, in our previous research, the VARMA model was only used to estimate the strength of the association among infectious diseases surveillance data in different regions and time. We did not further apply the VARMA model to make predictions. Therefore, the purpose of this study was to utilize the prediction function of the VARMA model to further explore the *forecasting performance* of the spatial transmission network for the infectious diseases across multiple regions.

The VARMA model is a multivariate extension of the classical time series autoregressive moving average (ARMA) model ^[9]. In this study, we compared the spatial transmission network with the classical ARMA model used as the benchmark model to evaluate its merits in infectious diseases forecasting. Meanwhile, sentinel surveillance data of influenza-like illness (ILI) as an example in Chengdu, Deyang and Mianyang of Sichuan Province were collected and analyzed in this study. Sichuan Province is one of the core regions for influenza surveillance in China. In terms of the

three specific included cities, Chengdu, Deyang and Mianyang are located next to each other sequentially, and accounting for 40% of the total population and 53% of the regional GDP of Sichuan Province, respectively. The representativeness of these areas not only guaranteed the applicability of this method in these regions, but also offered evidence for future application to other regions. Therefore, this study selected these three places as examples to illustrate the *forecasting performance* of spatial transmission network for infectious diseases across multiple regions.

2 Methods

2.1 Data collection and preparation

Ever since the outbreak of the influenza A (H1N1) in 2009, China has established a nationwide ILI sentinel surveillance system, which consists of 193 sentinel hospitals in 30 provinces at present. The identification of patients with ILI is based on the standard case definition of the WHO, that is, body temperature ≥ 38 degrees Celsius with either cough or sore throat, in the absence of an alternative diagnosis^[10]. Besides, the ILI% was widely used in influenza surveillance to reflect the local intensity of influenza epidemic, where $ILI\% = (\text{the number of ILI cases in out-patient and emergency departments}) / (\text{the total number of out-patient and emergency cases}) * 100\%$ ^[11-12]. Moreover, for the convenience of analysis, the logarithmic transformation of ILI% performed prior to the modelling process in this study, and $x_1(t), x_2(t), x_3(t)$ denoted the logarithmic transformed ILI% values of Chengdu, Deyang and Mianyang at the t -th week ($t = 1, 2, \dots, 416$), respectively. Accordingly, all the observed data at the t -th

week could be noted as $\mathbf{x}_t = \{x_1(t), x_2(t), x_3(t)\}$, where $\mathbf{x}_t$ was a vector with three series components (boldface notation indicated vectors and matrices in the paper).

Furthermore, the application of spatial transmission network required each component of time series to meet weak stationary at least, which meant the mean and covariance matrices of a time series were independent on time. The stationarity assumption could be tested by the unit-root test. If it was violated, one should seek other time series analysis techniques for help. For the sake of clarity, this study only focused on situations when both conditions were satisfied, and readers who might be interested in more complicated situations could refer to the work of other researchers ^[13].

2.2 The forecasting of spatial transmission network

2.2.1 The construction of spatial transmission network

The spatial transmission network contained two kinds of information. The first was structural information, which was related to the existence and direction of disease transmission between places in the network. The second was parametric information, which measured the intensity of disease transmission between different regions. Correspondingly, the construction of spatial transmission network consisted of structure learning and parameter learning, which extracted the above two kinds of information from the original data respectively ^[6]. We used the dynamic Bayesian network model for structure learning and the VARMA model for parameter learning. After the construction of spatial transmission network, the possible outbreak and prevalence of infectious diseases could be predicted by using the spatial transmission network. Since

the emphasis of this study was the *forecasting performance* of the spatial transmission network, and the prediction was based on parameter learning, we would focus on the VARMA model next.

The VARMA model is a flexible modelling framework that can comprehensively describe and predict the dynamic relations among each component of multivariate time series. However, unlike the ARMA-type models, the VARMA model encounters the problem of *identifiability*, which means the coefficients may not be uniquely determined. Therefore, the VARMA model must involve structural specification technique (such as the Kronecker indices method) before any of its further application. A general VARMA (p, q) model can be written as

$$\mathbf{x}_t = \mathbf{a}_0 + \sum_{i=1}^p \mathbf{a}_i \mathbf{x}_{t-i} + \boldsymbol{\varepsilon}_t - \sum_{j=1}^q \mathbf{b}_j \boldsymbol{\varepsilon}_{t-j} \quad (1)$$

where p and q are nonnegative integers, $\mathbf{a}_0$ is a three-dimensional constant vector, $\mathbf{a}_i$ and $\mathbf{b}_j$ are 3×3 constant matrices, and $\{\boldsymbol{\varepsilon}_t\}$ is a sequence of independent and identically distributed random vectors. Once the identification problem has been settled, equation (1) could be transformed into equation (2) to address its meaning in public health. In equation (2), the autoregressive coefficient $a_{i,pq}$ ($p, q=1, 2, 3$, where 1 stands for Chengdu, 2 for Deyang, and 3 for Mianyang) indicates the effect of the observation of the q -th city at time $(t-i)$ (i.e., $x_q(t-i)$) on the observation of the p -th city at time t (i.e., $x_p(t)$). Correspondingly, according to the definition of $\mathbf{x}_t$ and the law of logarithmic transformation, $\exp(a_{i,pq})$ could be interpreted as the risk ratio of the ILI% of the q -th city at time $(t-i)$ to that of the p -th city at time t .

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} = \begin{pmatrix} a_{0,1} \\ a_{0,2} \\ a_{0,3} \end{pmatrix} + \sum_{i=1}^p \begin{bmatrix} a_{i,11} & a_{i,12} & a_{i,13} \\ a_{i,21} & a_{i,22} & a_{i,23} \\ a_{i,31} & a_{i,32} & a_{i,33} \end{bmatrix} \begin{pmatrix} x_1(t-i) \\ x_2(t-i) \\ x_3(t-i) \end{pmatrix} + \begin{pmatrix} \varepsilon_1(t) \\ \varepsilon_2(t) \\ \varepsilon_3(t) \end{pmatrix} - \sum_{j=1}^q \begin{bmatrix} b_{j,11} & b_{j,12} & b_{j,13} \\ b_{j,21} & b_{j,22} & b_{j,23} \\ b_{j,31} & b_{j,32} & b_{j,33} \end{bmatrix} \begin{pmatrix} \varepsilon_1(t-j) \\ \varepsilon_2(t-j) \\ \varepsilon_3(t-j) \end{pmatrix} \quad (2)$$

In this study, the Kronecker index approach was used to perform structural specification of VARMA model^[7]. For a multivariate time series $\mathbf{x}_t$, the Kronecker index approach seeks to specify an index for each component of $\mathbf{x}_t$. These Kronecker indices jointly identify a VARMA model for $\mathbf{x}_t$. In other words, the Kronecker index approach specifies the maximum order of the AR and MA polynomials for each component of $\mathbf{x}_t$. It has been mathematically proved that the specified VARMA model by Kronecker index approach could overcome the problem of identifiability and reveal the hidden structure among multivariate time series.

The model fitting step was comprised of order determination and parameter estimation. First, in the part of order selection, the values of p and q in equation (1) and (2) were determined. Tiao and Tsay proposed to use the two-way p -value table for extended cross-correlation matrices to specify the order (p, q) ^[14]. Once the orders were determined, the parameters of VARMA model could be estimated by the conditional likelihood method.

2.2.2 Forecasting of the spatial transmission network

For VARMA (p, q) model, the l -step ahead forecast of $\mathbf{x}_{t+l}$ at the forecast origin t is

$$x_t(l) = E(x_{t+l} | F_t) = \psi_l \varepsilon_t + \psi_{l+1} \varepsilon_{t-1} + \dots \quad (3)$$

where F_t denotes the information available at t and ψ_l is the coefficient matrices.

We used the minimum mean-squared error criterion for the forecasts of the VARMA (p , q) time series $\mathbf{x}_t$. The l -step ahead forecast error is equation (4). Consequently, the covariance matrix of l -step ahead forecast error is equation (5).

$$e_t(l) = \varepsilon_{t+l} + \psi_1 \varepsilon_{t+l-1} + \dots + \psi_{l-1} \varepsilon_{t+1} \quad (4)$$

$$\text{Cov}[e_t(l)] = \sum \varepsilon + \psi_1 \sum \varepsilon \psi_1' + \dots + \psi_{l-1} \sum \varepsilon \psi_{l-1}' \quad (5)$$

After the spatial transmission network was built, future values of logarithmized ILI% could be forecasted by inputting the historical ones into the estimated model. In order to assess both the fitting and forecasting performance of the model, the whole time series data were split into the training set (which integrated the data from the 1st week of 2010 to the 26th week of 2017) and the testing set (data integrated from the second-half year of 2017). We used the first set for the test of model fitting and the second one for forecasting. In this way we could also assess both the fitting and forecasting performance of spatial transmission network in relative to the traditional ARMA model.

2.2.3 Performance validation

The performance validation step included the residual checking and performance comparison.

For the residual checking part, since a well-behaved VARMA model should ideally extract all the regular pattern information out of the original data, the residuals of the model were white noise (i.e., the data should be completely random series without any regular pattern information). Therefore, the Ljung-Box statistics was utilized in this

study to verify the model by testing whether residuals of the model were only white noise.

As for the performance comparison part, we compared the results of the spatial transmission network with those of the ARMA model. To this end, the accuracy and precision of methods were measured. The accuracy evaluated whether the fitted (forecasted) values of the logarithmized ILI% were close to the real ones, and the precision reflected the variation of the forecasted values. In this study, we adopted mean absolute error (MAE) and mean absolute percentage error (MAPE) to monitor accuracy, and the boxplot of forecasting standard errors to evaluate the precision. It is known that lower MAE and MAPE implies better accuracy, and lower forecasting standard error suggests better precision.

2.2.4 Performance interpretation

One of the major concerns of infectious diseases surveillance lies in how the epidemic of one place could influence its neighbouring areas, which could in turn be characterized by the spatial transmission network as the effect of changes in one variable on another variable. From this point of view, the spatial transmission network could be interpreted by equation (2), which indicated that the historical value in the j -th place would probably be useful in predicting current value in the i -th place $x_i(t)$ ($i \neq j$).

2.2.5 The implementation of forecasting in R

All statistical analyses were performed in R 3.5.0 (the R Foundation for Statistical Computing ^[15]), a free software environment for statistical computing and graphics. The computing packages, {MTS} for VARMA model and {STATS} for ARMA model, were downloaded and installed from the Comprehensive R Archive Network (CRAN) at <http://cran.r-project.org/>. Additionally, the VARMA and ARMA models could be estimated by the command *VARMA* and *arima*, respectively.

3 Results

3.1 Data description

The log-transformed ILI% series of actual values of Chengdu, Deyang and Mianyang could be seen from Fig.2, respectively. Stationarity could be visually verified since there was no dramatic change in each time series. Furthermore, the unit-root test showed that all the *P*-values were less than 0.05 (the corresponding Dickey-Fuller statistics were -4.8707, -5.1568 and -3.6973), which further confirmed that all the time series were stationary.

Besides, from the cross-correlation plots shown in Fig.1 (A)-(C), we could observe that there was a negative dynamic correlation between Chengdu and Mianyang, which was similar to the correlation between Deyang and Mianyang. In contrast, a positive dynamic correlation existed between Chengdu and Deyang. In summary, considering the stationarity and the existence of dynamic correlations among the logarithmic transformed ILI% time series had been verified, it was appropriate to use the VARMA model for further analysis.

Fig.1 The cross-correlation of log-transformed ILI% between (A)Chengdu and Deyang; (B)Chengdu and Mianyang; (C)Deyang and Mianyang.

3.2 Estimates of the VARMA model

In this study, we estimated the VARMA model for 1-390 weeks. The Kronecker index was estimated to be (1,1,1). Furthermore, it turned out that the VARMA (1,1) model with Kronecker index (1,1,1) served as the best model. As a result, equation (6) showed the estimated VARMA model for the training dataset (1-390 weeks).

$$\begin{cases} x_1(t) = -0.72 + 0.90x_1(t-1) - 0.06x_3(t-1) + \varepsilon_1(t) - 0.26\varepsilon_1(t-1) + 0.11\varepsilon_2(t-1) + 0.11\varepsilon_3(t-1) \\ x_2(t) = -1.58 + 0.09x_1(t-1) + 0.69x_2(t-1) - 0.06x_3(t-1) + \varepsilon_2(t) - 0.04\varepsilon_1(t-1) - 0.15\varepsilon_2(t-1) + 0.02\varepsilon_3(t-1) \\ x_3(t) = -0.68 - 0.04x_1(t-1) - 0.05x_2(t-1) + 0.94x_3(t-1) + \varepsilon_3(t) + 0.12\varepsilon_1(t-1) + 0.06\varepsilon_2(t-1) - 0.31\varepsilon_3(t-1) \end{cases} \quad (6)$$

$$\text{Residuals cov-matrix} = \begin{bmatrix} 0.09 & 0.01 & 0.00 \\ 0.01 & 0.12 & 0.01 \\ 0.00 & 0.01 & 0.05 \end{bmatrix}$$

According to the estimated results of equation (6), the logarithmic transformed ILI% of Chengdu had a positive lag-1 effect on Deyang (estimated coefficient of the log-transformed ILI%=0.09, RR=1.09). Furthermore, the fitted model also indicated that the current log-transformed ILI% was influenced by the historical data of the past week,

which was again in accordance with the report interval of influenza in previous literature (about one week) ^[16]. More explanations about the estimated model were further discussed in the section 3.4.

3.3 Performance validation

It could be generally seen from Fig.2 that both the fitted values and forecasted values approximated very well to the actual ones for Chengdu, Deyang and Mianyang. In addition, the forecasting part also exhibited that the 95% CI of forecasted values could basically contain the actual values, which suggested the spatial transmission network was also good at precision. In addition, more details about the residual checking and model comparison were given as below.

Fig.2 The fitted/forecasted time series vs. the actual values for (A)Chengdu, (B)Deyang and(C) Mianyang.

3.3.1 Residual checking

The Q statistics for 1-, 5- and 10-lag Ljung-Box tests were 2.19 ($df = 9$, $P = 0.99$), 35.70 ($df = 45$, $P = 0.84$) and 101.30 ($df = 90$, $P = 0.20$), respectively, which validated that the VARMA model had fully extracted the useful information out of the original data and that the left residuals were only white noise series. Meanwhile, compared with

Fig.1 (A)-(C), the Fig.3 displayed that the information of cross-correlation between dynamic time series had been fully extracted by the VARMA model.

Fig.3 The cross-correlation of the residuals between (A)Chengdu and Deyang; (B)Chengdu and Mianyang; (C)Deyang and Mianyang.

3.3.2 Performance comparisons

Comparisons of the spatial transmission network and ARMA model involved the accuracy comparison as well as the precision comparison. The accuracy comparison was conducted in both the fitting and forecasting datasets. Meanwhile, the precision comparison was merely implemented in the forecasting datasets since the forecasting of future events was the major concern for infectious diseases surveillance.

3.3.2.1 Accuracy comparison

Table 1 listed the comparison of the fitting MAE and MAPE for the two methods. Compared with the ARMA model, it could be seen that the spatial transmission network had different fitting performance for each city. For Deyang, the spatial transmission

network performed slightly worse than the ARMA model, albeit the relative changing rate for both MAE (0.0013%) and MAPE (0.1271%) were extremely small. Conversely, the spatial transmission network outperformed the ARMA model for the other two cities. For Chengdu, it showed that the spatial transmission network could reduce MAE and MAPE by 0.8460% and 0.6537% respectively compared to the ARMA model. Regarding to Mianyang, the spatial transmission network had the best fitting performance with the relative changing rates of MAE and MAPE both exceeded 1% in relative to those of the ARMA model. In summary, though the fitting mean squared errors (MSEs) of both methods were already low enough, it was worth mentioning that in such condition, the spatial transmission network could be even more accurate in terms of fitting performance.

Table 1 The comparison of the fitting MAE and MAPE of the two methods

City	MAE		*RCR of MAE (%)	MAPE		RCR of MAPE (%)
	ARMA	STN		ARMA	STN	
Chengdu	0.2219575	0.2200797	-0.8460	4.916080	4.883944	-0.6537
Deyang	0.2539881	0.2539913	0.0013	4.534722	4.540486	0.1271
Mianyang	0.1665659	0.1642758	-1.3749	4.022101	3.965030	-1.4189

Abbreviations: STN, spatial transmission network.

*RCR of MAE: relative changing rate= (the fitting MAE of VARMA model- the fitting MAE of ARMA model)/the fitting MAE of ARMA model*100%. The definition of RCR of MAPE was in similar way.

Table 2 compared the forecasting MAE and MAPE of the ARMA and spatial transmission network. It showed clearly that the spatial transmission network outperformed the ARMA model by at least 1% for both the MAE and MAPE. Combining Table 1 and Table 2 it could be seen that the spatial transmission network consistently outperformed ARMA for Chengdu and Mianyang, both in term of fitting

and forecasting. As for Deyang, although the spatial transmission network performed slightly worse than ARMA model in terms of fitting (with less than 1%), it should be noted that the forecasting MAE and MAPE of spatial transmission network for Deyang were reduced by as high as 5.6277% and 5.4052% with that of ARMA model.

While it was plausible to imply that the spatial transmission network was generally better than the ARMA model in accuracy, the results also indicated the robustness of spatial transmission network from fitting to forecasting. For example, it should be noted from Table 1 and Table 2 that the forecasting MAE (MAPE) were much higher than the fitting MAE (MAPE) in Mianyang. From Fig.2 (C), we could see that there was a dramatic upward trend of ILI% in Mianyang starting from the winter of 2017, which caused tendency change in distribution characteristics between the training set and the testing set. This somehow led to the large difference between forecasting MSE and fitting MSE in Mianyang. However, even under such an unfavourable situation for forecasting, the spatial transmission network could still reduce the forecasting MAE and MAPE of the ARMA model by 1.9881% and 1.3368%, which indicated that the spatial transmission network was more robust than the ARMA model in terms of accuracy.

Table 2 The comparison of the forecasting MAE and MAPE of the two methods

City	MAE		*RCR of MAE (%)	MAPE		RCR of MAPE (%)
	ARMA	STN		ARMA	STN	
Chengdu	0.1801900	0.1730801	-3.9458	4.948527	4.636171	-6.3121
Deyang	0.3446477	0.3252521	-5.6277	7.405478	7.005199	-5.4052
Mianyang	0.2693281	0.2639735	-1.9881	5.482122	5.408838	-1.3368

Abbreviations: STN, spatial transmission network.

*RCR of MAE: relative changing rate= (the fitting MAE of VARMA model- the fitting MAE of ARMA model)/the fitting MAE of ARMA model*100%. The definition of RCR of MAPE was in similar way.

3.3.2.2 Precision comparison

Since the precision of forecasting model could be represented by the forecasting errors at each time point, we plotted the boxplots of forecasting errors of spatial transmission network and ARMA model for each city (Fig.4). For both the spatial transmission network and the ARMA model, it could be seen that the standard errors were the highest in Deyang, and the lowest in Mianyang, which were in coincidence with the variation levels of the actual observations of each city in the second half year of 2017 (see in Fig.2). This was reasonable because there was no doubt that the variation level of actual observations would definitely affect the forecasting standard errors. Furthermore, it should be noted that no matter how the actual observations varied, the forecasting standard errors of the spatial transmission network were always smaller than those of the ARMA model. Therefore, it could be concluded that the spatial transmission network outperformed the ARMA model in terms of precision.

Fig.4 The Boxplots of forecasting errors of spatial transmission network and ARMA model for each city.

3.4 Interpretations of the results

Results of this study not only verified the accuracy and precision of the spatial transmission network, but also provoked inspirations on the spreading directions of influenza epidemics among the three cities in Sichuan province. From the estimated VARMA model in equation (6), it was plausible to speculate that the influenza virus may spread from Chengdu to Deyang, while there was no evidence to support that it could also spread from Mianyang to the other two cities. Besides the estimated results of this study, the speculation could also be supported by the following three reasons.

① Deyang is located right in the middle of Chengdu and Mianyang. In accordance with the First Law of Geography, which means everything is related to everything else, but near things are more related to each other ^[17], spatial effects of influenza might somehow explain correlations of epidemics among Chengdu, Deyang and Mianyang.

② It is well known that influenza is more likely to outbreak in areas with a dense population and heavy traffic. In fact, Chengdu and Mianyang are the two largest cities in Sichuan province, while Deyang is incomparable with them in terms of population density and traffic volume. Therefore, it was of practical significance to speculate that influenza might spread from either Chengdu or Mianyang to Deyang.

③ Though the regression coefficient between Mianyang and Deyang in equation (6) was negative (-0.06), it still could not be interpreted that the increase of ILI% in

Mianyang would reduce that in Deyang. Considering the regression coefficients involving Mianyang and the other two cities were all negative but small, a more plausible interpretation was that the influenza epidemic of Mianyang was different from those of Chengdu and Deyang, and that there might be no influenza spreading between Mianyang and the other two cities.

Meanwhile, it is worth mentioning that although the above speculation needs further laboratory and etiological verifications that are far beyond our current study, one thing to emphasize is that the speculation itself at least could shed light upon future studies, because once the spreading direction of influenza was determined, it would be much more beneficial for making decisions on early warning and prevention of influenza outbreaks.

4 Discussion

The results of this study illustrated that the spatial transmission network had advantages in *forecasting performance* of infectious diseases across multiple regions. The accuracy and precision of its *forecasting performance* were superior to the ARMA model as a univariate time series. Combined with our previous research results ^[6], that was, the spatial transmission network also showed good *interpretability* and *feasibility*, it could be seen that the spatial transmission network would be very helpful in guiding the practical prevention and control work of infectious diseases. According to the description of Stoto, a practical surveillance system should include three parts: continuous monitoring of multivariate data, applying algorithms to raise the alarm when

395 something unusual is happening, and a protocol on how to respond to an alarm ^[18]. It
396 followed from our results that the spatial transmission network could at least assist to
397 improve all the three parts of surveillance system. Firstly, as a multivariate time series
398 analysis model, the VARMA model could inherently integrate and extract information
399 from multivariate data, which not only limited to a single variable (e.g., ILI%), but
400 could also contain other-type variables (e.g., the data involves lots of factors like ILI%
401 as well as meteorological, social-economic and environmental factors). The second
402 advantage of spatial transmission network was that its results could suggest when and
403 where the next outbreak would probably occur, as well as how serious the next outbreak
404 would be, so it could provide suggestion and evidence on whether an alarm should be
405 raised. Last but not least, the information of direction and time-lag of influenza
406 transmission exhibited by spatial transmission network could also indicate the way
407 about how to respond to an alarm. For example, as the results of this study found that
408 the influenza virus might spread from Chengdu to Deyang, then it was plausible to
409 respond to the alarm by organizing epidemiological investigation and laboratory
410 analyses especially in Chengdu so that the potential outbreak of influenza might be
411 eliminated at the start point. Under such circumstances, the government could take
412 strongly targeted actions such as isolation, quarantine and social distancing to
413 efficiently achieve the goal of early warning and prevention.

414 Except for the results of this study, there are still other features which could also
415 guarantee the potential use of spatial transmission network in the practice of influenza
416 surveillance.

① From correlation to causation. The VARMA model could provide evidence of both strength of association and temporality, which are the two key points of the Hill's criteria for causality ^[19]. Meanwhile, under some mild conditions, the probability distribution of VARMA model could be reliably represented as a casual network, and the latter is a commonly used tool in causal inference ^[20]. To this end, it is plausible that the spatial transmission network could at least partly serves for the etiological study to screen crucial clues of the cause-and-effect relationships.

② The representativeness of this study. It is well known that China is a large country with great diversities among different areas in many aspects (i.e., environment, population, economy and social customs). From the perspective of the representativeness of Sichuan Province, it has the fourth largest population, the fifth largest land and the sixth highest GDP in China. Besides, the population in Sichuan province included 55 of China's 56 ethnic groups, and there existed many major types of landforms in Sichuan (i.e., mountains, hills, plateaus and plains). Therefore, the representativeness of Sichuan province makes the spatial transmission network be likely to be applied into the surveillance of influenza in other areas.

It should also be acknowledged that there are still some limits in our study. For example, we merely proposed our explanations based on the features of our analyzed results and some practical experience, so we still need relevant etiological and laboratory data to further identify our speculation about the spreading direction of influenza in Sichuan province. Besides, there is still a long way to go from model building to making causal inferences about influenza epidemics in the real world.

Therefore, it is highly expected that the spatial transmission network could provide a new way for causal inference in the surveillance of infectious diseases.

5 Conclusions

This study applied the spatial transmission network to forecasting the infectious diseases across multiple regions. The results illustrated that the spatial transmission network not only had good accuracy and precision in *forecasting performance*, but also could indicate the spreading directions of infectious diseases among multiple regions to a certain extent. Therefore, the spatial transmission network is a promising candidate to improve the surveillance work.

List of abbreviations

CDC: Centers for disease control and prevention

WHO: World Health Organization

ARMA: Autoregressive moving average

VARMA: Vector autoregressive moving average

ILI: Influenza-like illness

MAE: Mean absolute error

MAPE: Mean absolute percentage error

MSE: Mean squared error

STN: Spatial transmission network

458

459 **Declarations**

460 **Ethics approval and consent to participate**

461 The ILI surveillance was a routine surveillance activity. The analysis of ILI data was
462 not considered human subject research. No administrative permission was needed to
463 assess the data.

464 **Consent for publication**

465 Not applicable.

466 **Availability of data and materials**

467 The data that support the findings of this study are available from Sichuan Center for
468 Disease Control and Prevention but restrictions apply to the availability of these data,
469 which were used under license for the current study, and so are not publicly available.
470 Data are however available from the authors upon reasonable request and with
471 permission of Sichuan Center for Disease Control and Prevention.

472 **Competing interests**

473 The authors declare that they have no competing interests.

474 **Funding**

475 This research work was funded by Sichuan Science and Technology Program
476 (2020YFS0015, 2020YFS0091), Health Commission of Sichuan province (20PJ092),

National Natural Science Foundation of China (81602935), Chongqing Science and Technology Program (cstc2020jsexcylyhX0003), Sichuan University (2018hhf-26, GSSCU2018038) and Liang Shan Zhou Center for Disease Control and Prevention. The funders played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

HMW consulted the literature, analyzed the data, and was a major contributor in writing the manuscript. JQQ analyzed the data and wrote part of the manuscript. CL wrote part of the manuscript and checked the full manuscript. HLW analyzed the results. CHY collected the data. TZ analyzed the data and reviewed all the material. All authors contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

- [1] Centers for Disease Control and Prevention: Seasonal flu death estimate increases worldwide. <https://www.cdc.gov/media/releases/2017/p1213-flu-death-estimate.html>. Accessed 28 Aug 2020.

495 [2] Iuliano AD, Roguski KM, Chang HH, Muscatello DJ, Palekar R, Tempia S, et al. Estimates of
 496 global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet*.
 497 2018;391(10127):1285-1300.

498 [3] 2019-2020 U.S. Flu Season: Preliminary Burden Estimates.
 499 <https://www.cdc.gov/flu/about/burden/preliminary-in-season-estimates.htm> Accessed 28 Aug
 500 2020.

501 [4] World Health Organization: The strategic movement of WHO to deal with influenza
 502 pandemics. <http://www.who.int/csr/resources/publications/influenza/CHfinal.pdf?ua=1>. Accessed
 503 28 Aug 2020.

504 [5] Thacker SB, Choi K, Brachman PS. The surveillance of infectious diseases. *JAMA*.
 505 1983;249(9):1181-1185.

506 [6] Qiu J, Wang H, Hu L, Yang C, Zhang T. Spatial transmission network construction of
 507 influenza-like illness using dynamic Bayesian network and vector-autoregressive moving average
 508 model. *BMC Infect Dis*. 2021;21(1):164.

509 [7] Ruey S. Tsay. Multivariate time series analysis with R and financial applications. New Jersey:
 510 Wiley; 2016.

511 [8] García Nieto PJ, Sánchez Lasheras F, García-Gonzalo E, de Cos Juez FJ. PM10 concentration
 512 forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM,
 513 MLP, VARMA and ARIMA: A case study. *Sci Total Environ*. 2018;621:753-761.

514 [9] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M. Ljung. Time Series
 515 Analysis: Forecasting and Control. 5th ed. New Jersey: John Wiley and Sons Inc.; 2015.

516 [10] World Health Organization: WHO surveillance case definitions for ILI and SARI.
 517 https://www.who.int/influenza/surveillance_monitoring/ili_sari_surveillance_case_definition/en/.
 518 Accessed 28 Aug 2020.

519 [11] Harcourt SE, Morbey RA, Smith GE, Loveridge P, Green HK, Pebody R, et al. Developing
 520 influenza and respiratory syncytial virus activity thresholds for syndromic surveillance in England.
 521 Epidemiol Infect. 2019;147:e163.

522 [12] Zhu D, Lok C, Chao S, Chen L, Li R, Zhao Z, et al. Detection and characterization of type B
 523 influenza virus from influenza-like illness cases during the 2017-2018 winter influenza season in
 524 Beijing, China. Arch Virol. 2019;164(4):995-1003.

525 [13] Nakajima J. Bayesian analysis of multivariate stochastic volatility with skew return
 526 distribution. Economet Rev. 2017;36(5):546-62.

527 [14] Tiao GC, Tsay RS. Multiple time series modelling and extended sample cross correlations. J
 528 Bus Econ Stat. 1983; 1: 43–56.

529 [15] The R Project for Statistical Computing: R language and environment for statistical
 530 computing. <https://www.R-project.org/>. Accessed 28 Aug 2020.

531 [16] World Health Organization: Influenza (Seasonal). [https://www.who.int/en/news-room/fact-](https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal))
 532 [sheets/detail/influenza-\(seasonal\)](https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)). Accessed 28 Aug 2020.

- 533 [17] Tobler W R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic*
534 *Geography*. 1970; 46 Supp 1:234-240.
- 535 [18] Stoto A. Comment on: A Bayesian dynamic model for influenza surveillance by Sebastianiet
536 al. *Statistics in Medicine* (this issue). *Stat Med*. 2006;25(11):1817-1818.
- 537 [19] Science-based Medicine: Causation and Hill's criteria.
538 <https://sciencebasedmedicine.org/causation-and-hills-criteria/>. Accessed 28 Aug 2020.
- 539 [20] Lèbre S. Inferring dynamic genetic networks with low order independencies. *Stat Appl Genet*
540 *Mol Biol*. 2009;8:Article 9.

Figures

Figure 1

The cross-correlation of log-transformed ILI% between (A)Chengdu and Deyang; (B)Chengdu and Mianyang; (C)Deyang and Mianyang.

Figure 2

The fitted/forecasted time series vs. the actual values for (A)Chengdu, (B)Deyang and(C) Mianyang.

Figure 3

The cross-correlation of the residuals between (A)Chengdu and Deyang; (B)Chengdu and Mianyang; (C)Deyang and Mianyang.

Figure 4

The Boxplots of forecasting errors of spatial transmission network and ARMA model for each city.