AIMS Mathematics

AIMS Mathematics, 8(11): 25528–25549.
DOI: 10.3934/math.20231303
Received: 05 August 2023
Revised: 24 August 2023
Accepted: 27 August 2023
Published: 04 September 2023

http://www.aimspress.com/journal/Math

Research article

Influenza trend prediction method combining Baidu index and support vector regression based on an improved particle swarm optimization algorithm

Hongxin Xue[1,2,3,*], Lingling Zhang[1,2,3], Haijian Liang[4], Liqun Kuang[1,2,3], Huiyan Han[1,2,3], Xiaowen Yang[1,2,3] and Lei Guo[1,2,3]

[1] School of Computer Science and Technology, North University of China, Taiyuan, Shanxi 030051, China
[2] Shanxi Province's Vision Information Processing and Intelligent Robot Engineering Research Center, Taiyuan, Shanxi 030051, China
[3] Shanxi Key Laboratory of Machine Vision and Virtual Reality, Taiyuan, Shanxi 030051, China
[4] School of Software, North University of China, Taiyuan, Shanxi 030051, China

* **Correspondence:** Email: 18334792279@163.com.

Abstract: Web-based search query data have been recognized as valuable data sources for discovering new influenza epidemics. However, selecting search and query keywords and adopting prediction methods pose key challenges to improving the effectiveness of influenza prediction. In this study, web search data were analyzed and excavated using big data and machine learning methods. The flu prediction model for the southern region of China, considering the impact of influenza transmission across regions and based on various keywords and historical influenza-like illness percentage (ILI%) data, was built (models 1–4) to verify the factors affecting the spread of the flu. To improve the accuracy of the influenza trend prediction, a support vector regression method based on an improved particle swarm optimization algorithm was proposed (IPSO-SVR), which was applied to the influenza prediction model to forecast ILI% in southern China. By comparing and analyzing the prediction results of each model, model 4, using the IPSO-SVR algorithm, exhibited higher prediction precision and more effective results, with its prediction indexes including the mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) being 0.0596, 0.2441 and 0.1884, respectively. The experimental results show that the prediction precision significantly increased when the IPSO-SVR method was applied to the constructed ILI% model. A new theoretical basis and implementation strategy were provided for achieving more accurate influenza prevention and control in southern China.

Keywords: influenza; web search data; prediction model; principal component analysis; support vector regression; improved PSO algorithm
Mathematics Subject Classification: 97R20, 97R50

1. Introduction

Influenza (flu) is a very common respiratory infectious disease with high variability and infectivity. It spreads rapidly through droplets, with an extremely fast transmission speed and incubation period, which makes the influenza epidemic rapidly spread in a short time and pose a serious threat to human health [1]. According to statistics from the World Health Organization (WHO), there are an estimated 600 million to 1.2 billion cases of influenza worldwide each year. This number includes at least 3 million severe cases and complications associated with influenza, and the number of fatalities ranges from 250,000 to 500,000 [2,3]. According to statistics, approximately 84,200 to 92,000 people die from respiratory diseases caused by the flu in China every year, accounting for 8.2% of total deaths from respiratory diseases [4]. At the same time, the annual economic burden caused by influenza is about ¥ 26.381 billion in China, equivalent to 0.233‰ of GDP in 2021 [5], which is shocking. If active and effective prevention and control measures are not taken, influenza will continue to impose a serious health and economic burden on both China and the world. Therefore, preventing and controlling flu is important public health work that requires extensive attention and investment.

Several methods have been proposed for real-time detection and routine monitoring of flu activity. Traditional influenza surveillance systems primarily rely on reported influenza-like cases and virological data from health care providers, including hospitals, clinics and contract laboratories [6,7]. Although China has established a nationwide influenza surveillance system, the time taken to publicly report influenza cases is usually delayed by about 1–2 weeks. Furthermore, there are issues such as high operating costs, low coverage of the surveillance network, low efficiency in information reporting, over-reliance on historical influenza data without multidimensional data support and simplistic methods for data mining, prediction and early warning [8]. If it is possible to predict the flu trends in certain areas promptly and accurately, and take appropriate prevention and control measures before the outbreak of influenza, we can effectively control the spread of the disease and reduce the harm and economic losses caused by it.

Yang et al. [9] propose a comprehensive learning particle swarm optimization based machine learning (CLPSO-ML) framework incorporating support vector regression (SVR) and multilayer perceptron (MLP) for multi-step-ahead influenza prediction. Wang et al. [10] propose a new end-to-end spatiotemporal deep neural network structure for influenza risk prediction. The proposed model mainly consists of two parts. The first stage is the spatiotemporal feature extraction stage where two-stream convolutional and recurrent neural networks are constructed to extract different regions and time granularity information. Then, a dynamically parametric-based fusion method is adopted to integrate the two stream features and make predictions. Kumar et al. [11] propose a hybrid fuzzy time series forecasting model based on particle swarm optimization and the fuzzy c-mean technique, named as fuzzy time series particle swarm optimization extended fuzzy c-mean technique. Thomas et al. [12] develop methods for real-time prediction of the risk that an ongoing influenza epidemic will be exceptionally severe and for real-time detection of anomalous epidemics and use them for prediction and detection of anomalies for influenza epidemics in France. The quality of predictions is assessed on observed and simulated data. Wei et al. [13] aimed to enhance their prediction model by incorporating traditional hydrological and atmospheric data. Features, such as popular search keywords on Google Trends, public holiday information, population density, air quality indices, and the numbers of COVID-19 confirmed cases, were also used to train the model. Kara [14] introduced a hybrid method that combines long short-term memory (LSTM) neural network and genetic algorithm (GA) for multi-step influenza outbreak forecasting problems. Kumar et al. [15] propose a hybrid fuzzy time series model for the prediction of upcoming COVID-19 infection cases and deaths in India by using a modified fuzzy C-means clustering technique.

At present, some non-traditional methods for influenza monitoring have been developed. For example, Ackley et al. [16] conducted a comparative analysis by integrating data from smart thermometers and mobile applications with regional influenza and influenza-like illness (ILI) surveillance data from the California Department of Public Health. They utilized smart thermometer readings and mobile application data to predict regional influenza in California. The experimental results demonstrated that these data improved the predictive capability of influenza illness. Murayama et al. [17] utilized inter-regional commuting data as a representation of human mobility when building a regional influenza prediction model and used it as spatial information in graph convolutional network (GCN) to predict the geographical distribution of influenza patients. The results show that the GCN model based on commuting data significantly improves the prediction accuracy in both temporal and spatial dimensions, thus providing an appropriate prediction interval. Yang et al. [18] developed a comprehensive influenza monitoring framework by integrating electronic medical records (EMRs) from several hospitals in Taiwan and ILI data from the Taiwan Center for Disease Control and Prevention (TWCDC). This framework is scalable and can periodically integrate TWCDC ILI open data with EMRs across multiple hospitals to automatically monitor influenza activity and support early surveillance of influenza outbreaks. In addition, some researchers have achieved real-time monitoring and prediction of influenza activity by utilizing non-traditional data sources such as social media data [19], web search data [20], call center data [21], pharmacy sales data [22] and meteorological data [23].

To enhance the prediction and response capabilities to influenza outbreaks, numerous researchers and institutions are devoted to improving influenza prediction models. These models encompass prediction models based on machine learning and deep learning, alongside prediction models grounded in mathematical models. For instance, Lu et al. [24] proposed the ARGONet method, which combines two prediction approaches with machine learning to estimate local influenza epidemics in real-time. This method first extended the proven inference method for influenza activity, called ARGO, to various states in the United States, and incorporates information related to influenza, such as Google search frequency, electronic health records and historical flu trends. To enhance prediction accuracy, a spatial network method called Net was developed based on ARGO, which improved the influenza estimation of ARGO by combining the spatiotemporal patterns of influenza transmission in neighboring regions. In this study, the ARGO model alone outperformed the Google Flu Trend prediction system that operated from 2008 to 2015. Researcher Fred Lu stated that this new method may lay the foundation for effective prevention of infectious diseases. With the increasing availability of online search data and cloud-based electronic health records collected from medical service providers, this new model will be able to predict disease outbreaks and epidemics more accurately in the future. Zimmer et al. [25] combined the developed calibration and prediction framework with the established humidity-based propagation dynamics model to predict influenza. They found that incorporating daily near real-time internet search data improved the accuracy of short-term and medium-term predictions of influenza activity. Miliou et al. [26] proposed the use of retail market data to improve the prediction of seasonal influenza and developed a near-term forecasting and prediction framework that provided estimates of influenza incidence in Italy. They employed a SVR model to predict seasonal influenza incidence. The results quantitatively show the value of incorporating retail market data into the prediction model, which can serve as an agent for real-time analysis of epidemics. Huang [27] utilized a retrospective epidemiological survey method and based on the Baidu index of H7N9 avian influenza keywords and clinical symptoms keywords of H7N9 subtype avian influenza, established the SVR prediction model and multiple linear regression prediction model in different segments to analyze the fit degree. The results revealed that public search behavior, epidemic segment characteristics and the frequency of public search for clinical symptoms of infectious diseases significantly improved the capability of search engine big data to predict the epidemic trends of H7N9 subtype avian influenza.

Recent research indicates that flu transmission trends can be effectively monitored by integrating open-source search query data and machine learning methods. This approach not only enables the timely provision of useful information to the public and medical professionals for taking appropriate prevention and control measures but also holds tremendous potential. However, research in this field is still relatively limited domestically, which necessitates further exploration and development. Currently, research methods mainly focus on multiple correlation regression analysis [28,29], but this approach has some issues in predicting the trend of influenza transmission. For example, in a multiple linear regression model, multicollinearity among the independent variables may lead to model instability. Additionally, the relationship between ILI and related factors is influenced by various factors, which may not exhibit a simple linear relationship. Therefore, using conventional linear models for fitting may not achieve the desired predictive performance.

Overseas research has primarily focused on using Google search engine data and Twitter data [30–34], while in China, Baidu index has become one of the main sources of search engine data. As of July 2022, Baidu holds a dominant market share of 71.2% in the Chinese search engine market, far surpassing other search engines, which better reflects the level of attention that most Chinese people have towards the epidemic. Therefore, this study utilized web search data provided by Baidu index and ILI data, and constructed a nonlinear influenza prediction model suitable for the characteristics of southern China based on machine learning methods. By leveraging Baidu index and machine learning algorithms, the model can better predict the spread trend of influenza and provide relevant information in time to provide scientific support for influenza prevention and control efforts.

2. Data acquisition and processing

2.1. Data source

The official influenza like case data used in this study was obtained through the ILI weekly report released by the National Influenza Center of China [35]. The collection of this data relies on the collaboration of medical institutions at all levels, disease prevention and control centers, and sentinel hospitals for monitoring, summarizing and analyzing influenza data reported by sentinel hospitals across the country. In this paper, we collected 207 weeks of ILI data in southern regions of the China from the 1st week of 2018 to the 49th week of 2021. These official influenza sample case data are recognized as reliable sources widely used for research and monitoring of influenza transmission trends. By analyzing these data, we can obtain important information about the epidemic situation and changing trends of influenza in southern China.

The web search data originates from Baidu index of Baidu search engine [36]. It is a statistical index that comprehensively reflects the reference value of user interest and media attention to a specific keyword on a certain day. Based on the search volume of internet users on Baidu, the weighted sum of search frequency of each keyword in Baidu web search is analyzed and calculated. In this paper, we first conducted a long-tail keyword search using "flu" as the initial value on "Chinaz.com." We selected keywords with a whole network index greater than 200 and chose relatively original search terms related to influenza symptoms, treatment, preventive measures and other aspects. Then, we referred to the literature to summarize other keywords used in relevant studies. A total of 37 keywords that may be related to changes in the influenza epidemic trend were sorted out, as shown in Table 1.

Table 1. Baidu search keyword and number.

Number	English Name	Chinese Name	Number	English Name	Chinese Name
K1	How to prevent flu	如何预防流感	K20	Cold medicine	感冒药
K2	Flu prevention measures	流感的预防措施	K21	Contac	康泰克
K3	Influenza vaccine	流感疫苗	K22	Gankang	感康
K4	Avian influenza vaccine	禽流感疫苗	K23	Amoxicillin	阿莫西林
K5	Is a flu shot necessary	流感疫苗有必要打吗	K24	Lianhua Qingwen capsule	连花清瘟胶囊
K6	Avian influenza prevention	禽流感预防	K25	Tylenol	泰诺
K7	How to prevent colds	怎样预防感冒	K26	Influenza	流行性感冒
K8	Preventing avian influenza	预防禽流感	K27	Flu	流感
K9	Influenza A symptoms	甲型流感症状	K28	Swine flu	猪流感
K10	Flu symptoms	流感的症状	K29	Influenza virus	流感病毒
K11	Cold	感冒	K30	H1N1 influenza	甲流
K12	Viral cold	病毒性感冒	K31	What is H1N1 influenza	甲流是什么
K13	Stomach flu	肠胃感冒	K32	H1N1 influenza virus	甲流病毒
K14	Cold symptom	感冒症状	K33	Influenza A	甲型流感
K15	Fever	发烧	K34	Influenza A virus	甲型流感病毒
K16	Hot	发热	K35	Avian influenza	禽流感
K17	High fever	高烧	K36	H1N1	H1N1
K18	Influenza treatment	流感治疗	K37	H7N9	H7N9
K19	What medicine to take for the flu	流感吃什么药			

2.2. Data preprocessing

The Baidu index of flu-related keywords is counted on a daily basis. In order to conduct consistent analysis with other time series data, it needs to be aggregated on a weekly basis. Each keyword's weekly summaries are calculated separately. However, missing data were found when collecting the Baidu index for keywords. To improve the accuracy of the prediction of ILI, it is necessary to repair the raw data. To address this issue, the K-nearest neighbors (KNN) algorithm was employed to fill in the missing data in the Baidu index of keywords K1, K2, K4, K5, K6, K8, K9, K18, K31 and K32. The algorithm utilizes existing adjacent data points to infer the missing value and interpolates by finding neighbor data that is most similar to the missing data. This approach enables the estimation of the missing data and reduces its impact on the accuracy of ILI prediction. The repaired data allows for a more comprehensive analysis of the trends and changes in flu-related keywords, providing more accurate predictions and insights.

2.3. Keyword filtering

Research has shown that an increase in the number of keywords does not necessarily improve the model's fitting performance. In order to accurately select the influencing factors related to the predicted outcome variable ILI%, a correlation analysis was conducted by comparing ILI% with the curated Baidu search index of keywords. In this way, keywords that contribute to the prediction model can be screened and included in the prediction model. In this study, the IBM SPSS Statistics 26.0 statistical tool was used for conducting the correlation analysis. To preliminarily screen keywords, a minimum correlation coefficient of 0.5 between the time series of Baidu search index for keywords and ILI% was required. By conducting a ranking analysis based on the correlation between the Baidu search index for each keyword and ILI%, it was found that out of the 37 keywords, 17 keywords had correlation coefficients less than 0.5 with ILI%, while 20 keywords had correlation coefficients greater than 0.5. The specific analysis results are shown in Table 2, which will help in the further selection of the most relevant keywords to establish a more accurate prediction model.

Table 2. Results of keyword inter-correlation analysis.

Number	Correlation	Number	Correlation
K1	**0.71**	K20	**0.53**
K2	0.46	K21	**0.59**
K3	0.01	K22	**0.51**
K4	0.18	K23	**0.50**
K5	0.30	K24	0.08
K6	0.46	K25	**0.77**
K7	0.38	K26	0.46
K8	0.34	K27	**0.72**
K9	**0.68**	K28	0.00
K10	0.44	K29	**0.67**
K11	**0.58**	K30	**0.67**
K12	**0.75**	K31	**0.66**
K13	0.35	K32	**0.62**
K14	0.40	K33	**0.72**
K15	**0.85**	K34	**0.59**
K16	0.47	K35	0.42
K17	**0.81**	K36	0.38
K18	**0.71**	K37	0.40
K19	**0.65**		

Influenza viruses are primarily transmitted through airborne droplets produced by sneezing or coughing, as well as through direct contact between people or contact with objects contaminated by influenza viruses. With the rapid and frequent operation of modern transportation, the frequent flow of people and the transportation of various new types of food, previously localized infectious diseases may become widespread and epidemic diseases. In areas with frequent population mobility, if an influenza outbreak occurs in one region, other closely related regions are also likely to be affected. Therefore, it is of great significance to analyze the number of ILI in northern and southern

China. Figure 1 shows the trends of ILI in the southern and northern regions of China from week 1 in 2018 to week 49 in 2021. From the graph, it can be observed that the trend of rising and declining influenza activity levels in both southern and northern China is relatively consistent, and there is a strong correlation. This observation indicates that it is important to consider the impact of influenza transmission in the northern region on the southern region when modeling prediction models, and to assess its effect on the prediction effectiveness.

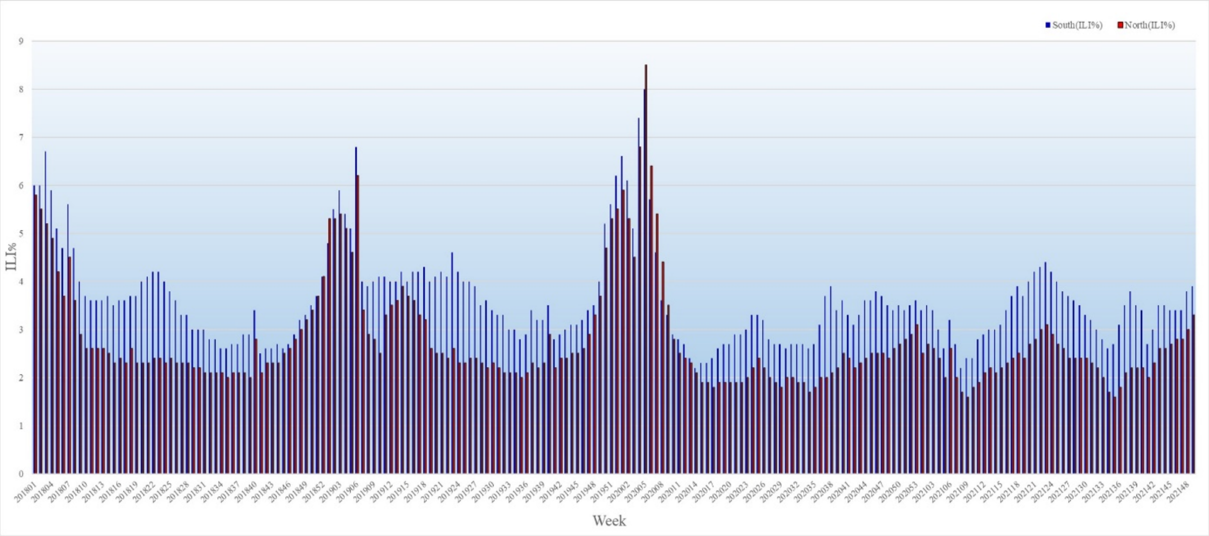

Figure 1. Trend of ILI in the southern and northern regions of China from week 1 in 2018 to week 49 in 2021.

2.4. Keyword time-delay correlation analysis

Due to the incubation period and subsequent disease development of influenza, the predictive factors generally exhibit time-delay characteristics. Therefore, the correlation trend between ILI% in the southern region of China and the Baidu search index of preliminary screened keywords was analyzed. The keywords "How to prevent influenza (K1)", "Influenza A symptoms (K9)", "Viral cold (K12)", "Fever (K15)", "High fever (K17)" and "Influenza A (K33)" were taken as examples. Figure 2 shows the distribution of ILI% and Baidu index of specific keywords in the southern region of China from week 1 in 2018 to week 49 in 2021. From the figure, it can be observed that the Baidu index of keywords K1, K9, K12 and K33 show a certain leading relationship compared to ILI%, while the Baidu index of keywords K15 and K17 exhibit relative synchronicity with ILI%. Based on the above analysis, it is evident that the influence of time lag should be considered when conducting keyword correlation analysis. Therefore, this study employed cross-correlation analysis to examine the time-lagged relationship between the selected keywords and ILI% in the southern region within a time range of 7 weeks before and after. The maximum absolute correlation coefficient for each keyword and ILI% was selected to ensure the correlation between the data. The results of the keyword cross-correlation analysis are shown in Table 3.

Keywords are classified into synchronous keywords, leading keywords and lagging keywords according to their temporal nature. From Table 3, it can be observed that as the number of lag days decreases, the correlation of each keyword gradually increases. Among them, 8 keywords reach the maximum value when the delay is 0, which belong to the "synchronous" keywords, including fever, high fever, what medicine to take for the flu, amoxicillin, flu, influenza virus, H1N1 influenza virus and influenza A virus. Additionally, there are 12 keywords that reach their maximum value when the delay is -1, which are "leading" keywords, including how to prevent flu, influenza A symptoms, cold, viral cold, flu treatment, cold medicine, Contac, Gankang, Tylenol, H1N1 influenza, what is H1N1 influenza and influenza A. Due to each keyword being highly correlated with ILI% at different lag times, the lag variable with the largest correlation coefficient was used to establish the model. This can more accurately reflect the association between keywords and ILI, thereby improving the accuracy of the prediction model.

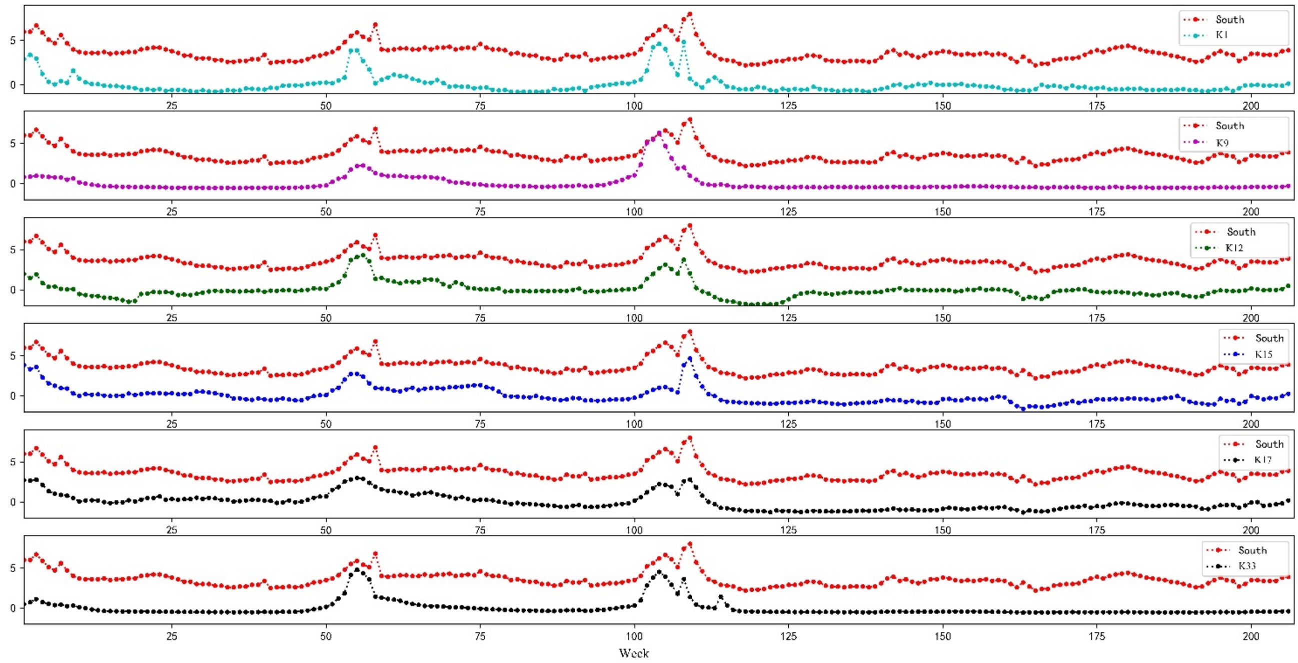

Figure 2. Distribution graph of ILI% and specific keyword Baidu index in the southern region of China.

Table 3. Results of time lag correlation analysis for the preliminary screened keywords.

Correlation Keyword	Delay weeks														
	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
K1	0.280	0.408	0.494	0.562	0.607	0.665	**0.742**	0.709	0.575	0.476	0.378	0.296	0.222	0.164	0.082
K9	0.401	0.518	0.592	0.651	0.683	0.709	**0.719**	0.684	0.596	0.498	0.391	0.294	0.209	0.136	0.067
K11	0.370	0.423	0.451	0.488	0.518	0.562	**0.591**	0.577	0.484	0.353	0.237	0.144	0.063	0.022	0.098
K12	0.281	0.366	0.444	0.545	0.624	0.705	**0.771**	0.751	0.651	0.540	0.415	0.285	0.164	0.051	0.062
K15	0.125	0.230	0.322	0.418	0.516	0.638	0.763	**0.848**	0.773	0.670	0.586	0.516	0.444	0.364	0.272
K17	0.268	0.371	0.463	0.542	0.619	0.697	0.768	**0.811**	0.745	0.645	0.546	0.459	0.371	0.283	0.198
K18	0.226	0.312	0.398	0.527	0.573	0.633	**0.683**	0.672	0.584	0.469	0.353	0.255	0.180	0.114	0.034
K19	0.213	0.285	0.350	0.426	0.499	0.568	0.629	**0.647**	0.582	0.457	0.315	0.223	0.160	0.112	0.063
K20	0.379	0.434	0.474	0.500	0.510	0.548	**0.561**	0.530	0.460	0.324	0.198	0.103	0.022	0.071	0.142
K21	0.445	0.498	0.529	0.558	0.580	0.612	**0.627**	0.591	0.526	0.411	0.298	0.209	0.130	0.041	0.034
K22	0.268	0.328	0.374	0.418	0.449	0.488	**0.520**	0.508	0.433	0.295	0.181	0.089	0.001	0.096	0.174
K23	0.201	0.247	0.282	0.322	0.352	0.400	0.470	**0.505**	0.470	0.378	0.298	0.230	0.162	0.084	0.001
K25	0.343	0.422	0.486	0.549	0.611	0.697	**0.768**	0.766	0.683	0.560	0.442	0.334	0.243	0.136	0.043
K27	0.027	0.102	0.178	0.264	0.350	0.463	0.643	**0.718**	0.640	0.602	0.547	0.491	0.425	0.358	0.244
K29	0.018	0.105	0.166	0.247	0.337	0.451	0.603	**0.670**	0.592	0.554	0.523	0.499	0.451	0.408	0.314
K30	0.294	0.379	0.451	0.542	0.613	0.660	**0.703**	0.670	0.595	0.510	0.408	0.306	0.210	0.130	0.046
K31	0.360	0.484	0.575	0.635	0.656	0.675	**0.683**	0.653	0.561	0.448	0.319	0.220	0.138	0.063	0.013
K32	0.168	0.222	0.288	0.379	0.455	0.526	0.599	**0.615**	0.583	0.520	0.443	0.363	0.281	0.211	0.125
K33	0.282	0.387	0.473	0.563	0.630	0.685	**0.740**	0.716	0.631	0.546	0.449	0.353	0.273	0.192	0.088
K34	0.092	0.163	0.257	0.357	0.455	0.541	0.610	**0.672**	0.658	0.644	0.634	0.618	0.622	0.621	0.549

3. Model

According to the analysis results of the above keywords, it can be observed that there is a significant positive correlation between the Baidu search index of 12 leading keywords and the weekly official reported ILI%. Based on this observation, the influenza prediction model (model 1) was first established using the leading keywords to verify whether search query data can reflect influenza transmission trends. Understanding the historical data of influenza is of great significance for predicting future trends. Therefore, model 2 was established to consider the influence of past influenza epidemics on the next moment's influenza to verify the impact of historical ILI% data in the southern region on ILI prediction. In addition, contact is an important pathway of influenza transmission, and with frequent population mobility, influenza can easily spread. Therefore, it is essential to incorporate the influenza level information from the northern region into the real-time influenza prediction model in the southern region to build model 3. The expression of the models are as follows:

$$\text{Model 1}:\quad ILI_t\% = \sum_{i=1}^{P} \alpha_i G_{i,t-1} + \varepsilon_t \ , \tag{1}$$

$$\text{Model 2}:\quad ILI_t\% = \sum_{i=1}^{P} \beta_i G_{i,t-1} + \sum_{j=1}^{M} \gamma_j S_{t-j} + o_t \ , \tag{2}$$

$$\text{Model 3}:\quad ILI_t\% = \sum_{i=1}^{P} \phi_i G_{i,t-1} + \sum_{j=1}^{M} \varphi_j S_{t-j} + \sum_{k=1}^{N} \eta_k N_{t-k} + \sigma_t \ , \tag{3}$$

where $ILI_t\%$ represents the ILI% of the southern region in the i-th week, $G_{i,t-1}$ denotes the Baidu search index of the i-th leading keyword, S_{t-j} represents the official ILI% of the southern region before the j-th week, N_{t-k} represents the official ILI% of the northern region before the k-th week, $P = 12$ indicates the number of leading keywords and M and N represent the lead orders of S_t and N_t, respectively. Through experimental verification, the model achieves the best predictive performance when $M = 4$ and $N = 3$, $\alpha_i, \beta_i, \gamma_j, \phi_i, \varphi_j$ and η_k are coefficients for each model, while ε_t, o_t and σ_t represent the residual terms of each model, respectively.

Wang et al. [37] suggested in their study that only leading keywords can be used to establish influenza prediction models. However, the results of cross-correlation analysis among keywords showed that the correlation coefficient between synchronous keywords and the number of ILI was all greater than 0.5 at a lag of 1 week. Therefore, this study considered incorporating the Baidu index of synchronous keywords at a lag of 1 week, along with leading keywords into the influenza prediction model to analyze their influence on the prediction accuracy of ILI. Considering that there are a large number of keywords and certain correlation among them, the information in the data overlaps to some extent. In order to reduce the number of variables and retain the main information, this study used principal component analysis to process the input Baidu keywords and extracted the principal components that contributed to 90% of the variance as the input variables for the model. After the analysis, when 7 principal components are selected, the cumulative contribution rate of keywords reached 95.58%. Therefore, these 7 principal components of the keywords were selected as inputs for the influenza prediction model (model 4). The specific model formula is as follows:

$$\text{Model 4:}\quad ILI_t\% = \sum_{l=1}^{Q} \mu_l Z_{l,t-1} + \sum_{j=1}^{M} \overline{\varphi_j} S_{t-j} + \sum_{k=1}^{N} \overline{\eta_k} N_{t-k} + \overline{\sigma_t}\,, \tag{4}$$

where $ILI_t\%$, S_{t-j}, N_{t-k}, M, N, μ_l, $\overline{\varphi_j}$, $\overline{\eta_k}$ and $\overline{\sigma_t}$ are the same as represented in Equations (1)–(3), $Z_{l,t-1}$ represents the value of the l-th principal component at the time $t-1$ and $Q=7$ indicates the number of search principal components included in the model.

4. Model prediction and analysis

4.1. Research method

4.1.1. Support vector regression

Support vector regression (SVR) is a machine learning method based on statistical learning theory. It employs the criterion of structural risk minimization, which aims to minimize the error of sample points while also maximizing the model's generalization ability. It is a convex quadratic optimization problem, ensuring that the extreme value found is the globally optimal solution [38]. SVR can be used to capture complex nonlinear relationships in the real world. Its main idea is to find a regression plane that minimizes the distance of all training points to that plane.

In a typical regression problem, given the training set:

$$G = \{(x_i, y_i)\}_i^l \subset R^d \times R\,, \tag{5}$$

where $x_i \in R^d$ is the input vector, $y_i \in R$ is the output variable and l represents the number of samples.

The modeling purpose of the nonlinear SVR is to map x into a high-dimensional feature space through a nonlinear mapping φ, and then determine the linear regression function $y = f(x)$ in that space to fit the data (x_i, y_i), which can be expressed as:

$$f(x) = \omega * \varphi(x) + b,\ \varphi : R^d \to F, \omega \in F\,, \tag{6}$$

where ω is the weight vector and b is the threshold, which are estimated by the training set G, $\varphi(x)$ represents the nonlinear mapping function that maps the input vector to a high-dimensional feature space F. Therefore, the linear regression in the high-dimensional feature space corresponds to nonlinear regression in the low-dimensional input space, while the inner product calculation between ω and $\varphi(x)$ in the high-dimensional feature space is ignored.

Based on the principle of structural risk minimization, the objective functions and constraints of SVR are defined as follows:

$$\min_{\omega, b, \xi_i, \overline{\xi_i}} \frac{1}{2}\omega^T \omega + C\sum_{i=1}^{l} (\xi_i + \overline{\xi_i})\,, \tag{7}$$

$$s.t.\ \begin{cases}[\omega^T * \varphi(x)+b]-y_i \le \varepsilon + \xi_i \\ y_i - [\omega^T * \varphi(x)+b] \le \varepsilon + \overline{\xi_i}\ , \\ \xi_i \ge 0, \overline{\xi_i} \ge 0, i=1,2,\cdots,l\end{cases} \tag{8}$$

where C is the trade-off parameter that adjusts the balance between regression error and regularization term, l is the number of training samples, $\xi_i(\overline{\xi_i})$ is the relaxation variable that allows for the error range of the regression function and $\varepsilon \ge 0$ is the parameter in the insensitive loss function of $\varepsilon -$, which is used to control the accuracy of the regression approximation.

By introducing Lagrange multipliers α and $\overline{\alpha}$, the quadratic programming problem can be optimized into a dual problem, then the dual problem of equation (7) can be written as:

$$\begin{aligned}&\max_{\alpha,\overline{\alpha}} \sum_{i=1}^{l} y_i(\overline{\alpha_i}-\alpha_i) - \varepsilon \sum_{i=1}^{l}(\overline{\alpha_i}+\alpha_i) \\ &-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\overline{\alpha_i}-\alpha_i)(\overline{\alpha_j}-\alpha_j)K(x_i,x_j) \\ &s.t.\ \sum_{i=1}^{l}(\overline{\alpha_i}-\alpha_i) \\ &C \ge \alpha_i, \overline{\alpha_i} \ge 0, i=1,2,\cdots,l\end{aligned} \tag{9}$$

where $\alpha = \{\alpha_1,\cdots,\alpha_l\}$ and $\overline{\alpha} = \{\overline{\alpha_1},\cdots,\overline{\alpha_l}\}$ are dual variables and $K(x_i,x_j)$ is the kernel function representing the inner product $\langle \varphi(x_i), \varphi(x_j)\rangle$.

By utilizing the Karush-Kuhn-Tucker (KKT) conditions to solve for α_i, $\overline{\alpha_i}$ and b in Equation (9), the regression function is as follows:

$$f(x) = \sum_{i=1}^{l}(\overline{\alpha_i}-\alpha_i)K(x_i,x)+b\ . \tag{10}$$

For the training of SVR method, the first step is to determine the kernel function. At present, several kernel functions have been proposed, but there is no theoretical solution for selecting the optimal kernel function, and the trial-and-error method is usually adopted [39]. In this paper, through iterative tests, radial basis function (RBF) is employed as the basic kernel function, expressed as follows:

$$K(x_i,x_j) = \exp(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2})\ . \tag{11}$$

4.1.2. Improved particle swarm optimization algorithm

Particle swarm optimization (PSO) algorithm, based on swarm intelligence, is one of the widely used methods in SVR parameter optimization calculation. It does not require gradient information during the iterative process and involves a relatively small number of adjustable parameters. This algorithm is known for its advantages such as ease of implementation, high efficiency and fast convergence speed [40,41]. The PSO algorithm can be described as follows: the particle swarm consists of m particles in the n-dimensional search space, The velocity state vector is composed of four parts: $x_i=(x_{i1},x_{i2},\cdots,x_{ij},\cdots,x_{in})^T, i=1,2,\cdots,m$ is the current position of the i-th particle in the search space, $v_i=(v_{i1},\cdots,v_{ij},\cdots,v_{in})^T$ is the velocity of the i-th particle, $p_i=(p_{i1},\cdots,p_{ij},\cdots,p_{in})^T$ represents the optimal position of the i-th particle at the current moment and $p_g=(p_{g1},\cdots,p_{gj},\cdots,p_{gn})^T$ represents the optimal position of the particle swarm in each iteration. The velocity and position of each particle are updated according to Equations (12) and (13):

$$v_{ij}^{k+1}=\omega v_{ij}^{k}+c_1r_1^k(p_{ij}^k-x_{ij}^k)+c_2r_2^k(p_{gj}^k-x_{ij}^k)\,, \tag{12}$$

$$x_{ij}^{k+1}=x_{ij}^{k}+v_{ij}^{k+1}\,, \tag{13}$$

where v_{ij}^k is the velocity of the j-th component of the i-th particle in the k-th iteration, ω is the inertia weight, c_1 and c_2 are cognitive learning factors and social learning factors and r_1^k and r_2^k are random numbers generated within the interval $(0,1)$.

Inertia weight ω plays a crucial role in the performance of PSO, as it balances the global search ability and local search ability of particles [42]. A large inertia weight enhances the algorithm’s global search ability, but it may lead to lower search efficiency. In contrast, a smaller inertia weight is beneficial for local search, but may lead to local optimality.

$$\omega=\omega_{\max}\cdot(w_{\min}/\omega_{\max})^{\frac{t-1}{(t_{\max}-1)^2}}\,, \tag{14}$$

where $\omega_{\max}$ and $w_{\min}$ are the maximum and minimum values of the inertia weight ω, respectively, t is the current iteration number and $t_{\max}$ is the maximum number of iterations. In Equation (14), ω gradually decreases during the search process, which satisfies the requirements of the adaptive process for the algorithm from global optimization to local optimization.

4.1.3. Improved particle swarm optimization-based SVR method

In the SVR method based on the RBF kernel function, C and σ (kernel width) are two adjustable parameters that play a crucial role in the performance of SVR [43,44]. In this study, an improved PSO algorithm is utilized to optimize the parameters C and σ of SVR. The method of optimizing SVR parameters using the improved PSO algorithm is referred to as IPSO-SVR, and the basic steps are summarized as follows:

Step 1: Initialize all the parameters of the algorithm, including the maximum number of iterations $t_{\max}$, population size, cognitive learning factor c_1 and social learning factor c_2, velocity range $[V_{\min},V_{\max}]$, etc.

Step 2: The population and speed are generated randomly, and the initial fitness value of each particle is calculated using Equation (15) for evaluation. x_i is set to p_i, and the particle with the best fitness is set to p_g.

$$fit_i = \frac{1}{m}\sum_{i=1}^{m}(\overline{y_i} - y_i)^2 , \tag{15}$$

where $\overline{y_i}$ is the predicted value and y_i is the true value.

Step 3: The velocity, position and inertia weights of the particles are updated according to Equations (12)–(14). Evaluate the fitness function for each particle and compare it with p_i. If the fitness value fit_i of the i-th particle is less than p_i, set x_i to p_i; otherwise, p_i is left unchanged. If the fitness value fit_i of the i-th particle is less than p_g, set x_i to p_g; otherwise, the original value is retained.

Step 4: Determine whether the termination conditions are met. If the condition is satisfied, proceed to the next step; otherwise, go back to step 2;

Step 5: The best parameters C_{best} and σ_{best} of the SVR model were obtained, an SVR model with C_{best} and σ_{best} as parameters was established by using the training set and the trained model was used to predict ILI% in southern China.

4.2. Model evaluation index

To validate the predictive performance of each model, the MSE, RMSE and MAE were used to evaluate the prediction results of each model, as shown in equations (16)–(18). MSE represents the mean of the squared prediction errors, RMSE represents the square root of the mean of the squared differences between predicted and true values, divided by the sample size m, which is used to measure the deviation of the overall prediction results from the actual values, and MAE is the mean of the absolute errors, accurately reflecting the actual predicted error situation.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y_i})^2 , \tag{16}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y_i})^2} , \tag{17}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - \widehat{y_i}\right| , \tag{18}$$

where m represents the number of samples, y_i represents the true value of ILI% and $\widehat{y_i}$ represents the predicted values of ILI%.

4.3. Prediction results and analysis

The prediction target of this study is the ILI% in the southern region of China. Due to the fact that both domestic and international scholars often adopt traditional multiple linear regression methods when choosing influenza trend prediction methods [28,29], this paper uses the prediction results of this method as a comparative baseline to analyze and compare them with the prediction results of SVR, GA-SVR, PSO-SVR and IPSO-SVR methods. The construction code of the model was implemented using Python 3.8.12 software. In this study, 164 weeks of data from week 1 of

2018 to week 7 of 2021 were selected as training samples and 42 weeks of data from week 8 of 2021 to week 49 of 2021 were selected as test samples.

The independent variables from models 1–4 were used as inputs for multiple linear regression, SVR, GA-SVR, PSO-SVR and IPSO-SVR, respectively, ILI_t as the output. Training samples were used to train each model, and the trained models were utilized to predict the ILI% for the southern region of China from week 8 to week 49 of 2021. The MSE, RMSE and MAE results of each model's test sample are shown in Table 4, where LR represents the prediction results of the multiple linear regression method.

Table 4. The evaluation index results for the five methods in models 1-4.

Model	Method	LR	SVR	GA-SVR	PSO-SVR	IPSO-SVR
Model 1	MSE	**0.2845**	0.3974	0.3570	0.3581	0.3027
	RMSE	**0.5334**	0.6304	0.5975	0.5984	0.5502
	MAE	**0.4216**	0.5426	0.5091	0.4642	0.4488
Model 2	MSE	0.0800	0.1053	0.0977	0.0931	**0.0765**
	RMSE	0.2829	0.3244	0.3126	0.3051	**0.2765**
	MAE	0.2227	0.2573	0.2455	0.2347	**0.2058**
Model 3	MSE	0.0736	0.0968	0.0821	0.0834	**0.0625**
	RMSE	0.2712	0.3112	0.2865	0.2888	**0.2501**
	MAE	0.2114	0.2394	0.2133	0.2163	**0.2033**
Model 4	MSE	0.0781	0.0834	0.0751	0.0723	**0.0596**
	RMSE	0.2794	0.2887	0.2740	0.2689	**0.2441**
	MAE	0.2177	0.2339	0.2178	0.2144	**0.1884**

Comparing the MSE, RMSE and MAE results of the five prediction methods for each model in Table 4, it can be observed that, when compared with the traditional multiple linear regression, SVR, GA-SVR and PSO-SVR methods, the prediction results of IPSO-SVR are the best in models 2–4, demonstrating superior prediction performance. Among the SVR, GA-SVR, PSO-SVR and IPSO-SVR methods, model 4 demonstrates the best predictive performance. However, within the LR method, model 3 exhibits the most favorable prediction effectiveness. From the IPSO-SVR prediction results of model 1, it can be observed that when ILI% is predicted by the leading keywords, although there is some discrepancy between the predicted results and the true values, the trend of the predictions is relatively consistent with the true values. By comparing the three evaluation index results of the IPSO-SVR algorithm in model 1 and model 2, it can be found that by adding historical ILI% data from the southern region, the MSE, RMSE and MAE index results of the model were reduced by 74.7%, 49.7% and 54.1%, respectively. This indicates that the historical ILI data contains a significant amount of influenza epidemic trend information.

By comparing the three evaluation index results of the IPSO-SVR algorithm in model 2 and model 3, it can be observed that adding historical ILI% data from the northern region led to a reduction of 18.3%, 9.5% and 1.2% in the model's MSE, RMSE, and MAE index, respectively. This indicates that the influenza epidemic in the northern region has some impact on the southern region, which means that influenza transmission can be affected by interregional transmission. Therefore, when analyzing and forecasting ILI, it is essential to consider not only the impact of ILI in the current region but also the influence of the epidemic situation in other regions on the current region.

By comparing the three evaluation index results of the IPSO-SVR algorithm in model 2 and model 3, it can be observed that adding the Baidu index of synchronous keywords from the previous week can reduce the MSE, RMSE and MAE index results of the model by 4.6%, 2.4% and 7.3%, respectively. This indicates that incorporating synchronous keywords into the model can improve its predictive accuracy. Therefore, when establishing influenza prediction model based on web search data, the information of synchronous keywords should not be directly excluded. Instead, the influence of synchronous keywords on influenza prediction should be further analyzed by constructing models to assess their impact.

The comparison of fitted values, actual values and predicted values for the training and testing samples using five forecasting methods in models 1–4 is shown in Figures 3–6. In each model prediction result graph, the subgraph is divided into two parts by a vertical deep red line along the horizontal axis: the left part shows the actual values (red) and fitted values of models 1–4 on the training sample, while the right part displays the actual values (red) and predicted values of models 1–4 on the testing samples. By comparing the prediction results of the five methods in model 1 to model 4, it can be found that the prediction output of model 4 is closer to the real value of both the training set and the test set. Regardless of the fitting and prediction time periods, the IPSO-SVR method in model 4 can capture the peaks and troughs of the time series curve of ILI, and the prediction effect is better than that of other models.

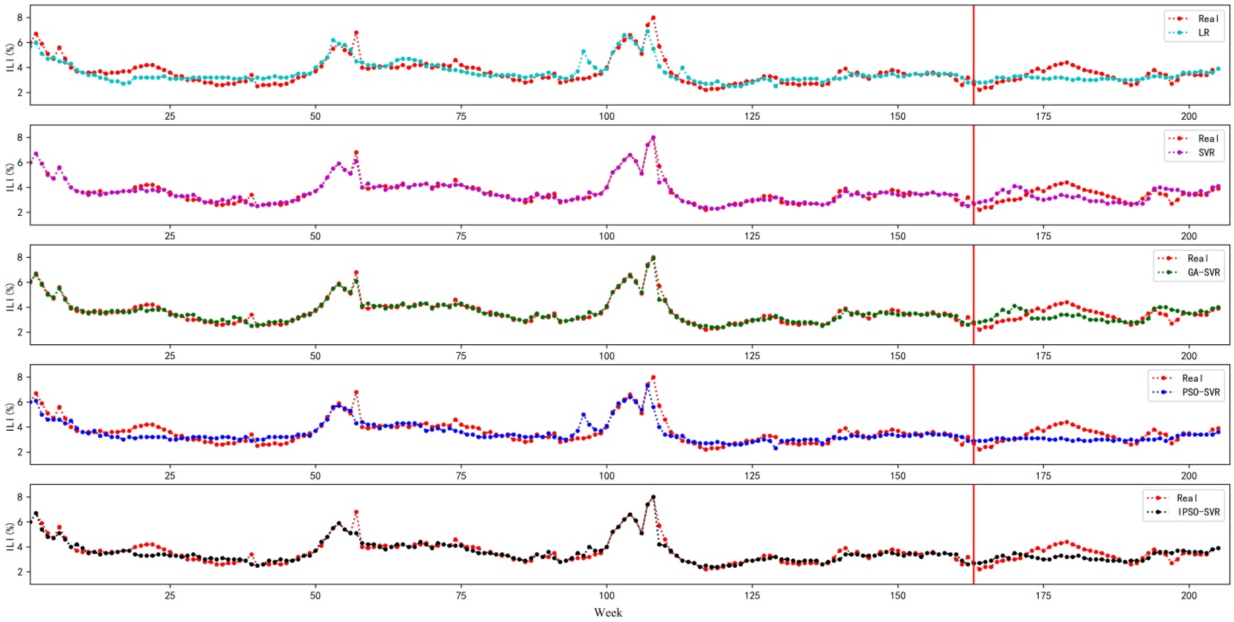

Figure 3. Training and testing results of the five methods in model 1.

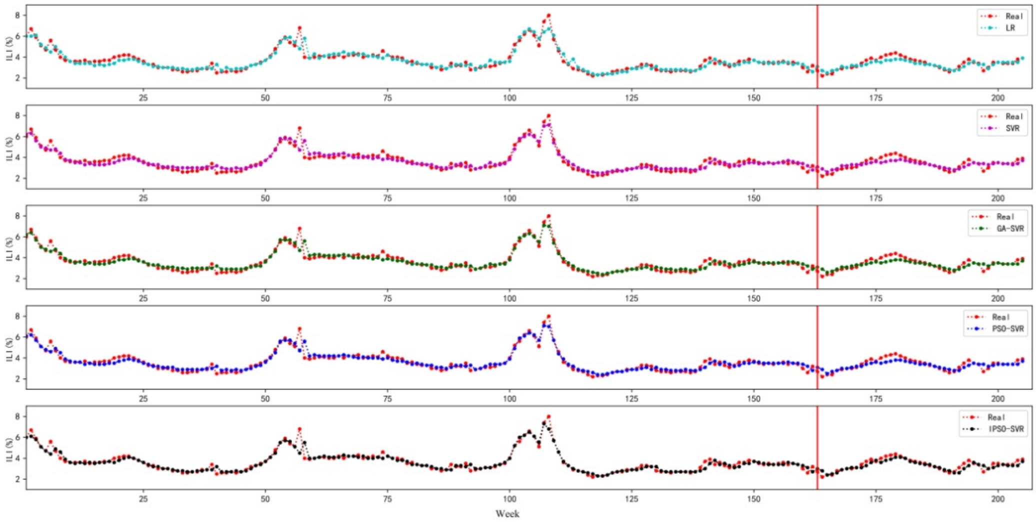

Figure 4. Training and testing results of the five methods in model 2.

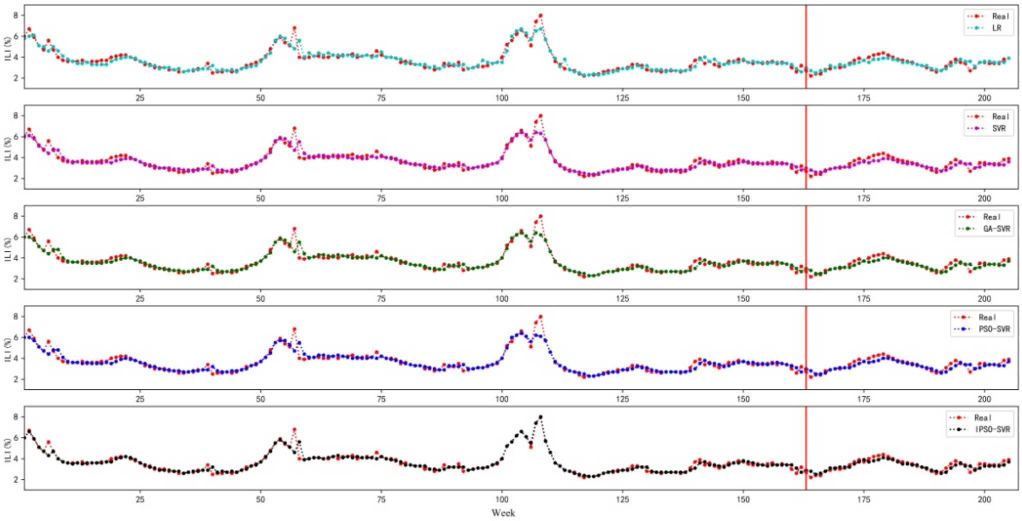

Figure 5. Training and testing results of the five methods in model 3.

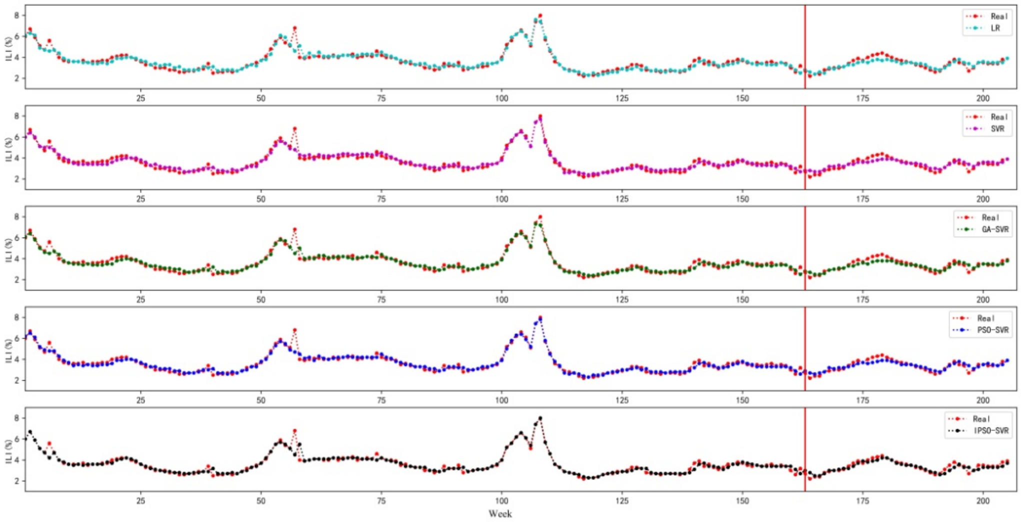

Figure 6. Training and testing results of the five methods in model 4.

5. Conclusions

Influenza is a common respiratory disease that can lead to illness and death in humans. Timely and accurate prediction of disease risks is essential for public health management and prevention. While various prediction efforts regarding infectious diseases have matured, the current infectious disease surveillance system model is excessively passive, heavily reliant on case reporting and there is a large time lag. Additionally, the geographical distribution and genetic diversity of novel influenza viruses are rapidly expanding, presenting a direct challenge to the existing disease control system in China. To achieve near real-time monitoring of influenza spread, both domestic and international scholars have proposed influenza prediction methods based on informal sources of data, such as news reports, social media data, online search query data and electronic health information records. However, there is few research on the domestic influenza epidemic in this field. Many existing methods solely utilize historical time series data for prediction, overlooking the impact of spatial correlations among neighboring regions and temporal correlations across different time periods. Additionally, influenza prediction methods often heavily rely on the use of multivariate linear regression techniques.

In this study, an attempt was made to identify significant keywords related to influenza, followed by an initial screening of these keywords. By analyzing the time-delay correlation between each keyword and ILI, the keywords were further filtered and screened. Secondly, based on the identified distinct types of keywords and considering the influence of influenza transmission between neighboring regions, the influenza prediction model suitable for the characteristics of the southern region of China was constructed. The model can comprehensively consider spatial and temporal correlations, providing a more accurate reflection of the influenza transmission trends in the region. Finally, an improved PSO-based SVR method was proposed for model prediction, and its prediction results were compared and analyzed with multiple linear regression, SVR, GA-SVR and PSO-SVR methods.

By comparing the prediction results of each model, the following conclusions were drawn: 1) The influenza epidemic in the northern region has some impact on the southern region, indicating that influenza transmission is influenced by interregional spread. 2) When establishing influenza prediction models based on web search data, the information of synchronous keywords should not be excluded directly. Instead, their impact on influenza prediction should be analyzed through further modeling. 3) The IPSO-SVR method used in model 4 can capture the peaks and troughs in the time series curve of ILI, which has higher prediction accuracy and a better effect, and can better reflect the real level of influenza.

In this study, the integration of Baidu search data and machine learning methods was employed to construct a series of influenza trend prediction models, along with the incorporation of the IPSO-SVR algorithm as a predictive tool. This innovative approach introduces a novel predictive framework to the field of influenza trend forecasting, providing essential decision support for public health management and epidemic prevention and control. By constructing various prediction models, this study has unveiled multiple factors that influence the spread of influenza, thereby enhancing our understanding of the mechanisms underlying influenza transmission. The significant impact of introducing the IPSO-SVR algorithm in enhancing the accuracy of predictions is particularly noteworthy. This optimization algorithm demonstrates promising potential in influenza trend prediction, offering a novel avenue to improve the accuracy of prediction outcomes. By incorporating this algorithm into the model, it becomes possible to capture the dynamic changes in influenza trends with greater precision, which provides new perspectives and approaches for research and application of influenza prediction.

The paper still has some limitations. For instance, the scope of the study is confined to influenza forecasting in the southern region of China, and the prediction performance in other regions has not undergone sufficient in-depth research. Further validation and expansion are necessary in this regard. In reality, there might be cases of cross-infection and mutual influence among different diseases. ILI% data could be affected by these underlying factors. Therefore, when conducting influenza prediction, it is crucial to take into account the impact of other relevant disease data to mitigate the prediction errors arising from multifactorial influences. In future research, further exploration can be conducted on how to incorporate additional disease data into the predictive model, aiming to enhance the accuracy of influenza trend prediction and further improve the reliability and applicability of the predictive model. Furthermore, the exploration of more advanced machine learning techniques and data analysis methods will be pursued to optimize the performance of the influenza prediction model. By introducing new technological approaches, there is a potential to further enhance the predictive capabilities of the model across various regions, offering more forward-looking and practical solutions for research and practical applications in the field of influenza prediction.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China granted [No. 62106238], Fundamental Research Program of Shanxi Province [No. 202103021224195], Shanxi Province Science and Technology Major Special Project "Revealing the Leader" Program [No. 202201150401021] and Provincial Natural Science Foundation of Shanxi [No. 202203021212138].

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. Z. Y. Zhao, M. M. Zhai, G. H. Li, X. F. Gao, W. Z. Song, X. C. Wang, Study on the prediction effect of a combined model of SARIMA and LSTM based on SSA for influenza in Shanxi Province, China, *BMC INFECT. DIS.*, **23** (2023), 71. https://doi.org/10.1186/s12879-023-08025-1
2. H. Gong, X. Shen, H. Yan, W. Y. Lu, G. J. Zhong, K. G. Dong, et al., Estimating the disease burden of seasonal influenza in China, 2006-2019, *Natl. Med. J. China*, **101** (2021), 560–567. https://doi.org/10.3760/cma.j.cn112137-20201210-03323

3. World Health Organization, Fact sheet on influenza (seasonal). Available from: https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)
4. L. Li, Y. Liu, P. Wu, Z. Peng, X. Wang, T. Chen, et al., Influenza-associated excess respiratory mortality in China, 2010-2015: a population-based study, *Lancet Public Health*, **4** (2019), e473–e481. https://doi.org/10.1016/S2468-2667(19)30163-X
5. National Bureau of Statistics, Gross domestic product. Available from: http://www.stats.gov.cn/xxgk/sjfb/zxfb2020/202212/t20221227_1891261.html
6. B. Jang, I. Kim, W. Jong, Effective training data extraction method to improve influenza outbreak prediction from online news articles: deep learning model study, *JMIR Med. Inf.*, **9** (2021), e23305. https://doi.org/10.2196/23305
7. L. Zhou, J. Li, D. Shi, L. Xu, S. X. Huang, Predicting Influenza Epidemic for United States, *Int. J. Environ. Heal. R.*, **32** (2022), 1231–1237.
8. P. Guo, J. J. Zhang, L. Wang, S. Y. Yang, G. F. Luo, C. Y. Deng, et al., Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model, *Sci. Rep-UK.*, **7** (2017), 46469. https://doi.org/10.1038/srep46469
9. S. Y. Yang, Y. K. Bao, Comprehensive learning particle swarm optimization enabled modeling framework for multi-step-ahead influenza prediction, *Appl. Soft Comput.*, **113** (2021), 107994. https://doi.org/10.1016/j.asoc.2021.107994
10. R. X. Wang, H. Y. Wu, Y. S. Wu, J. Zheng, Y. Li, Improving influenza surveillance based on multi-granularity deep spatiotemporal neural network, *Comput. Biol. Med.*, **134** (2021), 104482. https://doi.org/10.1016/j.compbiomed.2021.104482
11. N. Kumar, H. Kumar, K. Kumar, A study for plausible third wave of COVID-19 in India through fuzzy time series modelling based on particle swarm optimization and fuzzy c-means, *Math. Probl. Eng.*, **2022** (2022), 5878268. https://doi.org/10.1155/2022/5878268
12. M. Thomas, H. Rootzen, Real-time prediction of severe influenza epidemics using extreme value statistics, *J. R. Stat. Soc. C-Appl.*, **71** (2022), 376-394. https://doi.org/10.1111/rssc.12537
13. Y. C. Wei, Y. L. Ou, J. Q. Li, W. C. Wu, Forecasting the potential number of influenza-like illness cases by fusing internet public opinion, *Sustainability-Basel*, **14** (2022), 2803. https://doi.org/10.3390/su14052803
14. A. Kara, Multi-step influenza outbreak forecasting using deep LSTM network and genetic algorithm, *Expert Syst. Appl.*, **180** (2021), 115153. https://doi.org/10.1016/j.eswa.2021.115153
15. N. Kumar, H. Kumar, A novel hybrid fuzzy time series model for prediction of COVID-19 infected cases and deaths in India, *ISA T.*, **124** (2022), 69–81. https://doi.org/10.1016/j.isatra.2021.07.003
16. S. F. Ackley, S. Pilewski, V. S. Petrovic, L. Worden, E. Murray, T. C. Porco, Assessing the utility of a smart thermometer and mobile application as a surveillance tool for influenza and influenza-like illness, *Health Inform. J.*, **26** (2020), 2148–2158. https://doi.org/10.1177/1460458219897152
17. T. Murayama, N. Shimizu, S. Fujita, S. Wakamiya, E. Aramaki, Predicting regional influenza epidemics with uncertainty estimation using commuting data in Japan, *PLoS One*, **16** (2021), e0250417. https://doi.org/10.1371/journal.pone.0250417
18. C.Y. Yang, R. J. Chen, W. L. Chou, Y. J. Lee, Y. S. Lo, An Integrated Influenza Surveillance Framework Based on National Influenza-Like Illness Incidence and Multiple Hospital

Electronic Medical Records for Early Prediction of Influenza Epidemics: Design and Evaluation, *J. Med. Internet Res.*, **21** (2019), e12341. https://doi.org/10.2196/13699

19. S. I. Leuba, R. Yaesoubi, M. Antillon, T. Cohen, C. Zimmer, Tracking and predicting US influenza activity with a real-time surveillance network, *PLoS Comput. Biol.*, **16** (2020), e1008180.
20. B. Jang, L. Kim, J. W. Kim, Long-Term Influenza Outbreak Forecast Using Time-Precedence Correlation of Web Data, *IEEE T. Neur. Net. Lear.*, **34** (2023), 2400–2412. https://doi.org/10.1371/journal.pcbi.1008180
21. D. Viglino, A. Vesin, S. Ruckly, X. Morelli, R. Slama, G. Debaty, et al. Daily volume of cases in emergency call centers: construction and validation of a predictive model, *Scand. J. Trauma Resus.*, **25** (2017): 86. https://doi.org/10.1186/s13049-017-0430-9
22. A. H. Gutierrez, V. J. Rapp-Gabrielson, F. E. Terry, C. L. Loving, L. Moise, W. D. Martin, et al., T-cell epitope content comparison (EpiCC) of swine H1 influenza A virus hemagglutinin, *Influenza Other Resp.*, **11** (2018), 531–542. https://doi.org/10.1111/irv.12513
23. S. N. Chen, J. Xu, Y. S. Wu, X. Wang, S. S. Fang, J. Q. Cheng, et al., Predicting temporal propagation of seasonal influenza using improved gaussian process model, *J. Biomed. Inform.*, **93** (2019). https://doi.org/103144. 10.1016/j.jbi.2019.103144
24. F. S. Lu, M. W. Hattab, C. L. Clemente, M. Biggerstaff, M. Santillana, Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches, *Nat. Commun.*, **10** (2019), 1–10. https://doi.org/10.1038/s41467-018-08082-0
25. C. Zimmer, S. I. Leuba, R. Yaesoubi, T. Cohen, Use of daily Internet search query data improves real-time projections of influenza epidemics, *J. R. Soc. Interface*, **15** (2018), 1–7. https://doi.org/10.1098/rsif.2018.0220
26. I. Miliou, X. Xiong, S. Rinzivillo, Q. Zhang, G. Rossetti, F. Giannotti, et al., Predicting seasonal influenza using supermarket retail records, *PLoS Comput. Bilo.*, **17** (2021), e1009087. https://doi.org/10.1371/journal.pcbi.1009087
27. Z. Y. Huang, Exploration of the Accuracy of epidemic prediction based on the Baidu index——taking H7N9 subtype avian influenza in Guangdong Province as an Example, *Chinese Journal of Zoonoses*, **36** (2020), 962–968.
28. Y. Lu, S. Wang, J. Y. Wang, G. Y. Zhou, Q. Zhang, X. Zhou, et al., An Epidemic Avian Influenza Prediction Model Based on Google Trends, *Lett. Org. Chem.*, **16** (2019), 303–310. https://doi.org/10.2174/1570178615666180724103325
29. X. Y. Zhou, Y. Zhang, C. J. Shen, A. L. Liu, Y. M. Wang, Q. Yu, et al., Knowledge, attitudes, and practices associated with avian influenza along the live chicken market chains in Eastern China: A cross-sectional survey in Shanghai, Anhui, and Jiangsu, *Transbound. Emerg. Dis.*, **66** (2019), 1529–1538. https://doi.org/10.1111/tbed.13178
30. M. Athanasiou, G. Fragkozidis, K. Zarkogianni, K. S. Nikita, Long short-term memory-based prediction of the spread of influenza-like illness leveraging surveillance, weather, and twitter data: model development and validation, *J. Med. Internet Res.*, **25** (2023), e42519. https://doi.org/10.2196/42519
31. T. Lazebnik, S. Bunimovich-Mendrazitsky, S. Ashkenazi, E. Levner, A. Benis, Early detection and control of the next epidemic wave using health communications: development of an artificial intelligence-based tool and its validation on COVID-19 data from the US, *Int. J. Env. Res. Pub. He.*, **19** (2022), 16023. https://doi.org/10.3390/ijerph192316023

32. C. Wu, S. C. Kao, Knowledge discovery in open data for epidemic disease prediction, *Health Policy Techn.*, **10** (2021), 126–134. https://doi.org/10.1016/j.hlpt.2021.01.001
33. S. B. Choi, J. Kim, I. Ahn, Forecasting type-specific seasonal influenza after 26 weeks in the United States using influenza activities in other countries, *PLoS One*, **14** (2019), e0220423. https://doi.org/10.1371/journal.pone.0220423
34. A. Boukobza, A. Burgun, B. Roudier, R. Tsopra, Deep neural networks for simultaneously capturing public topics and sentiments during a pandemic: application on a COVID-19 Tweet data set, *JMIR Med. Inf.*, **10** (2022), e34306. https://doi.org/10.2196/34306
35. Chinese National Influenza Center, Weekly report of influenza-like cases. Available from: https://ivdc.chinacdc.cn/cnic/
36. Baidu search engine, Bai index. Available from: https://index.baidu.com
37. R. J. WANG, Machanism and empirical research on forecasting influenza epidemic fused with Baidu index, *Journal of the China society for scientific and technical information*, **37** (2018), 206–219.
38. N. Sultana, N. Sharma, K. P. Sharma, S. Verma, A Sequential Ensemble Model for Communicable Disease Forecasting, *Curr. Bioinform.*, **15** (2020), 309–317. https://doi.org/10.2174/1574893614666191202153824
39. H. S. Cai, X. D. Jia, J. S. Feng, W. Z. Li, Y. M. Hsu, J. Lee, Gaussian Process Regression for numerical wind speed prediction enhancement, *Renew. Energ.*, **146** (2020), 2112–2123. https://doi.org/10.1016/j.renene.2019.08.018
40. B. J. Zhang, L. Sun, W. B. Wang, Two stage prediction model of sunspots monthly value based on CEEMDAN and particle swarm optimization ELM, *IEEE Access*, **10** (2022), 102981–102991. https://doi.org/10.1109/ACCESS.2022.3206542
41. Y. P. Wen, Y. Wang, J. X. Liu, B. Q. Cao, Q. Fu, CPU usage prediction for cloud resource provisioning based on deep belief network and particle swarm optimization, *Concurr. Comp.-Pract. E.*, **32** (2020), e5730. https://doi.org/10.1002/cpe.5730
42. W. P. Gong, S. Tian, L. Wang, Z. B. Li, H. M. Tang, T. Z. Li, et al., Interval prediction of landslide displacement with dual-output least squares support vector machine and particle swarm optimization algorithms, *Acta Geotech.*, **17** (2022), 4013–4031. https://doi.org/10.1007/s11440-022-01455-2
43. Q. Ma, H. Wang, P. Luo, Y. S. Peng, Q. R. Li, Ultra-short-term Railway traction load prediction based on DWT-TCN-PSO_SVR combined model, *Int. J. Elec. Power.*, **135** (2022), 107595. https://doi.org/10.1016/j.ijepes.2021.107595
44. C. L. Dong, X. Meng, L. X. Guo, J. M. Hu, 3D sea surface electromagnetic scattering prediction model based on IPSO-SVR, *Remote Sens.-Basel.*, **14** (2022), 4657. https://doi.org/10.3390/rs14184657