Methods and Applications

A New Integrated Interpolation Method for High Missing Unstable Disease Surveillance Data — 12 Urban Agglomerations, China, 2009–2020

Yuanhao Shi[1,2]; Yilan Liao[1,#]

ABSTRACT

Introduction: The prevalence of unstable and incomplete monitoring data significantly complicates syndromic analysis. Many data interpolation methods currently available demonstrate inadequate effectiveness in overcoming this issue.

Methods: To improve the accuracy of interpolation, we propose the integration of the SHapley Additive exPlanation model (SHAP) with the structural equation model (SEM), forming a combined SHAP-SEM approach. A case study is then performed to assess the enhanced performance of this novel model compared to traditional methods.

Results: The SHAP-SEM model was utilized to develop an interpolation model employing data from the Chinese respiratory syndrome surveillance database. We executed three distinct experiments to establish the model datasets, comprising a total of 100 replicates. The performance of the model was evaluated using the root mean square error (RMSE), correlation coefficient (r), and F-score. The findings demonstrate that the SHAP-SEM model consistently achieves superior accuracy in data interpolation, which is evident across different seasons and in overall performance.

Discussion: We conclude that the SHAP-SEM model demonstrates an exceptional capacity for accurately interpolating volatile and incomplete data. This capability is crucial for developing a comprehensive database that is essential for conducting risk assessments related to syndromes.

Syndrome surveillance is crucial for the rapid detection and alerting of infectious disease outbreaks. Nonetheless, it often encounters challenges including uneven distribution of monitoring sites, irregular reporting schedules, and incomplete data (*1*). These factors hinder the ability to accurately delineate disease distribution temporally and spatially, and to discern patterns and anomalies. Traditional spatio-temporal interpolation methods (*2*) are ill-suited for addressing the volatility and gaps in disease data, particularly when integrating significant influencing factors. Conversely, the structural equation model (SEM) facilitates analysis of complex data interactions to unearth underlying relationships among variables (*3*), thus enabling more precise interpolation. We suggest employing a SHapley Additive explanation (SHAP)-based SEM for the spatiotemporal interpolation of unstable and incomplete data.

METHODS

Shapley Additive ExPlanation Model with The Structural Equation Model (SHAP-SEM)

The SHAP model is utilized to evaluate multiple variables and their correlations with observed variables (*4*). It quantifies the marginal contribution of each variable to the model output, thereby demonstrating its influence on the overall model. This study aims to explore the relationship between meteorological variables and the virus positivity rate to identify the most impactful combination of these factors. Subsequently, the SEM is applied to discern the connections between the factors and observed variables (*3*), facilitating spatio-temporal interpolation. The SEM illustrates the network of relationships between meteorological variables and the virus, satisfying interpolation requirements based on the model fit. For further information on the SHAP-SEM model and additional classical interpolation methods, readers are referred to the Supplementary Material (available at https://weekly.chinacdc.cn/).

Model Assessment

In this study, the root mean square error (RMSE), correlation coefficient (r), and F-score were utilized to assess the accuracy of the SAHP-SEM and other comparative models.

RMSE measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model can predict the target value (*5*).

The r quantifies the extent and direction of a linear relationship between two variables. Our research explores various methodologies, identifying that while Kriging interpolation displays a high determination coefficient, it unfortunately correlates poorly with the original dataset. This misalignment highlights its inability to accurately represent the underlying data trends. Consequently, due to the need for a precise reflection of these trends, we have chosen the r as our preferred metric.

The F-score is a composite metric that favors algorithms with higher sensitivity while penalizing those with higher specificity. It is derived from the precision and recall of the test. Here, precision refers to the number of true positive results divided by the total count of samples predicted as positive, inclusive of false positives. Recall, alternatively, is the number of true positives divided by the total number of samples that are correctly identified as positive (*6*). The F-score is formulated as follows (Equations 1–3):

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

TP, FP, and FN denote true positives, false positives, and false negatives, respectively, across different classes. Precision quantifies the accuracy across various classes, recall indicates the detection rate, and the F-score provides a balanced measure of precision and recall.

Empirical Study

This study utilizes two primary data types: symptom data and meteorological data. Symptom data was derived from the China CDC's febrile respiratory syndrome surveillance, encompassing patient demographic profiles, clinical characteristics, and laboratory results obtained from 330 hospitals nationwide (Supplementary Figure S1, available at https://weekly.chinacdc.cn/). Meteorological data were sourced from the Meteorological Station Sharing Service System (https://data.cma.cn/). Selected meteorological parameters included atmospheric pressure, average relative humidity, mean temperature, maximum and minimum temperatures, and the range between the highest and lowest temperatures. Details on the variable utilization within the model are described in the Supplementary Material. For analysis, the SHAP-SEM model was applied to the China respiratory syndrome monitoring database to assess its effectiveness in disease interpolation. The period of study spanned from January 1, 2009 to January 4, 2020, with quarters serving as the smallest time unit. The study's scope of respiratory syndrome detection included various viruses: influenza virus (IFV), human respiratory syncytial virus (HRSV), human parainfluenza virus (HPIV), human adenovirus (HADV), human metapneumovirus (HMPV), and human coronavirus (HCOV).

All analyses were conducted using R software (version 4.1.2; R Foundation for Statistical Computing, Vienna, Austria). The "shapr" package facilitated the implementation of the SHAP process, and the "lavvn" package was utilized to construct the SEM model.

RESULTS

Supplementary Figure S2A (available at https://weekly.chinacdc.cn/) illustrates the seasonal distribution of six viruses across various regions. In major Chinese urban areas, the average absence rate of respiratory syndrome is 32.49%. Notably, the Central and Southern Liaoning and the Beibu Gulf urban agglomerations exhibit the highest absence rates, reaching 69.22%. In contrast, the Beijing-Tianjin-Hebei, Yangtze River Delta, and Pearl River Delta urban agglomerations show the lowest rates, all below 30%. To develop training and validation sets for our model, we randomly sampled various proportions of complete data. The model's accuracy is assessed using RMSE, r, and F-score metrics. Each virus demonstrates a distinct peak during the spring and winter festivals. IFV, HRSV, and HPIV are the predominant viruses among the general population. Geographically, IFV and HPIV are more prevalent in the north, whereas IFV and HRSV are more common in the south, as depicted in Supplementary Figure S2B. This regional variation persists, with HADV being more prevalent in the north. After data interpolation, the results shown in Figure 1 confirm consistent proportions of the three

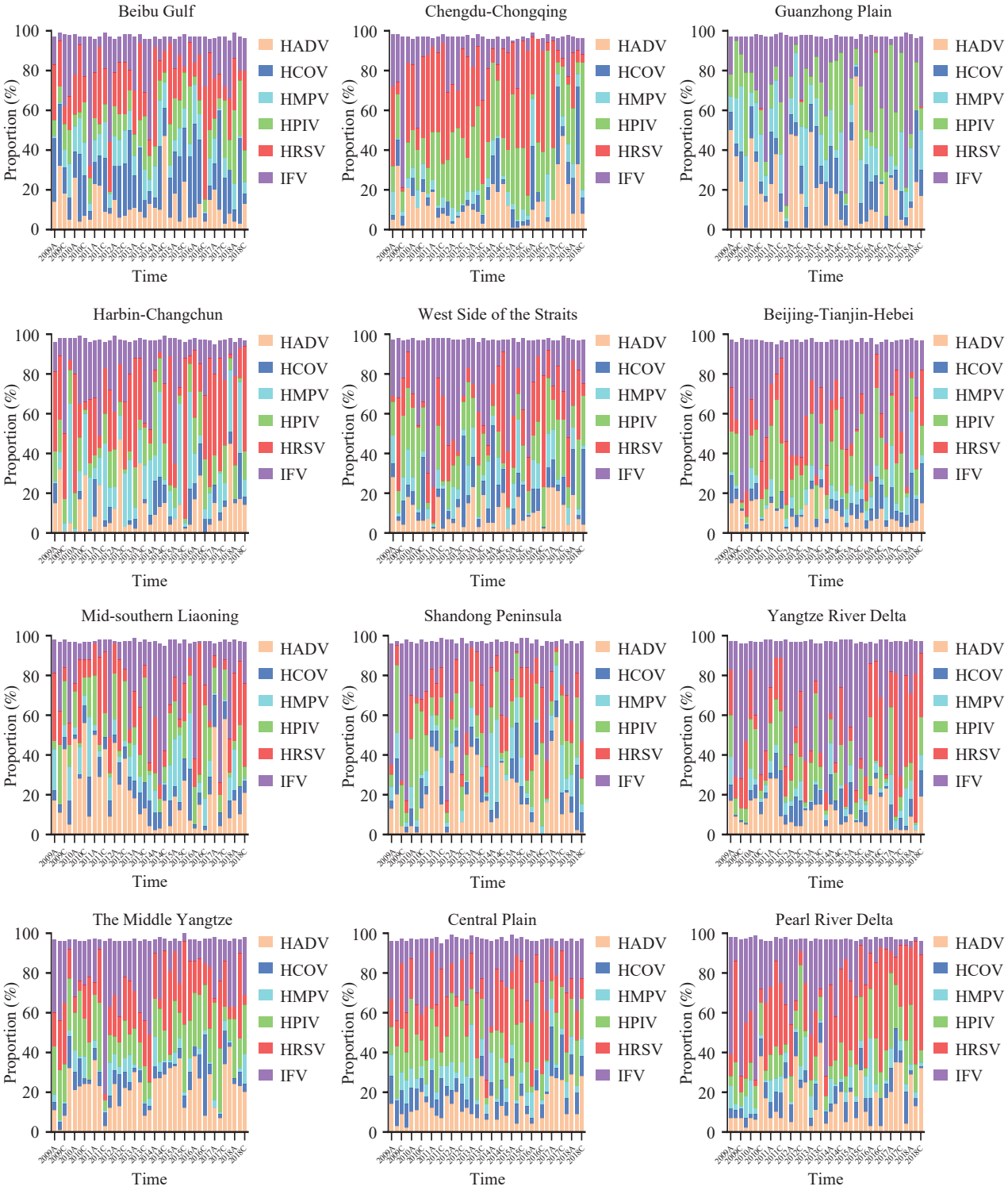


FIGURE 1. The virus structure spectrum of different urban agglomerations after interpolation.
Note: The letters "A" and "C" in the abscissa represent springand autumn quarters, respectively.
Abbreviation: HADV=human adenovirus; HCOV=human coronavirus; HMPV=human metapneumovirus; HPIV=human parainfluenza virus; HRSV=human syncytial virus; IFV=influenza virus.

primary viruses within the population. The north exhibits a higher prevalence of adenovirus as the primary virus, whereas the south displays a higher incidence of co-infections involving the IFV and syncytial virus.

The interpolation analysis delineates two distinct

TABLE 1. The average values of the three evaluation indicators under 100 repeated experiments for 5 models in each setting.

Methods	RMSE	r	F-score
Setting 1: Training=60%			
SHAP-SEM	5.813	0.710***	0.752
SEM	9.424	0.614**	0.651
Cokriging	8.273	0.429	0.634
Bayesian	10.174	0.494**	0.693
Sandwich	7.235	0.539***	0.621
Setting 2: Training=70%			
SHAP-SEM	5.157	0.734***	0.781
SEM	9.364	0.633**	0.684
Cokriging	8.047	0.457	0.634
Bayesian	9.154	0.518**	0.691
Sandwich	7.176	0.584***	0.633
Setting 3: Training=80%			
SHAP-SEM	5.081	0.767***	0.792
SEM	9.331	0.657*	0.708
Cokriging	7.524	0.461**	0.642
Bayesian	8.699	0.523**	0.701
Sandwich	6.926	0.601***	0.651

Abbreviation: RMSE=root-mean-square error; SHAP-SEM=SHapley Additive exPlanation model with the structural equation model; SEM=structural equation model.
* $P<0.05$;
** $P<0.01$;
*** $P<0.001$.

patterns in the dissemination of respiratory syndrome cases across China from 2009 to 2020. Supplementary Figure S2 categorizes Chinese urban agglomerations into three regions: the northern area, which includes the Central Plains; the central region, extending up to the middle reaches of the Yangtze River; and the combined southern and western areas, located beyond the middle reaches of the Yangtze River. In the first grouping, IFV predominated, followed by the parainfluenza virus, with adenoviruses appearing intermittently. The second group also had a predominance of IFV, with the syncytial virus as the second most prevalent, and adenoviruses appearing later in the sequence. In the third group, the syncytial virus was the most dominant, followed by the parainfluenza virus, with adenoviruses emerging in the first six months of the year. These findings have been validated in previous studies (*7*–*9*).

Table 1 presents the mean values of three evaluation metrics across 100 experimental replicates. The data demonstrates superior performance of the SEM integrated with the SHAP algorithm over traditional methods, notably in comparison to the conventional SEM, across all assessed metrics. This indicates enhanced and more consistent interpolation accuracy with the SHAP-SEM approach. Additionally, enlarging the training set size contributes to improved interpolation precision. Compared to the conventional SEM, the SHAP-SEM model excels in spatial distribution accuracy and primary virus classification accuracy. Overall, the SHAP-SEM model achieves robust performance across all indices. Kriging shows the lowest RMSE, but exhibits a decreased r value, whereas the Bayesian approach maintains a relatively high F-score.

Table 2 presents the mean values of evaluation metrics across various simulation settings for different quarters. The findings reveal that accuracy peaks during the winter and dips to its lowest in the autumn. The SHAP-SEM model exhibits superior performance

TABLE 2. Average values of the three evaluation indicators in different seasons under 100 repeated experiments for 5 models in each setting.

Quarters	Method	RMSE	r	F-score
Spring	SHAP-SEM	4.214	0.781***	0.811
	SEM	5.019	0.722**	0.736
	Cokriging	5.105	0.581*	0.662
	Bayesian	6.428	0.614**	0.705
	Sandwich	6.832	0.637**	0.651
Summer	SHAP-SEM	6.194	0.703**	0.710
	SEM	10.521	0.641**	0.682
	Cokriging	7.113	0.519	0.627
	Bayesian	9.245	0.588*	0.690
	Sandwich	8.194	0.572*	0.638
Autumn	SHAP-SEM	7.237	0.651**	0.631
	SEM	9.144	0.601*	0.614
	Cokriging	6.184	0.467	0.596
	Bayesian	7.965	0.501	0.587
	Sandwich	8.229	0.523*	0.577
Winter	SHAP-SEM	4.057	0.722***	0.753
	SEM	6.124	0.671***	0.707
	Cokriging	5.016	0.514*	0.636
	Bayesian	7.255	0.635**	0.641
	Sandwich	5.417	0.642**	0.639

Abbreviation: MSE=mean-square error; RMSE=root-mean-square error; SHAP-SEM=SHapley Additive exPlanation model with the structural equation model; SEM=Structural equation model.
* $P<0.05$;
** $P<0.01$;
*** $P<0.001$.

in the winter, markedly enhancing accuracy relative to competing models. Conversely, in the autumn, the sandwich interpolation method surpasses the SHAP-SEM model in terms of accuracy enhancement.

DISCUSSION

Missing data frequently complicates syndromic surveillance, obstructing the analysis of disease patterns and trends, thereby impeding efforts in disease prevention and control. Developing methods for data interpolation in environments characterized by unstable monitoring and significant data gaps presents a formidable research challenge. Conventional interpolation techniques are often inadequate in contexts involving complex interactions between diseases and their determinants. These methods generally underperform in addressing missing values within sparse datasets.

This study employed interpolation techniques to estimate the prevalence of primary viruses associated with seasonal respiratory syndrome across 13 major urban areas in China between 2010 and 2018, accounting for sparse data and missing values. The accuracy of these estimates was assessed using RMSE, r, and F-score. The results indicate that this method surpasses other approaches in enhancing the accuracy of data on primary respiratory syndrome viruses, achieving significant improvements in overall and seasonal accuracy.

Most spatiotemporal interpolation models for diseases incorporate both spatiotemporal autocorrelation and differentiation. The Co-kriging method (*10*) is a geostatistical technique leveraging correlations between various variables across different sites for spatial interpolation and prediction. However,

employing Co-kriging to estimate missing data introduces increased uncertainty because these estimations depend heavily on the availability of complete datasets and on the spatial correlation among variables. A lack of data in a dataset can undermine spatial autocorrelation, rendering predictions unreliable. Bayesian hierarchical models (*11*) attempt to manage missing data under the assumption of random data loss. If the missing data mechanism is misrepresented within the model, however, parameter estimations might be biased. Conventionally, it is assumed that data loss is random; however, overlooking the specific missing data mechanism can lead to biases in the estimated parameters due to improper data handling. The sandwich approach (*12*) further complicates the issue by treating missing and observed data as independent, disregarding any patterns or correlations in the data removal process. This can result in incorrect standard errors and inferences, particularly when the data deletion mechanisms are informative or directly linked to the variables of interest. Additionally, the complex dynamics of disease occurrence, infection, and transmission intersect variably with different factors. To enhance spatiotemporal interpolation accuracy, advanced techniques like deep learning, including random forest models (*13*) and regressive neural networks (*14*), have been utilized. Despite their effectiveness, these models are often complex and do not sufficiently address the multifaceted nature of disease prevention. This study proposes the integration of SEM with SHAP to discern crucial features and their interrelationships, thus tackling issues related to unstable and fragmented data in health monitoring. By resolving these issues and synthesizing them within our research, we can derive precise insights about syndromes, affected regions, and causal factors. This approach promises to yield scientifically based recommendations for efficacious local prevention and control strategies.

However, this study is subject to some limitations. The SHAP-SEM model necessitates a large sample size, which may not always be feasible with incomplete syndrome surveillance data. Additionally, it is crucial to recognize that viral activity is influenced by a range of risk factors. Future endeavors to integrate incomplete syndrome monitoring data with the SHAP-SEM model should include more factors. It is also important to note that the scalability of SHAP-SEM models may be compromised when handling large datasets. As the size of the dataset expands, the computational and memory demands escalate, potentially leading to extended processing times and heightened complexity.

In further studies, comprehensive descriptions and analyses of the syndrome's spatial and temporal distribution can be achieved through interpolation methods. Additionally, examining variations in viral activity and seasonal trends across different regions is possible. Building on this knowledge, identifying specific risk groups and areas becomes feasible, providing essential data to support targeted, time-sensitive, and location-specific prevention and control strategies.

Conflicts of interest: No conflicts of interest.

Funding: Supported by the Foundation of China (grant number 42171419) and National Science and Technology Major Project of China (grant number 2018ZX10713001).

doi: 10.46234/ccdcw2024.124

[#] Corresponding author: Yilan Liao, liaoyl@lreis.ac.cn.

[1] The State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China; [2] University of Chinese Academy of Science, Beijing, China.

Submitted: May 06, 2024; Accepted: July 01, 2024

REFERENCES

1. Jia P, Yang SJ. China needs a national intelligent syndromic surveillance system. Nat Med 2020;26(7):990. https://doi.org/10.1038/s41591-020-0921-5.
2. Koch T. Disease mapping and innovation: a history from wood-block prints to Web 3. 0. Patterns 2022;3(6):100507. https://doi.org/10.1016/j.patter.2022.100507.
3. Lee KY, Li LX. Functional structural equation model. J Roy Stat Soc Ser B Stat Methodol 2022;84(2):600 – 29. https://doi.org/10.1111/rssb.12471.
4. Liu YC, Liu ZH, Luo X, Zhao HJT. Diagnosis of Parkinson's disease based on SHAP value feature selection. Biocybern Biomed Eng 2022;42(3):856 – 69. https://doi.org/10.1016/j.bbe.2022.06.007.
5. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Res 2005;30(1):79 – 82. https://doi.org/10.3354/cr030079.
6. Zhang WG, He YW, Wang LQ, Liu SL, Meng XY. Landslide susceptibility mapping using random forest and extreme gradient boosting: a case study of Fengjie, Chongqing. Geol J 2023;58(6):2372 – 87. https://doi.org/10.1002/gj.4683.
7. Li ZJ, Zhang HY, Ren LL, Lu QB, Ren X, Zhang CH, et al. Etiological and epidemiological features of acute respiratory infections in China. Nat Commun 2021;12(1):5026. https://doi.org/10.1038/s41467-021-25120-6.
8. Wang JL, Chen T, Deng LL, Han YJ, Wang DY, Wang LP, et al. Epidemiological characteristics of imported respiratory infectious diseases in China, 2014−2018. Infect Dis Poverty 2022;11(1):22. https:

//doi.org/10.1186/s40249-022-00944-6.
9. Zhao YJ, Lu RJ, Shen J, Xie ZD, Liu GS, Tan WJ. Comparison of viral and epidemiological profiles of hospitalized children with severe acute respiratory infection in Beijing and Shanghai, China. BMC Infect Dis 2019;19(1):729. https://doi.org/10.1186/s12879-019-4385-5.
10. Xiao MY, Zhang GH, Breitkopf P, Villon P, Zhang WH. Extended Co-Kriging interpolation method based on multi-fidelity data. Appl Math Comput 2018;323:120 – 31. https://doi.org/10.1016/j.amc.2017.10.055.
11. Miller PC, Ren MD, Schlame M, Toth MJ, Phoon CKL. A bayesian analysis to determine the prevalence of barth syndrome in the pediatric population. J Pediatr 2020;217:139 – 44. https://doi.org/10.1016/j.jpeds.2019.09.074.
12. Wang JF, Haining R, Cao ZD. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. Int J Geogr Inf Sci 2010;24(4):523 – 43. https://doi.org/10.1080/13658810902873512.
13. Mariano C, Mónica B. A random forest-based algorithm for data-intensive spatial interpolation in crop yield mapping. Comput Electron Agric 2021;184:106094. https://doi.org/10.1016/j.compag.2021.106094.
14. Abdelaziz M, Wang TF, Elazab A. Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks. J Biomed Inform 2021;121:103863. https://doi.org/10.1016/j.jbi.2021.103863.

SUPPLEMENTARY MATERIAL

SHAP-SEM Model

The fundamental concept of SHapley Additive exPlanation model (SHAP) is to facilitate both global and local interpretations through the evaluation of the marginal contributions of variable features. This approach assists users in comprehending the relative importance of each feature, thereby enabling effective variable filtering.

The SHAP model utilizes pre-processed data for each variable alongside the target variable. The Shapley values of individual variables are then calculated to assess their relationships with the target variable (Equation 4).

$$\hat{\varphi}_j = \frac{1}{M}\sum_{m=1}^{M}\left(f\left(x_{+j}^{m}\right) - f\left(x_{-j}^{m}\right)\right) \tag{4}$$

Where $\hat{\varphi}_j$ represents the Shapley value of the feature of the variable, j, M represents the total number of instances, f represents the kernel function, x_{+j}^{m} represents the entire instance containing the feature of the variable, and j, and x_{-j}^{m} represents the entire instance without the feature of the variable j.

In this study, we utilized quarterly non-normalized positivity rates along with meteorological variables from various urban agglomerations to construct SHAP models. Employing the edge effect approach, we calculated the significance of various variables on respiratory viruses and their inter-variable interactions. The impact of different combinations of influencing variables on respiratory viruses is presented in Equation 5.

$$\hat{v}(S) = \frac{1}{M}\sum_{k=1}^{M}\left(f\left(x_S^{(i)} \cup x_C^{(k)}\right) - f\left(x^{(k)}\right)\right) \tag{5}$$

$\hat{v}$ represents changes in the respiratory virus in the S variable combination; S is a combination of different variables; x_S and x_C represent all instances that contain S combinations and instances that do not contain S combinations.

Structural equation models (SEMs) elucidate variations within variables and the covariation among them while investigating the relationships between observed and latent variables. SEMs typically comprise two components: the measurement model, which analyzes the relationships between latent variables and their indicators, and the path model, which explores the relationships among the variables themselves.

In this study, the meteorological variables from the SHAP model were used as input variables in the SEM model, with the non-normalized positivity rate of respiratory viruses serving as the observed variables. The framework of the bipartite SEM model was structured accordingly (Equations 6–7).

$$Y_i = \lambda_i \eta + \delta_i \tag{6}$$

$$\eta_i = \sum_{\substack{j=1 \\ j \neq i}}^{J} b_{ij}\eta_j + \epsilon_i \tag{7}$$

Where Y_i is the value of the *ith* observed variable; λ_i represents the factor load between the observed variable and the variable under the value of the *ith* observed variable; η represents the variable that passes the significance test of Shapley value; δ and ϵ represent the error of the measurement model and the path model respectively, and b represents the path coefficient between the variable i and variable j, which means the interaction between the two variables.

Cokriging Method

Cokriging, a geostatistical method initially developed for mining applications, is extensively used in soil science. The primary instrument in geostatistics is the semivariogram, which quantifies the spatial dependence among adjacent observations. The semivariogram, $\gamma(h)$, can be defined as half the variance of the difference in attribute values at all pairs of points separated by h as Equation 8 (*1*).

$$\gamma(h) = \frac{1}{2N(h)}\sum_{i=1}^{N(h)}\left[Z(x_i) - Z(x_i + h)\right]^2 \tag{8}$$

where $Z(x)$ indicates the magnitude of variables, and $N(h)$ is the total number of pairs of attributes that are separated by a distance h.

SUPPLEMENTARY TABLE S1. Geographical distribution of 12 urban agglomerations in China.

UAs	Cities/Municipalities
Central Plain UA	Handan, Xintai, Changzhi, Jincheng, Yuncheng, Bengbu, Fuyang, Suzhou, Bozhou, Liaocheng, Heze, Zhengzhou, Kaifeng, Luoyang, Pingdingshan, Anyang, Hebi, Xinxiang, Jiaozuo, Puyang, Xuchang, Leihe, Sanmenxia, Nanyang, Shangqiu, Xinyang, Zhoukou, Zhumadian
Beijing-Tianjin-Hebei UA	Beijing, Tianjin, Shijiazhuang, Tangshan, Qinhuangdao, Baoding, Zhangjiakou, Chengde, Langfang, Cangzhou
Guanzhong Plain UA	Xi'an, Tongchuan, Baoji, Xianyang, Weinan, Shangluo, Tianshui
Beibu Gulf UA	Zhanjiang, Maoming, Yangjiang, Nanning, Beihai, Fangchenggang, Qinzhou, Yulin, Chongzuo, Haikou
Harbin-Changchun UA	Changchun, Jilin, Siping, Liaoyuan, Songyuan, Yanbian, Harbin, Qiqihaer, Daqin, Mudanjiang, Suihua
Shandong Peninsula UA	Jinan, Qindao, Zibo, Dongying, Yantai, Weifang, Weihai, Rizhao
Chendu-Chongqing UA	Chongqing, Chendu, Zigong, Luzhou, Deyang, Mianyang, Suining, Neijiang, Leshan, Nanchong, Meishan, Yibing, Guangan, Dazhou, Yaan, Ziyang
West Side of Straits UA	Wenzhou, Quzhou, Lishui, Fuzhou, Xiamen, Putian, Sanming, Quanzhou, Zhangzhou, Nanping, Longyan, Ningde, Ganzhou, Shantou, Meizhou, Chaozhou, Jieyang
Pearl River Delta UA	Guangzhou, Shenzhen, Zhuhai, Foshan, Jiangmen, Zhaoqin, Huizhou, Shanwei, Heyuan, Qinyuan, Dongwan, Zhongshan, Yunfu
Mid-southern Liaoning UA	ShenYang, Dalian, Anshan, Fushun, BenXi, Dandong, Yinkou, Liaoyang, Panjin
Yangtze River Delta UA	Shanghai, Nanjing, Wuxi, Changzhou, Suzhou, Nantong, Yancheng, Yangzhou, Zhenjiang, Taizhou, Ningbo, Jiaxin, Huzhou, Shaoxin, Jinhua, Zhoushan, Taizhou, Hefei, Wuhu, Maanshan, Tonglin, Anqin, Chuzhou, Chizhou, Xuancheng
The Middle Yangtze UA	Nanchang, Jingdezhen, Pingxiang, Jiujiang, Xinyu, Yintan, Jian, Yichun, Fuzhou, Shangrao, Wuhan, Huangshi, Yichang, Ezhou, Jinmen, Xiaogan, Jinzhou, HUanggang, Xianning, Changsha, Zhuzhou, Xiangtan, Hengyang, Yueyang, Chnagde, Yiyang, Loudi

Abbreviation: UA=urban agglomeration.

Bayesian Hierarchical Model

The framework of Bayesian hierarchical modeling consists of a structured approach to model building where unobserved quantities are categorized into distinct levels with clear, scientifically interpretable functions and probabilistic connections that reflect the intrinsic characteristics of the data. This methodology has been successfully applied in the analysis of complex epidemiological, biomedical, environmental, and various other types of data (*2*).

Spatial Sandwich Model

The concept of sandwich space interpolation is based on dual-layer stratified statistics. This process constructs an information transfer function across three hierarchical levels: the object layer, the zoning layer, and the reporting layer. Initially, attributes at the object layer are categorized or divided into different zones (zoning layer), followed by stratified sampling within these zones. Specifically, at least two sample points per category are chosen to calculate the mean and variance for each group. Subsequently, these statistical measures are transferred to the reporting layer through overlaying it with the zoning layer. This allows for the compilation of mean and variance for each reporting unit, thereby facilitating the generation of a spatial interpolation map and its corresponding error distribution map (*3*).

Data and Descriptive Analysis

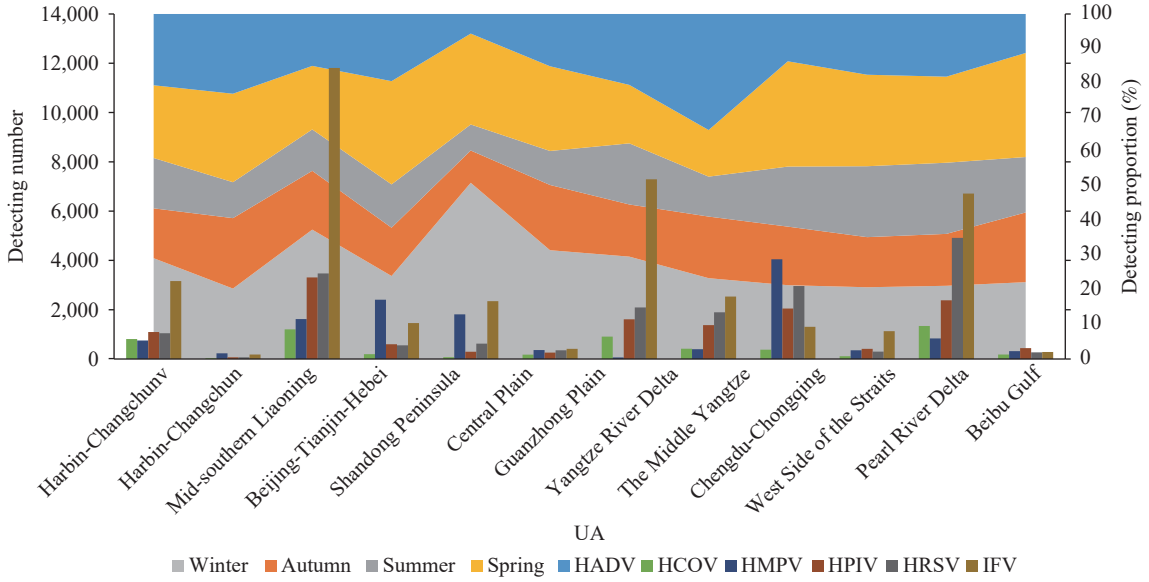
Meteorological data were gathered from 756 ground-based weather stations across China, accessed via the meteorological station sharing service system (https://data.cma.cn/). The selected meteorological variables for this study include atmospheric pressure, average relative humidity, average temperature, average maximum temperature, average minimum temperature, and the range between the highest and lowest temperatures. Daily meteorological data were consolidated into monthly averages. For other study areas, data were estimated using Kriging interpolation, and averages of urban meteorological parameters within urban conglomerates were calculated to derive the aggregate meteorological data for urban agglomerations.

The initial surveillance data indicated a total of 39,760 pneumonia cases, representing 19.63% of the 202,539 samples analyzed. Among the total study population, 76,450 were children under five years of age, constituting 37.75% of the total. Of these, 16,867 cases of pneumonia were children, making up 42.42% of all pneumonia cases. According to Supplementary Table S2, IFV and HRSV were the largest viruses detected in each season, accounting

SUPPLEMENTARY TABLE S2. Number of detections for each virus by season.

Virus	Spring	Summer	Autumn	Winter	Sum
IFV	3,372	3,141	3,397	9,399	19,309
HRSV	1,927	1,020	2,230	4,094	9,271
HPIV	2,002	1,997	1,647	1,301	6,947
HADV	1,509	1,462	1,238	1,589	5,798
HMPV	869	257	269	748	2,143
HCOV	760	850	636	668	2,914
Sum	10,439	8,727	9,417	17,799	46,382

Abbreviation: HADV=human adenovirus; HCOV=human coronavirus; HMPV=human metapneumovirus; HPIV=human parainfluenza virus; HRSV=human syncytial virus; IFV=influenza virus.



SUPPLEMENTARY FIGURE S1. Distribution of raw data and quarterly distribution by UAs.
Abbreviation: UA=urban agglomeration; HADV=human adenovirus; HCOV=human coronavirus; HMPV=human metapneumovirus; HPIV=human parainfluenza virus; HRSV=human syncytial virus; IFV=influenza virus.

for 41.63% and 19.99% of the total samples. IFV in winter accounted for the largest proportion of all viruses detected in all seasons with a percentage of 52.81% and 20.26% of the total annual samples.

Supplementary Figure S1 presents the original monitoring data for each urban agglomeration, segmented by quarter. The Beijing-Tianjin-Hebei, Pearl River Delta, and Yangtze River Delta urban agglomerations lead in monitoring volume. In contrast, the Mid-southern Liaoning and Guanzhong Plain urban agglomerations recorded the lowest sample numbers. Notably, the proportion of HMPV in the Shandong Peninsula, Central Plains, and Chengdu-Chongqing urban agglomerations is significantly higher than in others, while the prevalence of HADV and IFV is notably lower.

Model Setting

Our simulation study investigates six principal configurations by segmenting the quarter and urban agglomerations with monitoring data into separate training and validation sets. These configurations are subsequently applied to SHAP-SEM and other conventional interpolation models.

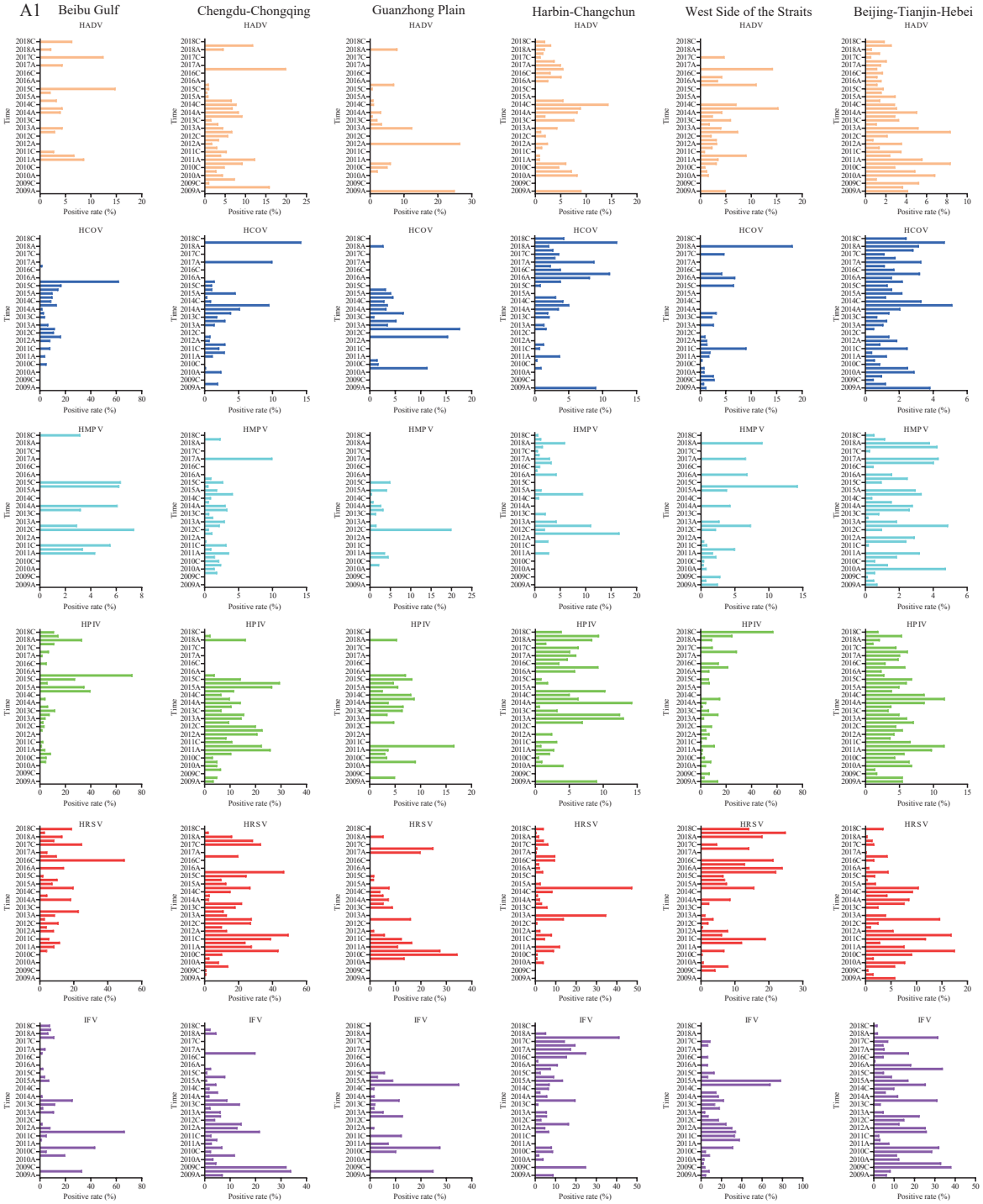
Setting 1: Training=60%, Valid=40%;

Setting 2: Training=70%, Valid=30%;

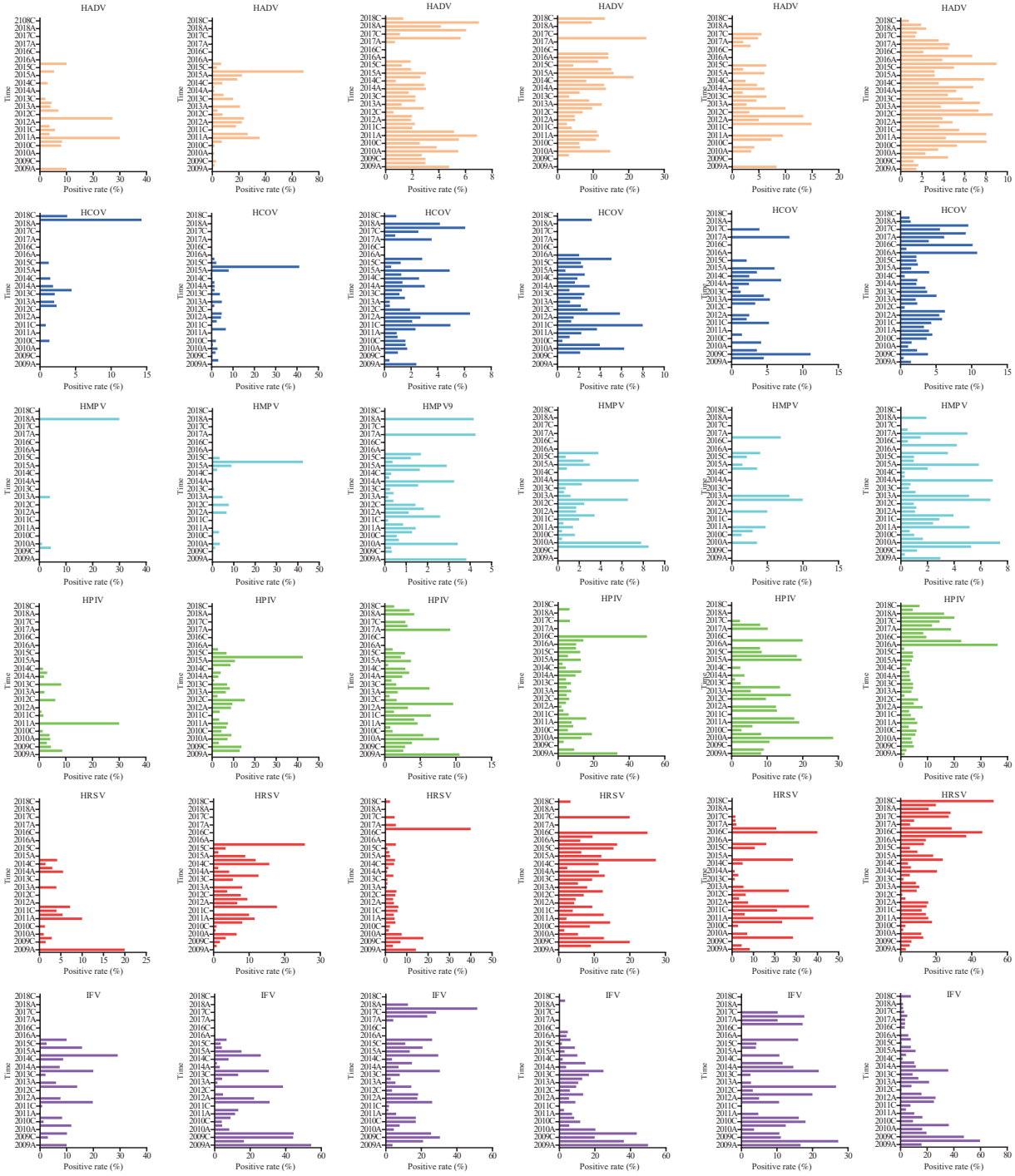
Setting 3: Training=80%, Valid=20%;

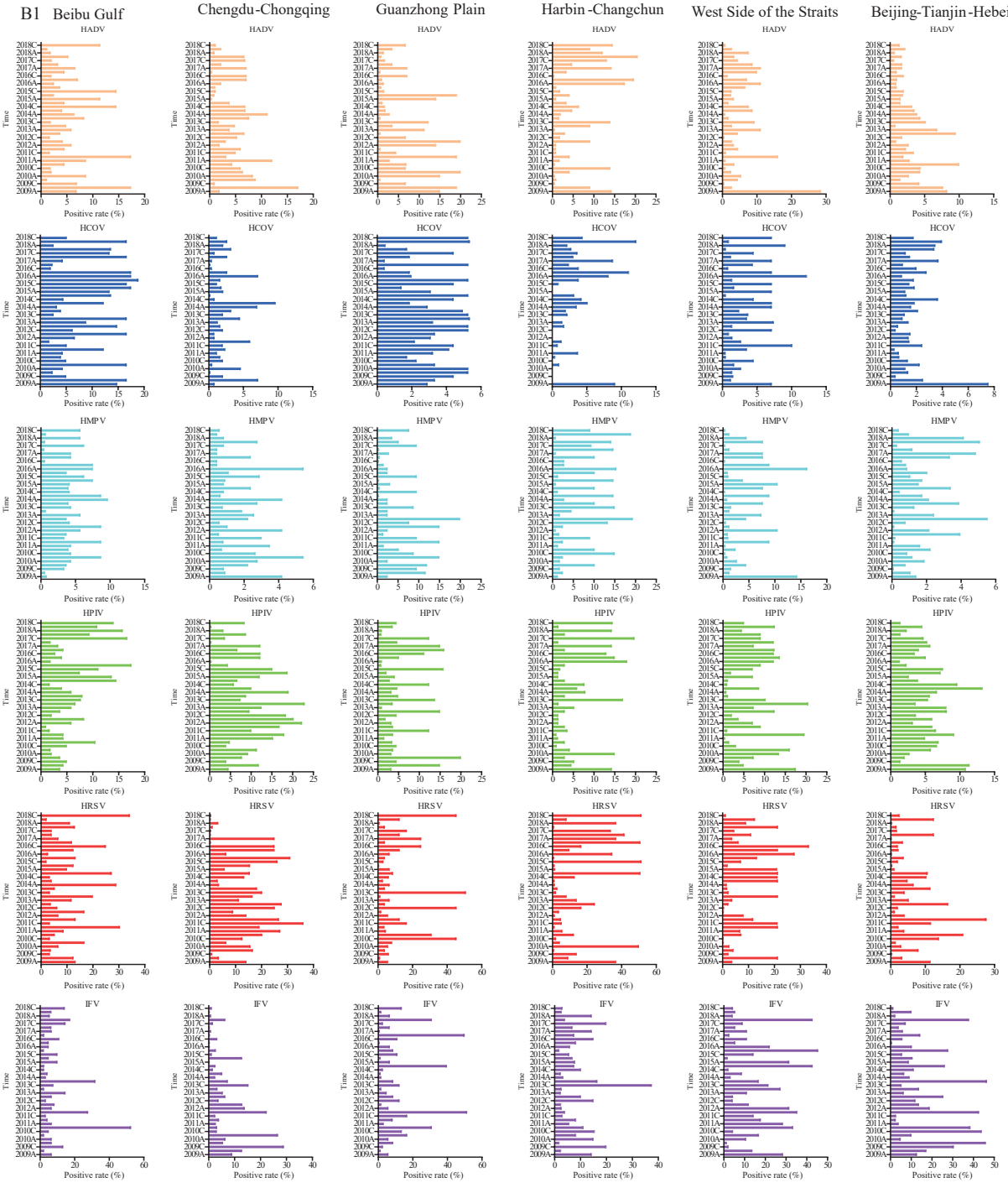
Utilization of Variables

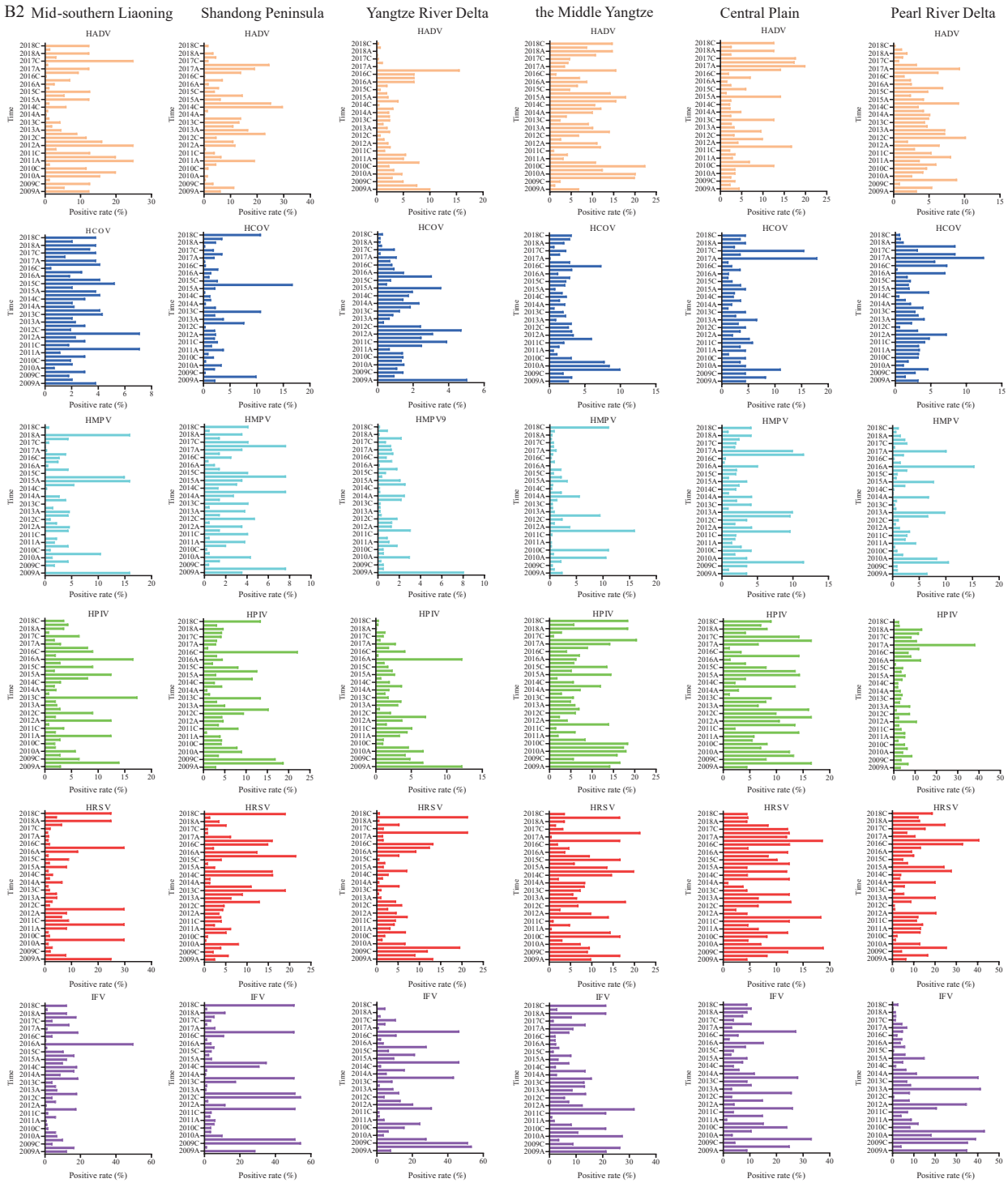
The Kriging interpolation method employs the covariance function of the virus under the assumption of second-order stationarity. Bayesian interpolation is based on Bayesian probability theory. Sandwich interpolation utilizes stratified sampling and considers spatial variations for interpolation. SEM is used to model and interpolate the network diagram of structural relationships, where meteorological elements and the virus are depicted as nodes, and their interconnections as edges. The SHAP-SEM approach quantifies the marginal effects of meteorological variables on the virus and identifies significant contributors for inclusion in further SEM analysis.



A2 Mid-southern Liaoning | Shandong Peninsula | Yangtze River Delta | the Middle Yangtze | Central Plain | Pearl River Delta




B1 Beibu Gulf
Chengdu-Chongqing
Guanzhong Plain
Harbin-Changchun
West Side of the Straits
Beijing-Tianjin-Hebei
HADV
HCOV
HMPV
HPIV
HRSV
IFV
Time
Positive rate (%)



SUPPLEMENTARY FIGURE S2. Virus positive rates in different urban agglomerations. (A) Before interpolation; (B) After interpolation.

Note: A1 and B1 include six UAs: Beibu Gulf, Chengdu-Chonqing, Guanzhong Plain, Harbin-Changchun, West side of the Straits, and Beijing-Tianjing-Hebei. A2 and B2 include other six UAs: Mid-southern Liaoning, Shandong Peninsula, Yangtze River Delta, the Middle Yangtze, Central Plain, and Pear River Delta. The letters “A” and “C” in the ordinate represent spring and autumn quarters, respectively.

Abbreviation: HADV=human adenovirus; HCOV=human coronavirus; HMPV=human metapneumovirus; HPIV=human parainfluenza virus; HRSV=human syncytial virus; IFV=influenza virus.

REFERENCES

1. Ahmadi SH, Sedghamiz A. Application and evaluation of kriging and cokriging methods on groundwater depth mapping. Environ Monit Assess 2008;138(1-3):357 – 68. https://doi.org/10.1007/s10661-007-9803-2.
2. Richardson S, Best N. Bayesian hierarchical models in ecological studies of health–environment effects. Environmetrics 2003;14(2):129 – 47. https://doi.org/10.1002/env.571.
3. Wang JF, Haining R, Cao ZD. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. Int J Geogr Inf Sci 2010;24(4):523 – 43. https://doi.org/10.1080/13658810902873512.