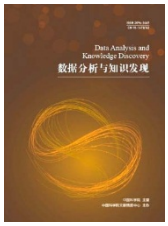

数据分析与知识发现
Data Analysis and Knowledge Discovery
ISSN 2096-3467,CN 10-1478/G2

《数据分析与知识发现》网络首发论文

题目：基于多机器学习方法联合的公共卫生风险预测研究——以兰州市流感预测为例

作者：柴国荣，王斌，沙勇忠

网络首发日期：2020-10-12

引用格式：柴国荣，王斌，沙勇忠．基于多机器学习方法联合的公共卫生风险预测研究——以兰州市流感预测为例．数据分析与知识发现.
https://kns.cnki.net/kcms/detail/10.1478.G2.20201012.1512.006.html

网络首发时间：2020-10-12 16:34:23
网络首发地址：https://kns.cnki.net/kcms/detail/10.1478.G2.20201012.1512.006.html

基于多机器学习方法联合的公共卫生风险预测研究——以兰州市流感预测为例

柴国荣[1,2,3]，王　斌[1,2,3]，沙勇忠[1,2,3]
[1]（兰州大学管理学院 兰州　730000）
[2]（兰州大学医院研究中心 兰州　730000）
[3]（兰州大学应急管理研究中心 兰州　730000）

摘要：[目的] 探索应用机器学习预测流感这类公共卫生风险的可行性和有效性。**[方法]** 首先，收集 2009—2016 年兰州市的流感和气象数据，拆分成 2009—2015 年和 2016 年两组，分别作为训练和验证数据；然后，分别基于 SARIMA、Kalman Filter 和 VAR 建立三种机器学习预测方法，并设计两种多方法联合预测策略；最后，评估、比较上述方法（策略）的预测性能。**[结果]** 在设定的全期（WP）、暴发期（OP）和稳定期（SP）三种场景下，SARIMA、VAR 和 Kalman Filter 方法的预测效果依次最好（*RMSE* 分别为 11.68、19.23、1.60；R^2 分别为 0.932、0.923、0.956）；多方法联合策略可进一步提升三种场景下的预测效果，其中联合策略 Comb_2 的表现更好（*RMSE* 分别为 10.82、14.68、1.38；R^2 分别为 0.942、0.934、0.963）。**[局限]** 相关数据限制，主要考虑了气象一类外部相关因素。**[结论]** 应用机器学习预测流感等公共卫生风险具有可行性和有效性，且潜力巨大。但目前面临的主要困境是多源数据缺乏，需要从技术、组织和制度层面打破数据壁垒，推动数据共享与开放。

关键词：机器学习；流感预测；公共卫生风险；风险预测
分类号：C916; D63　　文献标识码：A
DOI: 10.11925/infotech.2020.0754.

Public Health Risk Forecasting With Multiple Machine Learning Methods Combined：Case Study of Influenza Forecasting in Lanzhou, China

Chai Guorong[1,2,3]，Wang Bin[1,2,3]，Sha Yongzhong[1,2,3]
[1] (School of Management, Lanzhou University, Lanzhou 730000, China)
[2] (Hospital Management Research Center, Lanzhou University, Lanzhou 730000, China)
[3] (Research Center for Emergency Management, Lanzhou University, Lanzhou 730000, China)

Abstract: [Objective] This study tries to explore the practicability and effectiveness of forecasting public health risks with machine learning, taken influenza as an example. **[Methods]** First, we collected the data on influenza and meteorological factors during 2009 to 2016 in Lanzhou, China. Data from the year 2009 to 2015 were used as the training data and 2016 as the testing data. Then, based on SARIMA, Kalman Filter, and VAR, three machine learning methods for influenza prediction were put forward, respectively. Moreover, we designed two multi-method combined forecasting strategies. Finally, the forecasting performance of the above methods (strategies) was carefully evaluated and compared. **[Results]** The SARIMA, VAR, and Kalman Filter achieved best predict performance in the whole period (WP), outbreak period (OP), and

stabilization period (SP), with *RMSE* at 11.68, 19.23, 1.60, and R^2 at 0.932, 0.923, 0.956, respectively. The forecasting performance among all three scenarios was improved by our multi-method combined strategies, in which Comb_2 has better performance, with *RMSE* at 10.82, 14.68, 1.38, and R^2 at 0.942, 0.934, 0.963, respectively. **[Limitations]** Limited by the data, this study just considered meteorology factors as external factors. **[Conclusions]** Predicting public health risks (such as influenza) with machine learning is practicable, effective and has great potential. But a lack of multi-source data is the major dilemma. Therefore, to promote the open exchange and sharing of data, barriers should be broken at the technical, organizational, and institutional levels.

Keywords: machine learning; influenza forecast; public health risk; risk forecast

1 引言

2020 年，新冠疫情全球大流行再次警醒我们，世界已进入风险社会[1,2]，各种复杂社会风险可能随时随地暴发[3]。公共卫生风险作为人类面临的主要社会风险之一，往往不受疆界、阶层、种族、年龄和性别限制，迅速流行蔓延，危害人民的生命健康和财产安全，破坏社会经济系统[4]。以流行性感冒（后文简称“流感”）为例，其具有潜伏期短、传染性强、传播速度快的特点[5]，可在短时间内迅速暴发流行，造成严重的发病和死亡。如 1918 年西班牙大流感，在极短时间内形成全球大流行，累计造成 10 亿人感染，2500 万~1 亿人死亡（当时世界总人口约 17 亿）[6]。即使在医疗卫生条件极大改善的今天，全球每年仍有 29~65 万人死于流感[7]，绝大部分来自发展中国家[8]。另外，流感大流行还会引发经济停滞、民众恐慌，甚至社会失调，如大规模缺勤、缺课，医疗机构不堪重负，谣言四起，民众恐慌、囤积物资，生产活动停滞等[9]。人口众多、人员流动性大、民众风险意识不足等，造成我国社会的风险抗逆力普遍较低，长期饱受流感危害。因此，为了避免流感大暴发、降低损失，在流感防控中要坚持底线思维，做好充分准备，除了研发、生产、储备疫苗外，还应在现有流感监测的基础上积极展开流感预测[4]，提前预防干预，避免风险演变成危机[10]。

目前，世界各国流感防控和预警仍主要依赖于传统的流感监测[11]，政府和公共卫生部门根据流感监测系统（网络）上报的流感信息来评估近期流感活动，作出公共卫生和应急决策[12]。如中美两国均利用自己的流感监测网络，汇总和分析全国各地上报的流感信息（发病数、临床症状、病毒学实验室结果、入院和死亡统计等信息），监测和评估流感动向[13]。传统的流感监测为流感防控提供了大量信息和数据，发挥了重要作用。然而，由于监测数据从基层监测点（哨点医院和实验室）逐级上报到中央决策部门需要较长时间（约 1~2 周），监测数据往往滞后于流感实时状态[14]。这导致评估者无法及时、准确洞察流感发展态势，决策者无法果断做出最佳决策，错过防控“最佳窗口期”。面对这一不足，世界卫生组织（WHO）指出增强流感预测能力对流感防控意义重大[4]，学界和实践界就流感预测积极展开了探索。最初的流感预测模型是基于经典的“易感—感染—恢复”（Susceptible - Infected – Recovered，SIR）流行病学模型建立的，然而由于模型的关键参数是“区间人员流动”，不仅需要频繁调整，而且当时条件下很难实现实时获取，极大限制了它的应用。[15]然而，随着大数据和 5G 等信息技术的快速发展，其应用潜力巨大。随着互联网的快速普及，学者们开始探索基于互联网数据预测流感的方法，最具代表性的就是谷歌流感趋势（Google Flu Trends）。

2008 年，谷歌流感趋势基于互联网搜索数据预测了 2003-2015 年的流感，然而后来发现预测结果不准、预测方法存在缺陷[16]。在当今互联网环境下（如有目的的网络话题制造和引导），其缺陷可能会被进一步放大。还有学者尝试基于相关分析方法，通过分析外部变量（如病毒类型、病假、相关药物销售）[17]与流感（发病数或发病率）的相关关系来预测流感。然而，现实中这些外部变量数据很难实现实时获取。也有学者同样基于相关分析方法，利用气象因素（温度和湿度等）[18]预测流感，相较于上述外部变量，气象因素数据更容易做到实时监测和获取。总之，以往的预测方法或因假设和技术存在限制性，或因关键因素获取困难缺乏灵活性，都制约了实际效用。最近一些流感预测研究使用了差分整合移动平均自回归 ARIMA （Autoregressive Integrated Moving Average）[19,20]、卡尔曼滤波器（Kalman filter）[21]和机器学习[22]等方法，均取得了较好的预测效果，但仍有提升空间。另外，流感的高度敏感性，很难用单一固定模型或方法完全描述[21]。

基于上述背景，本研究收集2009-2016年的兰州市流感和部分气象因素数据，基于季节性差分自回归滑动平均 SARIMA（Seasonal Autoregressive Integrated Moving Average）、卡尔曼滤波器 KF（Kalman Filter）和向量自回归 VAR（Vector Autoregression）建立机器学习预测方法，检验三种机器学习方法独立和联合预测流感活动的可行性和有效性，并比较预测性能。本研究将为我国的流感预测和防控实践提供一种新思路，为我国公共卫生风险治理提供借鉴，助力我国国家治理现代化。

2 研究设计

2.1 数据来源

流感一般具有2到4天，最长7天的潜伏期，且从出现症状到就诊也存在1到2天的时差。因此，通常设置一周（7天）为一个流感预测期，本研究也遵循了该惯例，研究所涉及的流感和气象因素数据均为周数据。其中，流感数据为发病数（例），来源于兰州市疾病预防控制中心，时间范围为2009年01月01日至2016年12月31日；同期兰州市主要气象因素数据来源于甘肃省气象局网站（http://gs.weather.com.cn/），包括：温度（℃）、气压（hPa）、相对湿度（%）、风速（m/s）以及降雨量（mm）。最后，将上述数据按周汇总成共418周的周数据。

2.2 数据分析

数据分析包括：探索性数据分析、单方法独立预测、多方法联合预测，及预测结果的验证与比较四个阶段。

（1）探索性数据分析

描述主要研究变量的平均值、标准差、范围及四分位数，并绘制时间序列图，以了解数据的分布、周期和趋势特征。计算滞后一期气象因素与当期流感发病数之间的 Spearman 相关系数，以分析流感发病与气象因素间的原始关联关系，为后续预测分析提供支持。

（2）单方法独立预测

① 预测流程设计

为了提升预测性能，本研究设计了批处理学习与实时学习结合的有监督机器学习预测流程，如图 1 所示。该流程可以有效避免机器预测性能随时间推移而逐渐退化的情况。

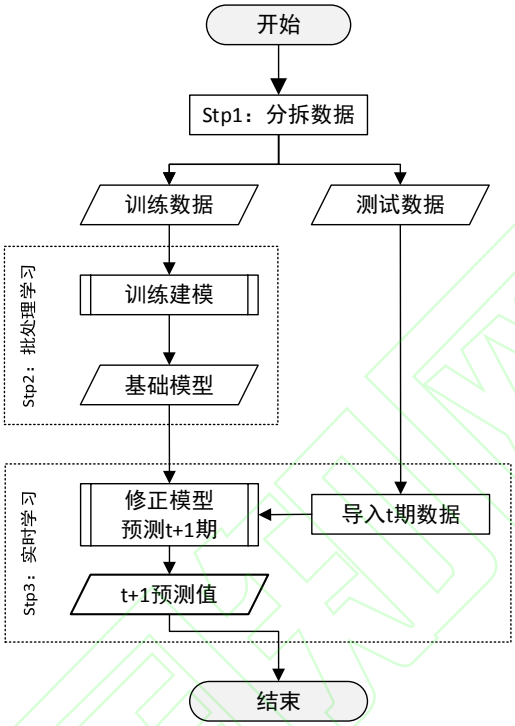

图 1　预测流程

Fig.1 Forecasting process

Step 1：将数据分拆成训练和测试数据两部分。其中，0～366 周（2009-2015 年）的数据作为训练数据，367～418 周（2016 年）数据作为测试数据。

Step 2：批处理学习。利用不同预测方法使用训练数据建立最优基础模型、估计关键参数；

Step 3：实时机器学习与预测。逐次导入 $t(t \in [367,418])$ 周测试数据，机器根据最优准则完成一次迭代学习，生成一个即时最优模型，并以此模型预测 t+1 周的流感发病数，即“实时学习，实时预测”。

② 预测方法与建模

结合探索性数据分析获得的数据特征，分别基于 SARIMA、KF 和 VAR 三种方法建立预测模型。

1) SARIMA

SARIMA 是在差分整合滑动平均自回归模型 ARIMA（Autoregressive Integrated Moving Average）的基础上考虑季节性（周期性）影响发展而来[23]，广泛应用于流行病预测，尤其是针对存在明显季节性（周期性）的流感、登革热、疟疾和手足口病等流行病[24]。一般表示为 $\mathrm{SARIMA}(p,d,q)\times(P,D,Q)s$，其中 p，d，q 分别为自回归、差分、移动平均的阶数；P，D，Q 分别为季节性自回归、差分、移动平均的阶数，s 为周期[25]。本研究首先以时间序列数据流感周发病数建立最优 SARIMA 模型，然后将气象因素作为外部变量引入模型，建立 SARIMAX 流

感预测模型，如式(1)：

$$\hat{y}_{t+1} = \frac{q(B)Q(B^s)}{p(B)P(B^s)\nabla^d \nabla_s^D} + X_t + \varepsilon \quad (1)$$

其中，$\hat{y}_{t+1}$为流感预测值，$q(B)$和$Q(B^s)$分别为移动平均和季节性移动平均多项式，$p(B^s)$和$P(B^s)$分别为自回归和季节性自回归多项式∇^d和∇_s^D分别为非季节性和季节性差分项，X_t为 t 期气象因素，ε 为残差。首先，批处理学习阶段，设置周期$s = 1, 2, \cdots, 104$，以 AIC（Akaike information criterion）最小为准则，利用 R 扩展包 forecast 中 auto.arima()函数建立不同周期的最优模型，再从这些模型中选择一个最优的作为基础模型，确定最佳周期s_o。实时学习阶段，逐期导入测试数据，以 AIC 最小为准则，由 auto.arima()函数自主选择即时最优模型，以此模型预测后一期流感发病数。

2) KF

KF（Kalman Filter）是一种高效率的递归滤波器，可以根据不同时期变量的观测值，考虑各时期的联合分布，产生对未来的预测值，其适用于平稳和非平稳态的预测[26]。KF 建立于隐马尔可夫模型（Hidden Markov Model，HMM）之上，其动态系统可以用一个马尔可夫链表示，该马尔可夫链建立在一个被高斯噪声干扰的线性算子上。在时间推移过程中，变化的线性算子作用于当前状态，从而催生一个新的状态，并带入新的控制信息和噪声。KF 模型假设 t+1 时刻的状态是从 t 时刻的状态演化而来，表示为式(2) [27]：

$$\hat{y}_{t+1} = \Phi_t y_t + \Pi_t u_t + w \quad (2)$$

其中，$\hat{y}_{t+1}$和y_t分别为 t+1 和 t 时刻的真实状态，Φ_t为状态变换模型（线性算子），Π_t为控制器，u_t为控制信息，w 为噪声。

本研究将 t 流感发病数作为 t 期状态，t 期气象因素作为控制信息u_t，建立模型预测 t+1 期状态——流感发病数，根据 KF 原理和算法[27]编写程序。现有研究表明 KF 模型的预测性能非常依赖迭代次数[28]。为此，首先在批处理学习阶段利用训练数据对其进行大规模训练，然后在实时学习阶段逐期导入测试数据，实时学习、修正模型，并预测后一期流感发病数。

3)VAR

VAR 是一种非结构化的多方程建模方法，在经济和金融领域被广泛使用，用于研究时间序列系统的预测和随机扰动对变量系统的动态影响。VAR 的核心思想是直接考虑研究变量的时间和时序关联关系。因此，其不但可以避免结构化建模中的内生性问题，而且估计研究变量动态变化时可以涵盖更多信息，获得更好的预测效果。[29] 本研究建立的 VAR 流感预测模型如式(3)：

$$\hat{y}_{t+1} = \alpha + \sum_{i=0}^{l} \beta_{t-i} y_{t-i} + \lambda_{t-i} X_{t-i} + \varepsilon \quad (3)$$

其中，$\hat{y}_{t+1}$为 t+1 期的流感预测值，y_{t-i}为不同滞后期流感的观测值，X_{t-i}为不同滞后期的气象因素，$l\ (l \in [0, t])$为模型最优时的滞后期，α 和ε 分别为截距和残差。首先，批处理学习阶段设定最大滞后期为 10 期，以 AIC 最小为准则，利用 R 扩展包 vars 中 VARselect()函数选择最优模型，确定模型最优时的滞后期参数lag_o；在实时学习阶段，设定模型滞后期$l = lag_o$，逐期导入测试数据，以 AIC 最小为准则，由 vars 包提供的 VAR()函数自主生成即时最优模型，以此模型预测后一期流感发病数。

（3）多方法联合预测

物理学中为了提高测量（实验）的结果准确性和有效性，通常的做法是多次测量（实验），然后取平均值。本研究借鉴该思想，通过分别计算上述三种方法预测值的算术平均数和几何平均数，设计了两种多方法联合预测的策略 Comb_1（取算术平均数）和 Comb_2（取几何平均数）。

（4）预测结果验证

为了全面评估预测效果，首先，绘制预测值与观察值的曲线图；然后，根据流感周期变化趋势设置全期（WP）、暴发期（OP）和稳定期（SP）三个场景，评估和比较上述三种独立方法和两种联合策略的预测效果。其中，全期是指整个测试期；暴发期是流感发病数变化剧烈的时段；稳定期是流感发病数平稳，变化较小的时段。另外，由于本研究是一个数值型预测问题，且观测值和预测值都存在等于 0 的情况。因此，参考已有研究 [30,31]，在众多指标中选择了被广泛使用且不受 0 影响的均方根误差 *RMSE*（Root Mean-Square Error）和决定系数 R^2（Coefficient of Determination）两个指标来评估预测效果。其中，*RMSE* 计算方法如式(4)，用于度量预测值与观测值的平均偏离程度，值越小表示平均偏离程度越小，即预测越准，预测效果越好；R^2（$R^2 \in [0,1]$）用于度量预测值与观测值的接近程度，即拟合优度，值越接近于 1，表示预测值越接近观测值，预测效果越好。

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=t}^{n}(y_t - y_t)^2} \qquad (4)$$

其中，n 表示预测样本的大小，y_t 和 $\hat{y}_t$ 分别表示 t 期的观测值和预测值。

以上预测建模过程主要基于 R（Version 3.4.4）实现，主要使用了 tidyverse、forecast、vars、DMwR、MASS、sandwich、strucchange 和 ggplot2 等功能扩展包。

3 研究结果

3.1 探索性数据分析

2009—2016 年，兰州市共报告流感 3956 例。描述性统计分析显示（表 1）：平均每周流感发病数约为 9 例（范围：0~241 例）；同期气象因素：温度、大气压、风速、相对湿度和降雨量的周平均值分别为：11.23 ℃（范围：-8.81~29.00 ℃）、846.78 hPa（范围：836.70~858.00 hPa）、1.24 m/s（范围：0.69~2.00 m/s）、50.46%（范围：18.06%~78.43%）和 0.82 mm（范围：0~17.89mm）。如图 2，兰州市流感和气象因素的变化均呈现明显的周期性，周期约为 1 年，冬、春两季是流感高发期。2016 年春季出现了一次暴发高峰，峰值为 631，明显高于其他年份同期。

表 1　流感与气象因素的统计性描述（2009—2016 年）

Table1 Descriptive statistics of influenza and meteorological factors, 2009–2016

	均值±标准差	最小值	最大值	百分位数		
				25%	50%	75%
流感发病数	9±20	0	241	2	5	10
温度	11.23±9.79	-8.81	29.00	2.18	12.93	19.92

大气压	846.78±4.48	836.70	858.00	843.51	846.87	849.70
风速	1.24±0.23	0.69	2.00	1.07	1.23	1.41
相对湿度	50.46±12.24	18.06	78.43	42.66	51.37	59.29
降雨量	0.82±1.67	0	17.89	0	0.14	1.00

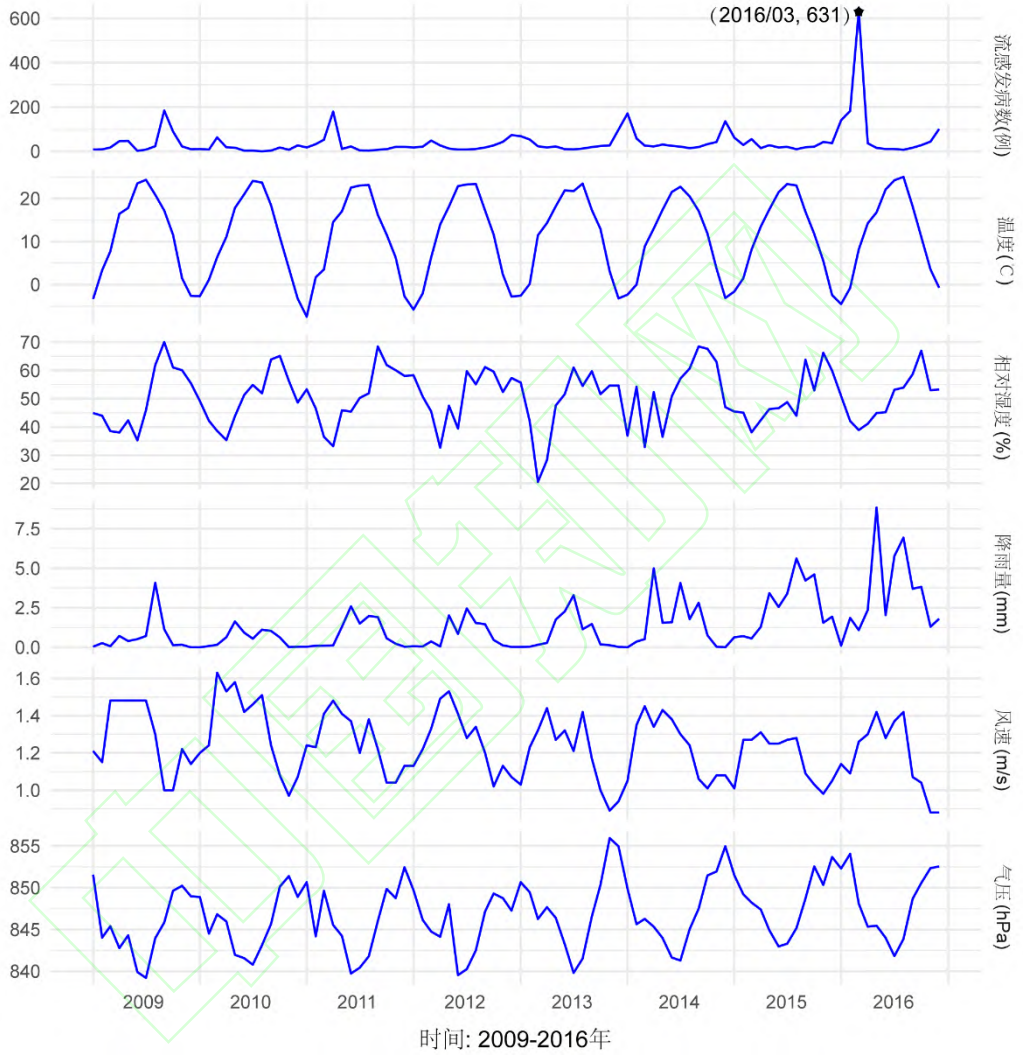

图 2　流感和气象因素的时间序列图（2009—2016）

Fig.2 Time Series chart of influenza and meteorological factors, 2009—2016

表 2 为滞后一期气象因素与当期流感发病数之间的 Spearman 相关系数，结果显示滞后一期气象因素与当期流感发病数的相关系数均具有统计学意义（$p < 0.05$），表明流感发病与气象因素之间存在关联关系。

表 2　滞后一期气象因素与流感发病数的 Spearman 相关系数（2009—2016 年）

Table2 Spearman correlations between weekly meteorological variables at lag of 1 week and influenza cases, 2009—2016

	温度	大气压	风速	相对湿度	降雨量
大气压	-0.741**				

风速	0.425**	-0.533**			
相对湿度	-0.032	0.166**	-0.439**		
降雨量	0.499**	-0.359**	0.135**	0.493**	
流感发病数	-0.482**	0.400**	-0.266**	-0.074*	-0.272**

*$p < 0.05$，**$p < 0.01$

3.2 预测结果及评估

图 3 分别展示了 367～418 周（2016 年）流感发病数的观测值和三种方法独立及联合预测的预测值。总体而言五组预测值与观测值都比较接近，变化趋势基本一致。因此，直观而言五种方法（策略）均可较准确地预测流感。

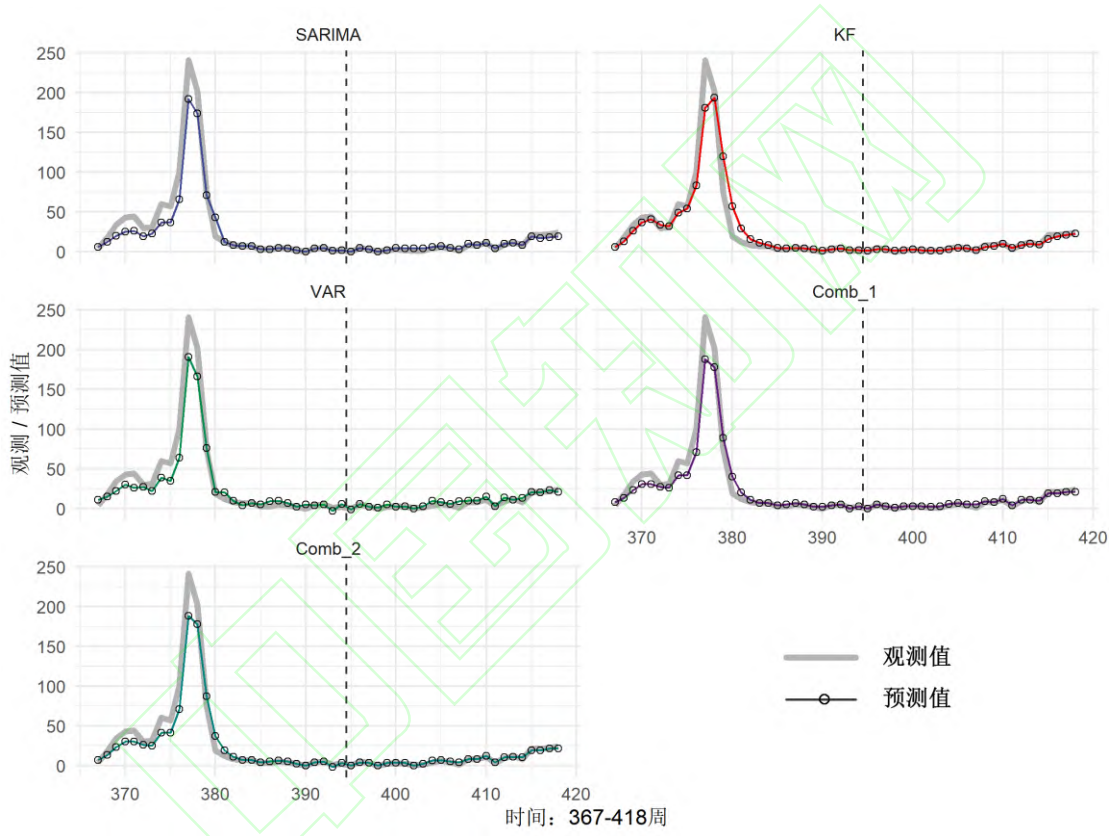

图 3　流感发病数的观测值和预测值（367～418 周）

Fig.3 Observed values and predicted values of influenza incidence, 367～418 weeks

设定第 367～418 周，即整个测试期为全周期场景（WP）。再根据流感的变化趋势，设定发病数波动剧烈的第 367-394 周为暴发期场景（OP），而发病数变化相对稳定的第 395～418 周为稳定期场景（SP）。三种场景下独立方法和联合预测策略的预测结果的 *RMSE* 和 R^2 如表 3 所示，比较可以发现：

第一，比较 SARIMA、KF 和 VAR 三种方法独立预测的效果，发现它们在不同场景下的预测表现不一致，各有所长。WP 场景下 SARIMA 的 *RMSE* 最小且 R^2 最大，预测效果最好，预测值与观测值平均偏差约为±11.68，预测值对观测值的解释度为 93.2%；在 OP 场景下，VAR 的 *RMSE* 最小且 R^2 最大，预测效果最好，平均偏差约为±19.23，解释度为 92.3%；SP 场景下 KF 的 *RMSE* 最小且 R^2 最大，预测效果最好，平均偏差约为±1.60，解释度为 95.6%。

第二，五种预测方法（策略）中，多方法联合的 Comb_2 在三种场景下的表现均为最好，Comb_1 次之，表明联合预测策略可以进一步提升预测性能。在 OP 场景下的性能提升尤为明显，两种联合预测策略的平均偏差约为±14.7，解释度均超过 93%，相较于三种独立预测方法，平均预测偏差下降了约 24%。

第三，纵向比较五种预测方法（策略）在三种场景下的表现，几乎都在 SP 场景下表现最好，WP 下次之，在 OP 下最差。

表3　各场景下独立方法和联合策略预测结果的*RMSE*和R^2

Table3 *RMSE* and R^2 for prediction results of independent methods and combined strategies in each scenario

		SARIMA	KF	VAR	联合预测	
					Comb_1	Comb_2
RMSE	WP	**11.68**	12.61	11.85	10.88	**10.82***
	OP	20.28	21.96	**19.23**	14.74	**14.68***
	SP	1.72	**1.60**	3.13	1.58	**1.38***
R^2	WP	**0.932**	0.921	0.930	0.941	**0.942**#
	OP	0.920	0.910	**0.923**	0.933	**0.934**#
	SP	0.918	**0.956**	0.832	0.953	**0.963**#

* 每行最小的 *RMSE*；# 每行最大的 R^2.

4 研究发现与启示

研究发现：第一，基于机器学习方法可以实现比较满意的流感预测；第二，与单一机器学习方法独立预测相比，多种机器学习方法联合预测可以有效提升预测能力，获得更好的预测结果。具体而言：分别基于 SARIMA、KF 和 VAR 建立的三种机器学习方法均可较好地预测流感发病数和活动趋势，而且面对不同场景表现各有所长。上述三种方法的联合可以进一步提升预测效果，尤其在暴发场景下尤为明显。造成上述结果的主要原因是：三种方法的基本原理不相同。其中，SARIMA 主要基于周期变化，KF 主要基于临近值的波动，VAR 则基于多因素的时序关联关系，这导致不同方法在不同场景下的表现各有千秋，而多方法联合正好可以通过取长补短来有效规避该问题，进而提升预测能力，得到更准确的预测结果。另外，研究结果也表明有监督的批处理学习和实时学习的结合，可以长期保持预测性能，避免随时间推移预测性能衰退，预测结果偏差越来越大的情况。

基于上述发现，本研究认为面对流感这类“灰犀牛”式公共卫生风险，可以在传统监测的基础上，考虑通过多方法联合策略建立和提升实时预测能力，从而构建常态与应急结合的风险治理体系。而这种多方法联合策略不仅需要多种先进分析方法和技术的支撑，更需要多源数据的实时共享与融合。从当前现实情况来看，实现这一构想的最大困难不是缺乏工具——方法和技术，而是缺乏原料——多源数据。目前多源数据缺乏的根本原因不是数据收集困难或不存在，而是因为缺乏数据共享与开放。因此，需要从技术、组织和制度全面入手，打破数据壁垒、联通数据孤岛。在技术层面，充分利用监测传感、5G、大数据、人工智能等新兴信息技术，实现多源异构数据的实时获取、处理和分析；在组织层面，打破条块分割形成的部门壁垒，建立由政府主导，卫生、医院、气象、环境、交通和个人等多元主体参与的数据共享与开放网络；在制度层面，建立、完善和落实信息公开、数据共享和数据开放的政策和标准。

5 结语

不断暴发的各种流行病疫情都在不断提醒我们，风险社会并非危言耸听，而是真真切切地威胁着每个人。面对流感这类公共卫生风险，仅依靠传统监测手段很难做到及时、准确地研判，尤其是面对短时间内快速传播、大规模感染的突发情况。而这种突发情况恰恰又是其主要破坏力之所在，一旦失控将引发重大公共危机。幸运的是监测传感、大数据、机器学习和 5G 等新兴技术的发展，不仅推动了生活、工作和组织等方式的深刻变革，也为流感等的预测提供了新思路和新方法。本研究以预测兰州市流感发病数为例，探索了应用机器学习预测流感活动的可行性和有效性。研究表明机器学习方法可以较好地预测流感，同时与单一方法相比，多方法联合可以进一步提升预测能力。目前流感预测面临的最大困难是多源数据缺乏，需要全社会共同努力，从技术、组织和制度层面推动数据共享与开放。本研究尚存在不足之处，如仅利用了气象一种外部因素，但对于深入探索利用机器学习预测公共卫生风险提供了一定借鉴。

参考文献：

[1] Beck U. World Risk Society[M]. Nanjing: Nanjing University Press, 2004.

[2] 范如国. “全球风险社会”治理:复杂性范式与中国参与[J]. 中国社会科学, 2017, (2): 65-83+206. (Fan Ruguo. Governance of “Global Risk Society”: The Paradigm of Complexity and Chinese Participation [J]. Social Sciences in China, 2017, (2): 65-83+206.)

[3] Giddens A. The consequences of modernity[M]. Redwood City: Stanford University Press, 1990: 325-327.

[4] WHO. Up to 650 000 people die of respiratory diseases linked to seasonal flu each year[EB/OL]. (2017-12-13).[2020-07-01]. http://www.who.int/mediacentre/news/statements/2017/flu/en/.

[5] 李兰娟, 任红. 传染病学[M]. 北京: 人民卫生出版社, 2013. (Li Lanjuan, Ren Hong. Infectious diseases [M]. Beijing: People’s medical publishing house Co., Ltd, 2013.)

[6] Barry J M. The Great Influenza: The Story of the Deadliest Pandemic in History[M]. London: Penguin Books, 2005.

[7] WHO. Influenza (Seasonal)[EB/OL]. (2018-10-06).[2020-04-06]. http://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal).

[8] Nair H, Brooks W A, Katz M, et al. Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis[J]. Lancet, 2011, 378(9807): 1917-1930.

[9] Thompson W W, Eric W, Praveen D, et al. Estimates of US influenza‐associated deaths made using four different methods[J]. Influenza Other Respir Viruses, 2009, 3(1): 37-49.

[10] 张海波. 社会风险研究的范式[J]. 南京大学学报(哲学·人文科学·社会科学), 2007, 44(2): 136-144. (Zhang Haibo. Paradigms for Societal Risk Studies[J]. Journal of Nanjing University (Philosophy, Humanities and Social Sciences), 2007, 42(2): 136-144.)

[11] 中国国家流感中心. 流感监测[EB/OL]. (2019-12-25).[2020-08-09]. http://www.chinaivdc.cn/cnic/lgwd/ptlg/201912/t20191225_209368.htm. (Chinese National Influenza Center. Influenza surveillance [EB/OL]. (2019-12-25).[2020-08-09]. http://www.chinaivdc.cn/cnic/lgwd/ptlg/201912/t20191225_209368.htm.)

[12] Lu F, Hou S, Baltrusaitis K, et al. Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis[J]. JMIR Public Health and Surveillance, 2018, 4: e4.

[13] Centers for Disease Control and Prevention of US. U.S. Influenza Surveillance System: Purpose and Methods[EB/OL]. (2020-07-08).[2020-08-18]. https://www.cdc.gov/flu/weekly/overview.htm.

[14] Biggerstaff M, Johansson M, Alper D, et al. Results from the second year of a collaborative effort to forecast

influenza seasons in the United States[J]. Epidemics, 2018, 24: 26-33.
[15] Yang W, Karspeck A, Shaman J. Comparison of Filtering Methods for the Modeling and Retrospective Forecasting of Influenza Epidemics[J]. PLOS Computational Biology, 2014, 10(4): e1003583.
[16] Olson D R, Konty K J, Paladini M, et al. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales[J]. PLOS Computational Biology, 2013, 9(10): e1003256.
[17] Kim M-J, Nembhard H, Lambert B, et al. A syndromic surveillance system for clinical and non-clinical health data[J]. IIE Transactions on Healthcare Systems Engineering, 2011, 1: 37-48.
[18] Soebiyanto R P, Adimi F, Kiang R K. Modeling and Predicting Seasonal Influenza Transmission in Warm Regions Using Climatological Parameters[J]. PLOS ONE, 2010, 5(3): e9450.
[19] 李若曦, 王晓岗, 陈黎黎,等. ARIMA 模型在北京市丰台区流行性感冒预测中的应用[J]. 职业与健康, 2018, 34(6): 792-795+799. (Li Ruoxi, Wang Xiaogang，Chen Lili, et al. Application of ARIMA model in forecasting incidence of influenza in Fengtai District of Beijing [J]. Occup and Health, 2018, 34(6): 792-795+799.)
[20] 周美兰, 周志华, 罗美玲,等. 湖南省哨点医院流感样病例 SARIMA 模型预测[J]. 实用预防医学, 2018, 25(3): 370-373. (Zhou Meilan, Zhou Zhihua, Luo Meiling, et al. Prediction of influenza-like illness in sentinel hospitals in Hunan Province by SARIMA model[J]. Practical Preventive Medicine, 2018, 25(3): 370-373.)
[21] Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza[J]. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(50): 20425-20430.
[22] Ben-Nun M, Riley P, Turtle J, et al. National and Regional Influenza-Like-Illness Forecasts for the USA[J]. bioRxiv, 2018.
[23] Box G E, Jenkis G M. Time series analysis for casting and control[M]. San Francisco: Holden-day, 1970.
[24] Liu S, Chen J, Wang J, et al. Predicting the outbreak of hand, foot, and mouth disease in Nanjing, China: a time-series model based on weather variability[J]. International Journal of Biometeorology, 2017, 62.
[25] Du Z, Xu L, Zhang W, et al. Predicting the hand, foot, and mouth disease incidence using search engine query data and climate variables: an ecological study in Guangdong, China[J]. Bmj Open, 2017, 7(10): e016263.
[26] Kalman R E. A New Approach to Linear Filtering and Prediction Problems[J]. Journal of Basic Engineering Transactions, 1960, 82: 35-45.
[27] Bishop G, Welch G. An introduction to the kalman filter[J]. Proc of SIGGRAPH, Course, 2001, 8(27599-23175): 41.
[28] Pei S, Kandula S, Yang W, et al. Forecasting the spatial transmission of influenza in the United States[J]. Proceedings of the National Academy of Sciences, 2018, 115(11): 2752-2757.
[29] Monogan J. Vector Autoregression[M]. Mauritius: Betascript Publishing, 2010: 678-699.
[30] Guo P, Liu T, Zhang Q, et al. Developing a dengue forecast model using machine learning: A case study in China[J]. Plos Neglected Tropical Diseases, 2017, 11(10): e0005973.
[31] 陈东, 王建冬, 李慧颖,等. 融合机器学习算法和多因素的禽肉交易量预测方法研究[J]. 数据分析与知识发现, 2020, 4(7): 18-27. (Chen Dong, Wang Jiandong, Li Huiying,et al. Forecasting Poultry Turnovers with Machine Learning and Multiple Factors [J]. Data Analysis and Knowledge Discovery, 2020, 4(7): 18-27.)

(通讯作者：沙勇忠，E-mail: shayzh@lzu.edu.cn)

基金项目：本文系国家自然科学基金项目“空间受限大型复杂项目的安全与进度集成管理研究”(项目编号:71472079)；国家中央高校基本科研业务费重点项目“创新驱动的我国

建筑业转型发展研究”(项目编号: 18LZUJBWZD07)；教育部哲学社会科学重大课题攻关项目“大数据驱动的城市公共安全风险研究”(项目编号:16JZD023) 的研究成果之一。

作者贡献声明：

柴国荣、沙勇忠：提出研究思路，设计研究方案，论文最终版本修订；
王　斌：清洗和分析数据，论文起草，论文最终版本修订。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储，E-mail: Chaigr@lzu.edu.cn。
[1] 柴国荣, 王斌, 沙勇忠. 兰州市流感和气象数据_2009-2016.CSV.
[2] 柴国荣, 王斌, 沙勇忠. 兰州市流感预测数据_2016.CSV.