·预防医学·

应用乘积季节模型与指数平滑模型预测上海市肺结核疫情

卞子龙[1,2,3],卓莹莹[2],贺志强[2],张 枫[1],蔡奇慧[1],吴 静[1*]

[1]南京医科大学生物医学工程与信息学院,[2]公共卫生学院,江苏 南京 211166;[3]浙江大学公共卫生学院,浙江 杭州 310058

[摘 要] **目的**:探讨利用两种时间序列模型预测上海肺结核发病趋势的可行性,为制定上海地区肺结核预防控制策略提供科学依据。**方法**:收集上海市2007年1月—2018年6月传染病历史疫情资料,对2007年1月—2017年12月肺结核月发病人数分别采用自回归移动平均模型(ARIMA)乘积季节模型与指数平滑模型进行拟合,预测2018年1—6月的肺结核月发病人数,并与真实值进行比较。**结果**:ARIMA$(0,1,1)\times(1,1,2)_{12}$为乘积季节模型的最优模型,均方根误差(RMSE)为76.27,2018年1—6月预测值的相对误差和为0.402;运用指数平滑法构建的最优模型是Holt-Winters加法指数平滑,均方根误差(RMSE)为69.61,2018年1—6月预测值的相对误差和为0.292。**结论**:指数平滑模型拟合效果较好,预测精度更高,对上海市肺结核疫情的防控具有一定的指导意义。

[关键词] ARIMA模型;指数平滑模型;肺结核;预测

[中图分类号] R183.3 [文献标志码] A [文章编号] 1007-4368(2021)02-268-06

doi:10.7655/NYDXBNS20210223

Application of multiple seasonal model and exponential smoothing model in predicitng pulmonary tuberculosis epidemic in Shanghai

BIAN Zilong[1,2,3],ZHUO Yingying[2],HE Zhiqiang[2],ZHANG Feng[1],CAI Qihui[1],WU Jing[1*]

[1]School of Biomedical Engineering and Informatics,[2]School of Public Health,Nanjing Medical University,Nanjing 211166;[3]School of Public Health,Zhejiang University,Hangzhou 310058,China

[**Abstract**] **Objective**:To explore the feasibility of two time series models for predicting tuberculosis epidemic in Shanghai,so as to provide a scientific reference for the prevention and control of tuberculosis in Shanghai. **Methods**:The epidemiological data of tuberculosis from Jan.2007 to Jun.2018 in Shanghai was collected. The monthly cases of tuberculosis from Jan.2007 to Dec.2017 was fitted by multiple seasonal ARIMA model and the exponential smoothing model. We predicted the monthly number of tuberculosis cases from Jan. to Jun. 2018 using the established models and compared the results to the real values. **Results**:ARIMA$(0,1,1)\times(1,1,2)_{12}$ was the optimal ARIMA model whose RMSE was 76.27 and the sum of the relative error from Jan. to Jun.2018 was 0.402;The optimal model constructed by exponential smoothing was Holt-Winters additive exponential smoothing model with RMSE of 69.61 and a sum of relative error of 0.292 from Jan. to Jun.2018. **Conclusion**:The exponential smoothing model could be fitted more effectively and had higher predictive accuracy. In short,it has important guiding significance for the prevention and control of tuberculosis in Shanghai.

[**Key words**] ARIMA model;exponential smoothing model;pulmonary tuberculosis;prediction

[J Nanjing Med Univ,2021,41(02):268-273]

肺结核是由结核杆菌侵入人体肺部引起的一种慢性呼吸道传染病,其传染性强,易反复发作,可导致癌变恶化的发生,被列入全球十大致死传染病之一。据世界卫生组织估计,2017年全球结核病人数约为17亿,约有1 000万新结核病患者[1]。中国是全球22个结核病高负担国家之一,患者数量占全球的9%,居全世界第2位[2]。中国国家卫生委员会发

[**基金项目**] 国家自然科学基金青年项目(61901225);江苏省高校自然科学基金(19KJD330001);南京医科大学科技发展基金面上项目(2017NJMU005);南京医科大学教育研究课题(2019LX070)

*通信作者(Corresponding author),E-mail:wujing@njmu.edu.cn

布的2017年全国法定传染病疫情概况显示，全国(除港澳台地区)全年共报告肺结核发病835 193例，在乙类传染病中发病数和死亡数均居第2位，已经成为我国重点关注的公共卫生问题。上海市疾病预防控制中心公布的上海市传染病疫情报告显示，2017年全市居民新登记肺结核3 624例，发病率24.9/10万，较2016年下降3.0%；外来流动人口新登记肺结核2 821例，发病数较2016年下降2.6%。虽然上海肺结核疫情已得到了有效的控制，但由于耐多药肺结核的流行、人口老龄化加速以及外来人口流动性增加等问题，上海的结核病防治工作又面临着新的严峻考验，上海地区的结核病防控依然不容轻视[3-4]。

时间序列是指将相同统计指标的数值按其发生的时间先后顺序排列而成的数列，对其分析的主要目的是根据对已有历史数据规律的挖掘从而实现对未来的预测，故近年来被越来越广泛地应用在传染病的发病预测中。本文采用两种时间序列模型——自回归移动平均模型(autoregressive integrated moving average model，ARIMA)乘积季节模型与指数平滑模型对上海市2007年1月—2017年12月肺结核月发病人数进行拟合分析，预测2018年1—6月的肺结核月发病人数，并与实际值进行比较，探讨这两种模型在上海市肺结核疫情预测中的效果，确定预测肺结核发病趋势的最优模型，为上海市肺结核防控工作提供科学依据。

1 资料和方法

1.1 资料

数据资料来源于上海市疾病预防控制中心网站(网址：http://www.scdc.sh.cn/)2007年1月—2018年6月上海市法定报告传染病疫情资料，其中2007年1月—2017年12月的肺结核发病数据用于建立模型，2018年1—6月的数据用于验证模型的预测效果，从而确定最优模型。

1.2 方法

1.2.1 ARIMA乘积季节模型

ARIMA是由美国统计学家Box和英国统计学家Jenkins提出的著名时间序列预测模型之一，又称Box-Jenkins模型。本研究应用同时带有季节性与趋势性的ARIMA乘积季节模型ARIMA$(p,d,q)\times(P,D,Q)_s$，其中参数p、d、q分别为非季节自回归阶数、非季节差分阶数、非季节移动平均阶数，P、D、Q分别为季节自回归阶数、季节差分阶数、季节移动平均阶数，s为季节周期[5-6]。

ARIMA模型的基本思想是，将预测值随时间迁移而形成的数据序列视为一个随机序列，用相对应的数学模型来描述该序列中的自相关性。当模型被识别后，就可从该时间序列的过去值及现在值来预测未来值。建立ARIMA时间序列模型可归纳为3个主要步骤：①数据的预处理(序列的平稳化)：观察时序图、自相关分析图判断平稳性，通过相应差分进行序列的平稳化，进行白噪声检验；②模型的识别、定阶与模型参数估计：采用Box-Jenkins方法建立ARIMA时间序列分析模型，也就是立足于考察数据的样本自相关、偏相关函数判断相应的阶数，季节长度s可由实际应用背景的分析得到；③模型的诊断检验及预测：典型方法是对观测值和模型拟合值的残差进行白噪声分析，同时可以结合赤池信息准则(Akaike information criterion，AIC)、Schwarz贝叶斯准则(Schwarz Bayesian criterion，SBC)，选取较优模型进行预测[7]。

1.2.2 指数平滑模型

指数平滑法是布朗(Robert G.Brown)提出的一种在移动平均法的基础上发展而来的时间序列分析预测方法，通过计算指数平滑系数，配合以时间序列预测模型对未来的现象做出预测。事实上，大多数随机事件，一般都是近期的数据会对现在的影响大，远期的数据会对现在的影响小。指数平滑法的基本思想就是考虑时间间隔对时间发展的影响，并且各期权重随时间间隔的增大呈指数衰减。指数平滑法的预测步骤为：①绘制序列图；②根据序列图确定有效参数；③绘制拟合曲线图，并观察拟合效果；④建立指数平滑模型，对数据进行预测。

根据序列是否具有长期趋势与季节效应，可以把序列分为3大类，采用3种不同的指数平滑模型进行序列预测，具体模型选择见表1[8]。

表1 指数平滑预测模型的使用场合

Table 1 The usage scenarios of exponential smoothing model

预测模型选择	长期趋势	季节效应
简单指数平滑	无	无
Holt两参数指数平滑	有	无
Holt-Winters三参数指数平滑	无/有	有

指数平滑模型含有常规参数、趋势参数和季节参数等3个重要参数，在通常情况下，应综合运用整体均值、整体趋势以及季节性进行预测，通过不同参数值的组合进行拟合。在选择较优模型时，通过比较均方根误差(root mean square error，RMSE)、平

均绝对误差百分比(mean absolute percent error, MAPE)、平均绝对误差(nean absolute error, MAE)的数值,综合选取最优模型,并对模型的预测效果进行评价。

1.3 统计学方法

应用SAS 9.4建立ARIMA乘积季节模型;应用R 3.5.0建立Holt-Winters三参数指数平滑模型。

2 结 果

2.1 数据的初步分析

考虑到上海市经济发达,人口流动性非常高,因此分析本市居民与外来人口和整体发病数之间的关系。为了直观地比较,利用软件画出本市居民与外来人口随时间变化的堆积面积图(图1)。由图可知,本市居民与外来人口发病数占比随时间变化相对比较平稳,呈季节性趋势,也就是整体与部分的病例数的趋势相对一致,因此可将上海市的病例数进行整体分析。

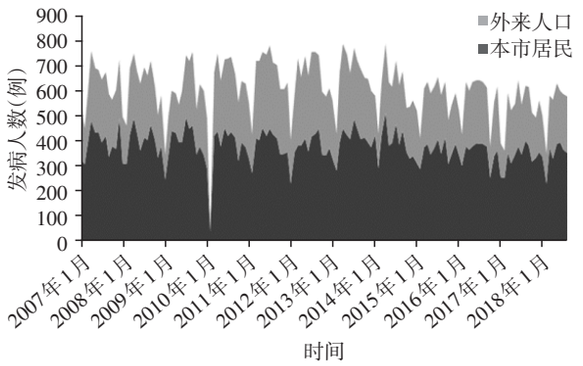

图1 上海市2007—2017年本市居民与外来人口肺结核的发病数

Figure 1 The number of pulmonary tuberculosis cases among residents and migrants in Shanghai from 2007 to 2017

2.2 ARIMA模型

2.2.1 数据预处理

绘制2007年1月—2017年12月肺结核发病数的时序图。从图2可以观察到,肺结核发病数随时间变化总体上呈下降的长期趋势,并且序列取值以12个月为周期呈现出有规则的上下波动。具体地,肺结核发病数从每年的1—2月开始上升,在该年的3—4月先达到1个高峰,继而波动式下降,在11—12月份左右略有上升后再持续下降到次年1—2月份,跌落谷底。

由原始序列图可知序列不平稳,存在周期性。

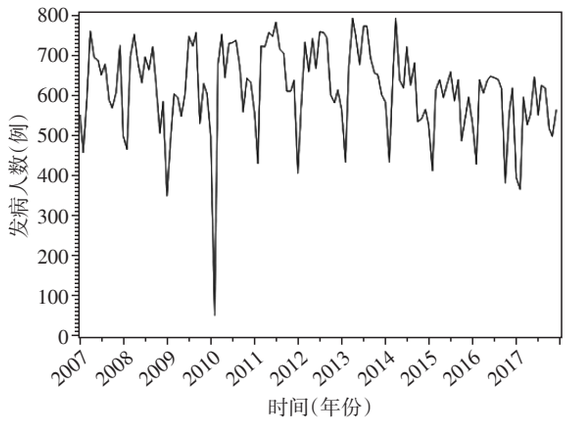

图2 2007—2017年上海市肺结核月发病数时间序列图

Figure 2 Time series of monthly reported number of pulmonary tuberculosis cases in Shanghai from 2007 to 2017

对该序列进行白噪声检验,其自相关检查的P值均<0.05,具有统计学意义,判定上海市肺结核月发病数的时间序列属于非白噪声序列。再对序列作1阶12步差分,提取其趋势效应和季节效应后,时序图基本平稳(图3)。

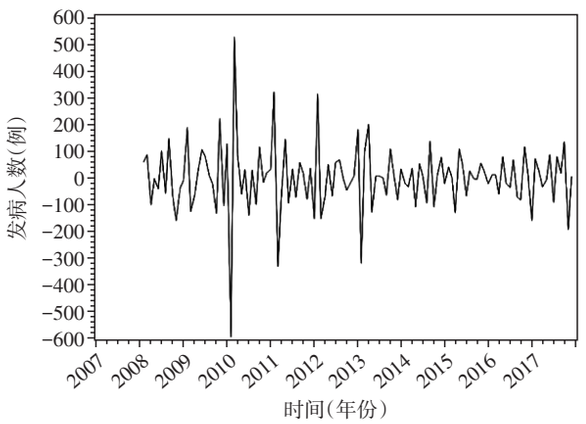

图3 2007—2017年上海市肺结核发病数差分图

Figure 3 Difference of the number of pulmonary tuberculosis cases in Shanghai from 2007 to 2017

2.2.2 模型识别与定阶

序列具有连续相关性和季节性,说明适合ARIMA乘积季节模型ARIMA$(p,d,q)\times(P,D,Q)_s$。经1阶12步差分处理后,序列的长期趋势和季节周期性被很好地消除,故判断d=1,D=1,s=12。根据差分后序列的自相关函数(ACF)图和偏自相关函数(PACF)图(图4),ACF图显示延迟1阶自相关系数显著非零,PACF图显示延迟1、2阶偏自相关系数均大于2倍标准差,故q可能取0、1,p可能取0、1、2。此外,考虑序列的季节自相关特征,差分后的ACF

图显示延迟12阶自相关系数显著非零，而且延迟24阶自相关系数也并未完全落入2倍标准差范围，由此判断Q可能取值为0、1、2。差分后的PACF图显示延迟12阶和延迟24阶的偏自相关系数均显著非零，故P可能取值为0、1、2。

接着对所有可能合理的模型进行拟合，选出符合建模标准的几种组合。经过几种模型的比较(表2)，选取AIC为1 375.27，SBC为1 383.61均最小的相对最优模型ARIMA(0，1，1)×(1，1，2)$_{12}$，参数估计均具有统计学意义。

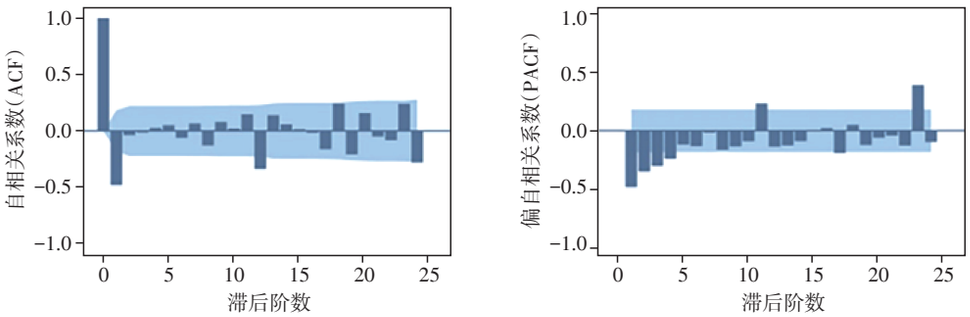

图4　序列的自相关函数(ACF)图与偏自相关函数(PACF)图

Figure 4　Autocorrelation function(ACF)and partial autocorrelation function(PACF)of series

表2　ARIMA(p，d，q)×(P，D，Q)$_s$各种模型比较

Table 2　Comparison of various models of ARIMA(p，d，q)×(P，D，Q)$_s$

模型	方差	标准误	AIC	SBC
ARIMA(0，1，1)×(0，1，1)$_{12}$	6 236.34	78.97	1 379.53	1 385.09
ARIMA(0，1，1)×(1，1，2)$_{12}$	5 967.94	77.25	1 375.27	1 383.61
ARIMA(0，1，1)×(2，1，1)$_{12}$	6 037.19	77.70	1 376.65	1 384.98
ARIMA(1，1，2)×(2，1，1)$_{12}$	6 072.58	77.93	1 378.31	1 389.43

表3　残差的自相关检查

Table 3　Autocorrelation check of residuals

滞后阶数	χ^2值	自由度	P值
6阶	1.20	3	0.753 2
12阶	4.99	9	0.835 3
18阶	12.63	15	0.631 2
24阶	20.27	21	0.504 0

2.2.3　模型的诊断检验及预测

残差的自相关检验表明序列均为白噪声，均有$P>0.05$，表示应该接受残差不相关的零假设，认为残差序列为纯随机序列，所建立的ARIMA(0，1，1)×(1，1，2)$_{12}$模型是合适的。此外，从残差自相关图中也能看出残差序列值都落入置信区间内，符合要求。

2.3　指数平滑模型

根据上海市肺结核月报告发病例数序列并结合指数平滑法的使用方法，选取Holt-Winters加法指数平滑法模型、Holt-Winters乘法指数平滑法模型进行尝试，对2007年1月—2017年12月数据进行拟合。利用R 3.5.0自动拟合出两种模型的最优参数，选择最优参数模型进行比较(表4)。结果显示，Holt-Winters加法指数平滑法建立模型RMSE、MAPE均比Holt-Winters乘法指数平滑法模型的评价参数小，分别是71.99、14.19%。这表明Holt-Winters加法指数平滑法模型比较好。

2.4　两种较优模型的对比

采用MAE、RMSE、MAPE评价两种较优模型的拟合效果(表5)。同时，根据两种模型对上海市2018年1—6月肺结核发病数预测值及其相对误差，比较两者预测效果(表6)。

由表5可知Holt-Winters加法指数平滑法模型的MAE、MRE、RMSE均比较小，在本例中该模型要优于ARIMA(0，1，1)×(1，1，2)$_{12}$模型。同时，表6具体展示了两种模型的预测值与相对误差，也表明Holt-Winters加法指数平滑法模型的总的相对误差要比ARIMA(0，1，1)×(1，1，2)$_{12}$模型的小。

最后，绘制这两种模型的拟合曲线与真实值进行更为直观的比较(图5)，发现Holt-Winters加法指

表4　最优参数设置和效果评价

Table 4　Optimal parameter setting and effect evaluation

方法	α	β	γ	RMSE	MAPE(%)
Holt-Winters加法指数平滑法	0.110	0.022	0.237	71.99	14.19
Holt-Winters乘法指数平滑法	0.096	0.022	0.220	72.82	14.24

表5 两种模型的拟合效果比较

Table 5 Comparison of the fitting effects of the two models

指标	ARIMA $(0,1,1)\times(1,1,2)_{12}$	Holt-Winters加法指数平滑法
RMSE	76.27	69.61
MAE	57.62	50.17
MAPE(%)	16.00	14.58

数平滑模型整体与真实曲线相对拟合效果较好。

3 讨 论

肺结核作为一种成因复杂的慢性传染病,其发病率高、病死率高,严重威胁人类的健康。随着人们生活水平的提高与医疗技术的进步,结核病在一定程度上得到了控制,但与此同时,耐多药肺结核的流行、人口老龄化加速,又给肺结核的防治工作带来了新的挑战。此外,流动人口对肺结核防治知识掌握率较低,结核病就诊率低,这也是肺结核防治工作中的难题[9]。上海市地理位置独特,经济发达,人口流动性大,上海地区结核病的防治工作尤其需要重视。

表6 两种模型对预测值的误差比较

Table 6 Comparison of relative errors on predicted values between two models

时间	实际值	ARIMA $(0,1,1)\times(1,1,2)_{12}$		Holt-Winters加法指数平滑法	
		预测值	相对误差	预测值	相对误差
2018年1月	503	408.45	0.190	436.29	0.130
2018年2月	351	343.36	0.020	357.41	0.020
2018年3月	583	556.58	0.050	582.61	0.001
2018年4月	566	567.09	0.002	603.86	0.070
2018年5月	630	555.54	0.120	587.72	0.070
2018年6月	603	593.81	0.020	603.48	0.001
相对误差和	—	—	0.402	—	0.292

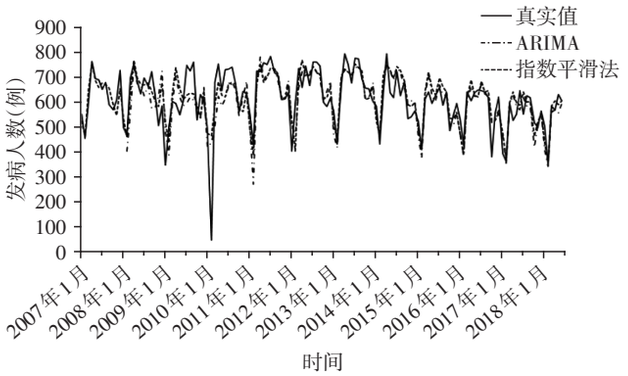

图5 实际值与两种模型拟合值序列比较图

Figure 5 Actual value series and fitted value series of the two models

目前对传染病疫情进行预测的模型有动态因子模型、BP人工神经网络模型、马尔可夫链(Markov模型)、广义回归神经网络(GRNN)模型、ARIMA模型等[10-14]。本文所采用的ARIMA模型和指数平滑模型都属于时间序列模型,其基本原理都是通过揭示历史数据随时间的变化规律,确定已有时间序列的变化模式,并将这种规律外延,以此来预测未来现象[12]。ARIMA模型综合考虑了序列的趋势、周期的变化及随机干扰等情况,并用模型参数进行了量化,能较好地反映时间序列的趋势和变化,并能判断季节效应、趋势效应和随机波动等因素。但ARIMA模型依赖于大量的历史数据,建模过程较复杂,模型的参数常不容易确定,且要求序列在相当一段时间内保持相对的平稳才能实现预测的精确有效,故在拟合模型之前,通常需要对非平稳序列进行差分等转换。指数平滑模型是基于对移动平均预测方法的改进,其综合运用临近的值、整体趋势和季节性来进行预测分析,按均方误差最小的原则来确定平滑系数,对近期的值给予更大的权重,因此近期数据对结果影响较大,而远期数据则影响较小,适用于分析随时间变化不大的数据。相对于ARIMA模型,指数平滑模型的建模过程更简单一些[15]。对于上海市肺结核发病的历史数据,本文的研究结果显示,$ARIMA(0,1,1)\times(1,1,2)_{12}$模型RMSE为76.27,2018年1—6月预测值的相对误差和为0.402;运用指数平滑法构建的最优模型是Holt-Winters加法指数平滑,其RMSE为69.61,2018年1—6月预测值的相对误差和为0.292,效果优于ARIMA乘积季节模型。由此推断,Holt-Winters加法指数平滑模型更适合用于上海市肺结核疫情的短期预测。

ARIMA乘积季节模型与指数平滑模型均以历史数据为基础，建模前提是数据的外延，若外界影响因素突然变化，或是有新变量引入，都会对模型预测效果造成大的影响，降低预测效能。因此这两个模型更加适用于时间序列的短期预测，对序列的更进一步预测，需要及时更新数据资料，添加新的实际值对模型进行修正，然后重新拟合预测。另外，影响肺结核发病的因素繁多，欲研究其他因素对肺结核发病序列的影响，可以考虑结合其他模型建立多因素模型共同分析。

[参考文献]

[1] KYU H H，MADDISON E R，HENRY N J，et al. The global burden of tuberculosis：results from the Global Burden of Disease Study 2015[J]. Lancet Infect Dis，2018，18(3)：261-284

[2] World Health Organization. Global tuberculosis report 2018[EB/OL]. [2019-08-12]. https://www.who.int/tb/publications/global_report/en/

[3] 王 华，包训迪，刘 双，等. 线性探针技术快速检测肺结核耐药性临床研究[J]. 临床肺科杂志，2016，21(5)：856-857

[4] 杨天池，洪 航，陈 同，等. 人口流入城市肺结核流行特征、时空分布及其社会影响因素分析[J]. 中国人兽共患病学报，2017，33(9)：800-804

[5] 王 超，丁 勇，陆 群，等. ARIMA乘积季节模型在我国甲肝发病预测中的应用[J]. 南京医科大学学报(自然科学版)，2014，34(1)：75-79

[6] 张文娟，刘文东，胡建利，等. 基于ARIMA模型的江苏省梅毒疫情预测[J]. 南京医科大学学报(自然科学版)，2017，37(5)：649-652

[7] 卞子龙，汤佳琪，倪春辉，等. 应用ARIMA-GRNN组合模型分析江苏尘肺病发病情况[J]. 环境与职业医学，2019，36(8)：755-760

[8] 王 燕. 应用时间序列分析[M]. 北京：中国人民大学出版社，2016：196-201

[9] 单富强. 流动人口结核病流行特征与管理现状[J]. 中国卫生产业，2018，15(20)：182-183

[10] 朱奕奕，赵 琦，冯 玮，等. 应用指数平滑法预测上海市甲型病毒性肝炎发病趋势[J]. 中国卫生统计，2013，30(1)：31-33

[11] 潘姣姣，董柏青，吕 炜，等. 三种时间序列模型探讨1989~2012广西肺结核发病趋势[J]. 中国卫生统计，2012，29(6)：868-870

[12] 严 婧，杨北方. 指数平滑法与ARIMA模型在湖北省丙型病毒性肝炎发病预测中的应用[J]. 中国疫苗和免疫，2017，23(3)：292-297

[13] 杨丽娟，段 禹，张燕杰，等. 动态因子模型在安徽省乙类传染病发病情况分析中的应用[J]. 中国卫生统计，2017，34(6)：853-856

[14] 张生奎，王镇德，杨 荔，等. 基于SARIMA-ERNN组合模型预测我国细菌性痢疾发病率[J]. 南京医科大学学报(自然科学版)，2019，39(6)：925-931

[15] 汪 鹏，彭 颖，杨小兵. ARIMA模型与Holt-Winters指数平滑模型在武汉市流感样病例预测中的应用[J]. 现代预防医学，2018，45(3)：385-389

[收稿日期] 2020-03-27

(上接第257页)

is volume superior to area or diameter?[J]. J Am Coll Cardiol，2006，47(5)：1018-1023

[14] MA Y，HOU Y，MA Q，et al. Compressed SENSE single-breath-hold and free-breathing cine imaging for accelerated clinical evaluation of the left ventricle[J]. Clin Radiol，2019，74(4)：325 e329-325 e317

[15] SOHRABI S，HOPE M，SALONER D，et al. Left atrial transverse diameter on computed tomography angiography can accurately diagnose left atrial enlargement in patients with atrial fibrillation[J]. J Thorac Imaging，2015，30(3)：214-217

[16] SHANG Y，ZHANG X，LENG W，et al. Left atrium passive ejection fraction is the most sensitive index of type 2 diabetes mellitus-related cardiac changes[J]. Int J Cardiovasc Imaging，2018，34(1)：141-151

[17] MORRIS D A，BELYAVSKIY E，ARAVIND-KUMAR R，et al. Potential usefulness and clinical relevance of adding left atrial strain to left atrial volume index in the detection of left ventricular diastolic dysfunction[J]. JACC Cardiovasc Imaging，2018，11(10)：1405-1415

[收稿日期] 2020-10-30