·方法研究·

ARIMA 模型在天津市结核病发病预测中的应用

李晓蓉, 庞学文, 于燕明, 高丽, 张丹, 万莹, 李敬新
天津市结核病控制中心,天津 300041

摘要: **目的** 构建天津市结核病月发病数的 ARIMA 模型,为结核病防控工作提供参考。 **方法** 采用 SPSS 16.0 统计软件包对天津市 2005 年 1 月-2016 年 7 月结核病月发病数资料建立最佳 ARIMA 预测模型,利用 2016 年 8 月-2017 年 7 月结核病月发病数对模型进行效果评价,并对 2017 年 8 月-2018 年 7 月结核病月发病数进行预测。 **结果** 建立的 ARIMA(0,1,1)(0,1,1)模型是拟合天津市结核病月发病人数的最优模型,利用 2016 年 8 月-2017 年 7 月结核病月发病数对模型进行效果评价,发病人数在 3-6 月有一个发病高峰,符合历年结核病发病趋势,且实际发病数均落在预测值 95%可信区间内,实际发病人数与预测发病人数的相对误差绝对值中位数为 2.49%,模型具有较高的预测精度。 **结论** ARIMA(0,1,1)(0,1,1)模型能够较精确的预测天津市结核病月发病情况,可为结核病的预防和控制提供重要理论依据。

关键词: ARIMA; 时间序列分析; 结核病; 预测
中图分类号:R521 **文献标识码:**B **文章编号:**1006-3110(2018)12-1536-03 DOI:10.3969/j.issn.1006-3110.2018.12.038

结核病是一种严重危害人类健康的慢性呼吸道传染病,因其传染性强、人群普遍易感、发病后治疗时间长、治愈率偏低等特点而受到特别重视,我国作为全球结核病高负担国家之一,为加强对结核病的管理和监测,从 2004 年开始建立以网络为基础的“疾病监测信息管理系统”和“结核病管理信息系统”,为实时的统计结核病的发病信息,制定相应的预防控制策略提供了便利。但是缺乏有效的预测预警分析手段,结核病防控工作只能处于被动应付的地位。本研究通过时间序列分析,利用 2005 年 1 月-2016 年 7 月天津市结核病月发病人数资料构建自回归滑动平均混合模型,对未来天津市结核病的月发病数进行预测,为今后的结核病防控工作提供依据。

1 资料与方法

1.1 *数据来源* 数据来源于结核病信息管理系统中 2005 年 1 月-2017 年 7 月登记的天津市结核病月发病资料。

1.2 *方法*

1.2.1 ARIMA 模型的理论基础 自回归滑动平均混合模型(autoregressive intergrated moving average, ARIMA)是由博克思(Box)和詹金斯(Jenkins)于上世纪 70 年代初提出的著名时间序列预测方法,所以又称为 Box-jenkins 模型、博克思-詹金斯法。该方法将预测对象随时间推移而形成的数据序列视为一个随机序列,即去除个别因偶然原因引起的观测值外,时间序列是一组依赖与时间 t 的随机变量,这组随机变量所具有的依存关系或者自相关性表征了预测对象发展的延续性,而这种依存关系或者自相关性一旦被相应的数学模型描述出来,就可以从时间序列的过去值及现在值来预测未来值。ARIMA 模型可分为自回归模型(AR),滑动平均模型(MA)和自回归滑动平均混合模型(ARIMA)[1]。

1.2.2 ARIMA 模型的建立 以 2005 年 1 月-2016 年 7 月天津市结核病月发病数为应变量,利用 SPSS 16.0 统计软件包时间序列分析模块的“专家建模器”自动拟合最优 ARIMA(p, d, q),(P,D,Q)。其中 p 和 P 分别代表自回归阶数和季节性自回归阶数,d 和 D 分别代表差分阶数和季节性差分阶数,q 和 Q 分别代表移动平均阶数和季节性移动平均阶数。利用 2016 年 8 月-2017 年 7 月结核病月发病数对模型预测效果进行评价,对 2017 年 8 月-2018 年 7 月结核病月发病数进行预测,并生成这些预测值的 95%可信区间。

1.3 *统计分析* 所有资料均采用 SPSS 16.0 统计软件包进行分析,$P<0.05$ 为差异有统计学意义。

2 结 果

2.1 *结核病月登记发病数变化趋势* 2005 年 1 月-2016 年 7 月天津市结核病月登记发病数总体呈波动性下降趋势。2006-2008 年出现一个发病高峰,随后几年趋于下降趋势。这 12 年期间每个月都有结核病病例登记,且月登记数呈现周期性波动,在大部分年份中,结核病登记主要集中在 3-6 月,具有季节性,

作者简介:李晓蓉(1988-),女,硕士,医师,主要从事结核病控制工作。
通信作者:庞学文,E-mail:pxw912@126.com。

见图 1。

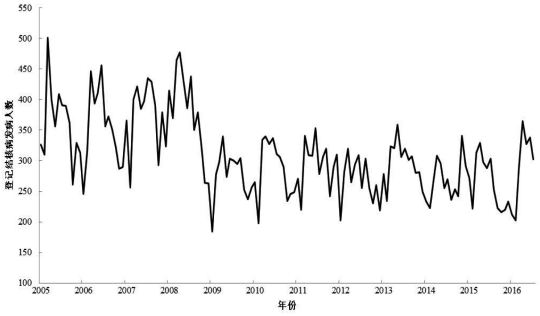

图 1　天津市 2005 年 1 月-2016 年 7 月结核病月登记发病数时序图

2.2　模型构建

2.2.1　ARIMA 月发病数模型　以 2005 年 1 月-2016 年 7 月结核病月发病数为应变量，在 SPSS 16.0 软件中时间序列模块的“专家建模器”构建最佳 ARIMA 模型。最终确定的天津市结核病月发病数 ARIMA 模型为 ARIMA(0,1,1)(0,1,1)，并绘制了残差序列的自相关函数(ACF)和偏自相关函数(PACF)图，残差的 Box-Ljung Q 统计量无统计学意义(Q = 26.086，P = 0.053)，表明剩余项独立，残差序列是白噪声。因此，可认为用 ARIMA(0,1,1)(0,1,1)模型进行天津市结核病月发病数的预测是比较理想和合理的。见图 2。

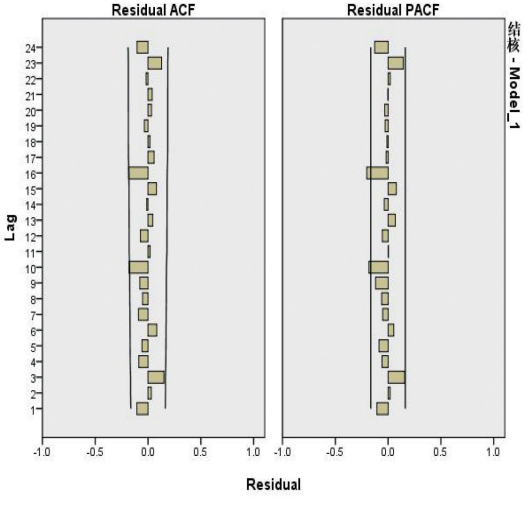

图 2　模型 ARIMA(0,1,1)(0,1,1)的残差序列的自相关函数和偏自相关函数图

2.2.2　模型预测效果评价　将 2016 年 8 月-2017 年 7 月的月发病数进行回代预测，预测效果表明结核病发病人数在 3-6 月有一个发病高峰，实际发病人数与预测发病人数的相对误差绝对值中位数为 2.49%，预测结果符合实际发病人数的变化规律，且 2016 年 8 月-2017 年 7 月的实际发病人数均在预测发病人数的 95%可信区间内，说明模型具有较精确的模拟效果，可以对天津市结核病的发病数进行较准确的预测。见图 3 和表 1。

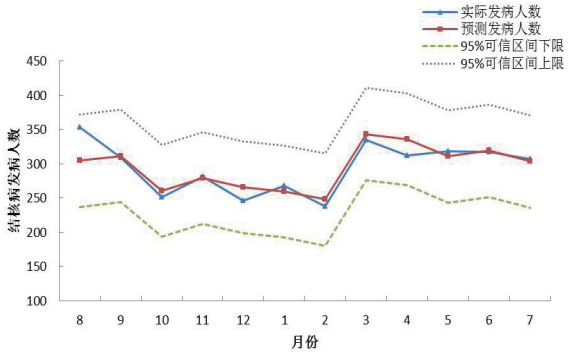

图 3　2016 年 8 月-2017 年 7 月天津市结核病发病人数预测结果

表 1　2016 年 8 月-2017 年 7 月天津市结核病发病人数预测值与实际值的比较

月份	实际发病人数	预测发病人数	95%可信区间		相对误差(%)
			LCL	UCL	
8	353	305	237	372	-13.69
9	309	311	244	379	0.78
10	251	261	193	328	3.83
11	281	279	212	347	-0.61
12	246	266	198	333	8.05
1	268	260	192	327	-3.13
2	238	248	181	316	4.29
3	335	343	276	411	2.49
4	312	336	268	403	7.65
5	318	311	244	379	-2.17
6	317	319	252	387	0.73
7	307	304	236	371	-1.03
中位数	-	-	-	-	2.49

2.2.3　模型预测应用　用 ARIMA(0,1,1)(0,1,1)模型对 2017 年 8 月-2018 年 7 月结核病月发病人数进行预测，预测期间的结核病月发病人数将在 234~334 例之间，见图 4、表 2。

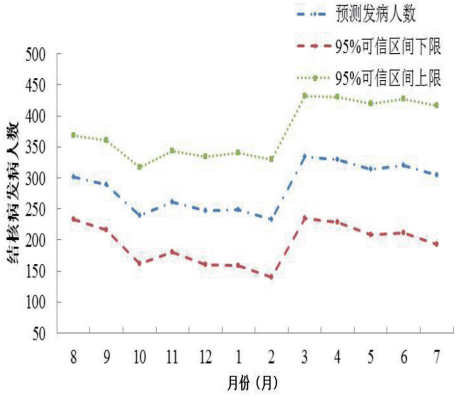

图 4　2017 年 8 月-2018 年 7 月天津市结核病发病人数预测结果

表 2 2017 年 8 月-2018 年 7 月天津市结核病月发病人数预测结果

月份	预测发病人数	95%可信区间	
		LCL	UCL
8	301	234	369
9	289	216	361
10	239	162	317
11	262	180	344
12	247	161	334
1	249	159	340
2	234	140	329
3	334	235	432
4	329	228	431
5	314	209	419
6	320	211	428
7	305	193	417

3 讨 论

近年来,统计预测方法越来越多的应用于疾病的预测,通过对疾病发病的预测,为制定疾病预防控制措施提供理论依据,实现早预防,早控制。统计预测方法有很多,根据分类可分为定性预测和定量预测,目前用的主要是定量预测方法,定量预测是根据历史数据,运用数学方法进行预测的方法,包括指数平滑法、马尔科夫链预测法、灰色预测法、ARIMA 预测法、神经网络预测法等。在疾病预测方面,ARIMA 法综合考虑了序列的趋势变化、周期变化及随机干扰,并借助模型参数进行了量化表达[2],在预测精度方面,ARIMA 模型对噪声进行了分析处理,只剩下当时和历史无关的白噪声,排除了人们主观判断的随意性,因此具有较高的精确度,目前,ARIMA 法已经广泛的应用于疾病特别是传染病的预测[3-9]。

本文应用 SPSS 统计软件包中的时间序列模块的“专家建模器”过程自动拟合最优的 ARIMA 模型,免去了传统的对模型结果和参数间的相关性进行反复调试、检验识别的复杂过程,使得模型构建过程操作简单可行。经过拟合,最终确定 ARIMA(0,1,1)(0,1,1)为最优的结核病月发病数模型。一方面,通过将 2016 年 8 月-2017 年 7 月结核病月发病数进行回代预测,预测结果较好的拟合了实际发病人数的变化规律,说明模型具有较高的精确度,适用于对天津市结核病未来月发病数的预测。另一方面,本文通过 ARIMA(0,1,1)(0,1,1)预测了 2017 年 8 月-2018 年 7 月天津市结核病的月发病数,可以通过预测值来估计未来结核病的流行强度,若 2017 年 8 月-2018 年 7 月实际发病人数在预测发病人数的 95%可信区间内,表明当月的结核病疫情基本正常,若发现实际发病人数处于预测发病人数的 95%可信区间外,则提示有可能发生结核病疫情暴发,应予以重视。因此,可以根据本文所拟合的 ARIMA 模型,对天津市结核病未来月发病人数进行早期预测、预警,及早采取相应的控制措施,为进一步加强结核病防控工作提供参考依据。

由于所建的模型是以历史数据为基础建立的,而结核病的发生规律不是一成不变的,会受到多种因素的影响,因此要不断的加入新的监测数据对模型进行不断的修正,以提高模型预测的精确度,正确掌握天津市结核病疫情的变化趋势,及早采取针对性措施控制结核病疫情的发生,减少给病人、家庭和社会带来的严重危害。

参考文献

[1] 刘桂芬,刘玉秀,仇丽霞,等. 医学统计学[M]. 第 2 版. 北京:中国协和医科大学出版社, 2009:354-365.
[2] 王振龙,胡永宏. 应用时间序列分析[M]. 北京:科学技术出版社, 2007:42-124.
[3] 刘继恒,白春林,孙要武,等. 应用 ARIMA 模型预测肺结核报告发病率的研究[J]. 中国热带医学,2014,14(9):1067-1070.
[4] 郑慧敏,薛允莲,黄燕飞,等. ARIMA 模型在深圳市法定传染病发病趋势预测的应用[J]. 实用预防医学,2016,23(2):240-243.
[5] 徐娜,霍飞,刘长娜, 等. ARIMA 模型在梅毒预测中的应用[J]. 疾病监测,2011,26(2):103-105,109.
[6] 张光,孙良,谢金贵,等. ARIMA 模型在阜阳市手足口病发病数预测中的应用[J]. 安徽预防医学杂志,2015,21(4):231-234.
[7] 孟蕾,王新华,白亚娜,等. 甘肃省哨点医院流感样病例 ARIMA 模型预测[J]. 中国公共卫生,2014,30(2):228-230.
[8] 王怡,张震,范俊杰,等. ARIMA 模型在传染病预测中的应用[J]. 中国预防医学杂志,2015,16(6):424-428.
[9] 刘继恒,贺圆圆,张皓,等. 基于时间序列模型对甲型病毒性肝炎的预测研究[J]. 实用预防医学,2017,24(8):1009-1011.

收稿日期:2017-10-20