·基础理论与方法·

自回归移动平均混合模型在中国道路交通伤害预测中的应用

庞媛媛　张徐军　涂志斌　崔梦晶　顾月

【导读】 探讨时间序列分析的自回归移动平均混合模型(ARIMA)在中国道路交通伤害(RTI)预测中的应用。收集1951－2011年中国道路交通伤害资料，进行时间序列分析，建立ARIMA模型。构建得到RTI事故起数ARIMA(1,1,0)预测模型为$Y_t = e^{Y_{t-1}+0.456\nabla Y_{t-1}+e_t}$，其中，$e_t$为随机误差，模型残差序列为白噪声，Ljung-Box检验$P>0.05$，统计量无统计学意义，拟合效果良好。应用该模型预测2011年中国RTI事故起数，预测值与实际观测结果相符，实际观测值在预测值95%*CI*内。用该模型预测2012年中国RTI事故起数，预测值(95%*CI*)为207 838(107 579～401 536)。应用ARIMA模型能较好地预测中国道路交通伤害情况。

【关键词】 道路交通伤害；时间序列分析；自回归移动平均混合模型；预测

Autoregressive integrated moving average model in predicting road traffic injury in China

PANG Yuan-yuan, ZHANG Xu-jun, TU Zhi-bin, CUI Meng-jing, GU Yue. Injury Prevention Research Institute/School of Public Health, Southeast University, Nanjing 210009, China

Corresponding author: ZHANG Xu-jun, Email: xjzhang@seu.edu.cn

【Introduction】 This research aimed to explore the application of autoregressive integrated moving average (ARIMA) model of time series analysis in predicting road traffic injury (RTI) in China and to provide scientific evidence for the prevention and control of RTI. Database was created based on the data collected from monitoring sites in China from 1951 to 2011. The ARIMA model was made. Then it was used to predict RTI in 2012. The ARIMA model of the RTI cases was $Y_t = e^{Y_{t-1}+0.456\nabla Y_{t-1}+e_t}$ (e_t stands for random error). The residual error with 16 lags was white noise and the Ljung-Box test statistic for the model was no statistical significance. The model fitted the data well. True value of RTI cases in 2011 was within 95%*CI* of predicted values obtained from present model. The model was used to predict value of RTI cases in 2012, and the predictor (95%*CI*) was 207 838 (107 579–401 536). The ARIMA model could fit the trend of RTI in China.

【Key words】 Road traffic injury; Time series analysis; Autoregressive integrated moving average model; Forecasting

道路交通伤害(RTI)是目前全球第十位死因，已成为不可忽视的社会安全和公共卫生问题。若不实施有效的干预，到2020年，RTI将成为全球第三位死因[1]。RTI绝大多数发生在发展中国家，只约有10%发生在发达国家[1]。据WHO预计，同1990年相比，2020年道路交通死亡在发展中国家将平均上升80%，在发达国家将下降近30%[1]。我国是最大的发展中国家，也是RTI最多的国家之一，随着经济迅速发展，机动车数量不断快速增长，但交通管理的改善和道路建设的发展却明显滞后，RTI及伤亡人数呈不断上升趋势[2]。我国RTI死亡60%发生在16～45岁的中青年人，对劳动生产力人口造成严重影响[3,4]。因此有必要研究RTI的流行规律及其发展趋势。本研究通过分析我国1951－2010年RTI监测资料，运用时间序列分析中的自回归移动平均混合模型(ARIMA)，对RTI建模拟合，并预测2011年我国RTI的发生情况。

基本原理

1. 模型及其公式：包括自回归模型(AR)$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + e_t$；移动平均模型(MA)$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$；ARIMA $Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$。其中，p、q分别为AR和MA的阶数，$e_t$为误

DOI: 10.3760/cma.j.issn.0254-6450.2013.07.018

作者单位：210009 南京，东南大学伤害预防研究所 公共卫生学院流行病与卫生统计学系

通信作者：张徐军，Email: xjzhang@seu.edu.cn

差或偏差，表示不能用模型说明的随机因素。运用ARIMA预测方法的前提是作为预测对象的时间序列是零均值的平稳随机序列。因此，在应用ARIMA模型前，应对时间序列先进行零均值化和差分平稳化处理。一阶差分：$\nabla Y_t=Y_t-Y_{t-1}(t>1)$；二阶差分：$\nabla^2 Y_t=\nabla Y_t-\nabla Y_{t-1}(t>2)$；依此类推。差分后，模型为ARIMA(p,d,q)，其中p是自回归的阶，d是差分次数，q是移动平均的阶。

2. 建立模型：以我国1951—2010年道路交通伤害监测数据为基础建立数据库，运用SPSS 13.0统计软件进行数据处理与分析，建立ARIMA模型。①数据预处理：通过序列图和自相关系数判断平稳性，若平均值和方差始终为常数，则为平稳序列，否则应用自然对数转换、差分等方法将其平稳化；②模型的识别、定阶与参数估计：采用Box-Jenkins法进行阶数识别和定阶，采用最小二乘法或非线性估计法进行参数估计；③模型的拟合优度检验：对观测值和拟合值的残差进行分析，若残差序列为白噪声序列，则所建模型为最终模型，否则需重复上述步骤，重新建模，直至残差序列为白噪声序列。

3. 预测：利用所建立的ARIMA模型对我国1951—2010年RTI情况进行回代预测，并预测2011年我国RTI情况。

实例分析

1. 资料：

(1)资料来源：数据来源于公安部交通管理局道路交通事故统计年报。主要包括1951—2011年各年度道路交通事故起数、死亡人数、受伤人数、万车死亡率和10万人口死亡率。

(2)缺失值处理：数据缺失3.28%，比例较小，可用线性插值法对缺失值进行填补，使数据信息完整并符合时间序列分析的特点。

2. RTI事故起数趋势分析：由图1可见，1951—2010年我国RTI事故起数呈明显的非平稳性，需对其进行差分以平稳时间序列，并通过对数转换减小方差波动。

3. 建立模型：

(1)数据平稳化：对RTI事故起数数据进行自然对数变换，再进行一阶差分。数据预处理后，序列自相关系数落入随机区间(图2)，时间序列平稳。

(2)模型的识别、定阶和参数估计：根据数据预处理后序列自相关和偏自相关函数图(图2、3)，初步选定ARIMA(1,1,0)、ARIMA(1,1,1)、ARIMA(0,1,1)3个模型进行拟合预测。应用SPSS 13.0统计软件进行参数估计(表1)。ARIMA(1,1,1)参数无统计学意义，除去该模型，ARIMA(1,1,0)、ARIMA(0,1,1)常数项无统计学意义，去掉常数项，再次进行检验，结果见表2。比较模型的AIC和SBC值，结果显示ARIMA(1,1,0)的AIC、SBC值最小，表明该模型最适合本研究(表3)。根据参数估计结果，ARIMA模型为：$\nabla Y_t=0.456\nabla Y_{t-1}+e_t$，$Y_t'=Y'_{t-1}+0.456\nabla Y_{t-1}+e_t$；事故起数预测公式：$Y_t=e^{Y_{t-1}+0.456\nabla Y_{t-1}+e_t}$。

(3)模型的拟合优度检验：分析残差序列，判断模型优劣。由图4可见，残差序列自相关和偏自相关系数均在95%*CI*内，残差为白噪声。Ljung-Box检验$P>0.05$，统计量无统计学意义，说明残差为随机

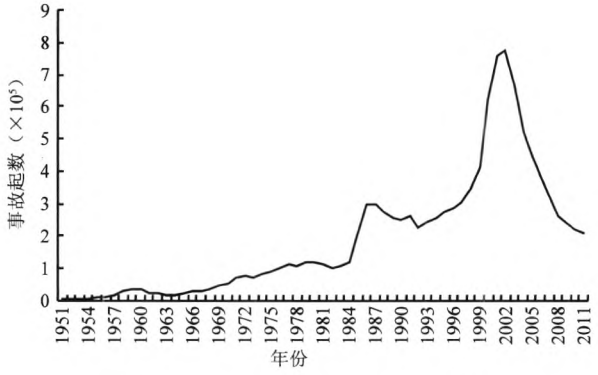

图1 1951—2011年中国RTI事故起数

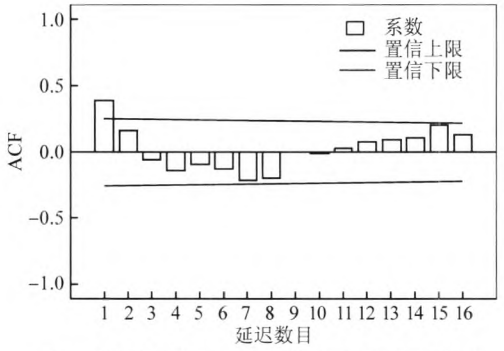

图2 数据预处理后序列自相关系数函数图

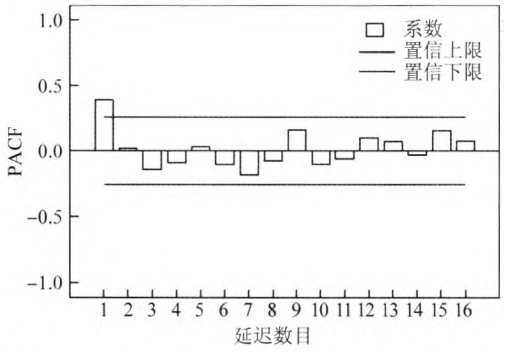

图3 数据预处理后序列偏自相关系数函数图

表 1　ARIMA 参数估计

参数	ARIMA(1,1,0)			ARIMA(1,1,1)			ARIMA(0,1,1)		
	估计	*t* 值	*P* 值	估计	*t* 值	*P* 值	估计	*t* 值	*P* 值
AR1	0.403	3.331	0.002	0.367	1.192	0.238	–	–	–
MA1	–	–	–	−0.045	−0.136	0.893	−0.383	−3.054	0.003
常数	0.056	1.411	0.164	0.056	1.418	0.162	0.057	1.705	0.094

表 2　ARIMA 调整参数估计

参数	ARIMA(1,1,0)			ARIMA(0,1,1)		
	估计	*t* 值	*P* 值	估计	*t* 值	*P* 值
AR1	0.456	3.924	0.000	–	–	–
MA1	–	–	–	−0.426	−3.476	0.001

表 3　ARIMA 拟合优化结果

指标	ARIMA(1,1,0)	ARIMA(0,1,1)
log likelihood	15.834	14.715
AIC	−29.669	−27.430
SBC	−27.591	−25.353

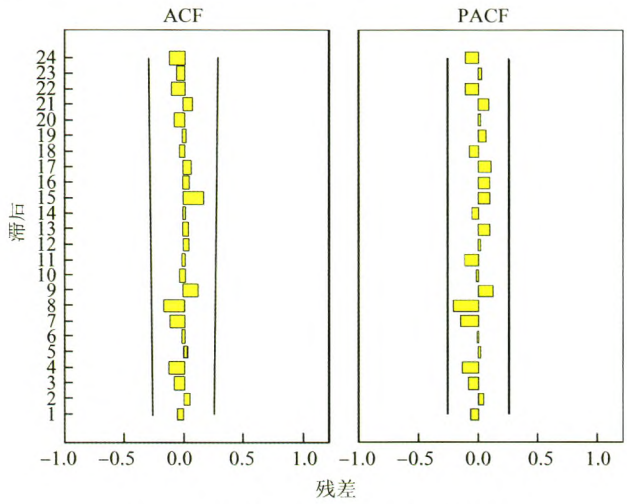

图 4　残差自相关和偏自相关函数图

误差，所建模型为最终模型。

4. 模型的预测：应用所建模型对 1951—2010 年我国 RTI 事故起数的时间序列进行回代预测，并预测 2011 年 RTI 发生情况。如图 5 所示，1951—2010 年我国 RTI 事故起数观测值与预测值基本相符，预测值的 95%*CI* 范围较小，表明该模型可较好地模拟我国 RTI 事故起数变化情况。对 2011 年我国 RTI 事故起数进行预测，预测值及其 95%*CI* 为 211 430（145 638 ~ 306 945），预测结果与实际观测相符，观测值在预测值的 95%*CI* 内。用该模型预测 2012 年我国 RTI 事故起数，预测值及其 95%*CI* 为 207 838（107 579 ~ 401 536）。

5. 万车死亡率和 10 万人口死亡率建模预测：应用上述相同方法对 1951—2010 年我国 RTI 的万车死亡率和 10 万人口死亡率建模，结果万车死亡率模型为 ARIMA(0,1,0)，预测公式为 $Y_t = e^{Y_{t-1} - 0.064 + e_t}$，预测结果见图 6；10 万人口死亡率模型为 ARIMA(4,1,0)，预测公式为 $Y_t = e^{Y_{t-1} + 0.063 - 0.453 \nabla Y_{t-4} + e_t}$，预测结果见图 7。

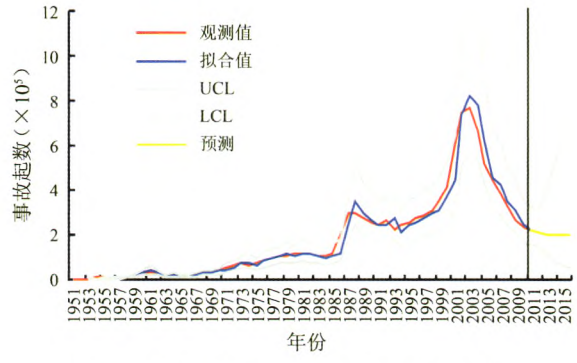

图 5　应用 ARIMA 预测我国 RTI 事故起数

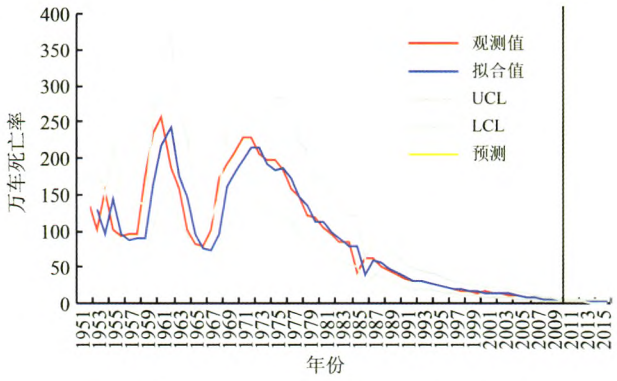

图 6　应用 ARIMA 预测我国 RTI 万车死亡率

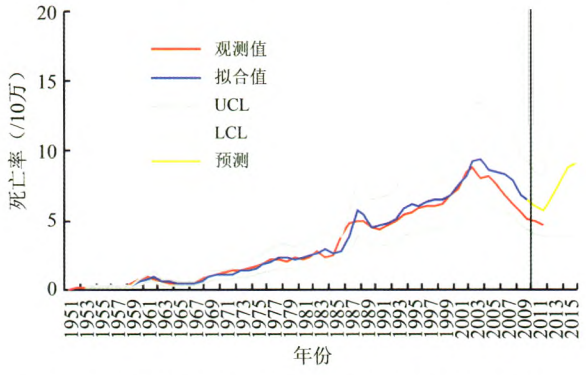

图 7　应用 ARIMA 预测我国 RTI 10 万人口死亡率

讨　论

在 RTI 预防控制中有效、准确的预测具有重要意义。RTI 的发生受多种因素影响，运用多元回归等静态因果结构模型预测往往很困难，而时间序列分析不是考虑变量间的因果关系，而是重点观察变量在时间上的发展变化规律并建立分析模型，进行趋势预测[5]。本研究运用时间序列分析分析我国 1951—2011 年 RTI 并建立模型进行预测。

ARIMA 模型是 Box-Jenkins 方法中重要时间序

列分析预测模型，已被广泛应用于时间序列分析。在时间序列资料的典型特征较难辨别的情况下，ARIMA模型能综合考虑序列的趋势变化、周期变化和随机干扰，借助参数进行量化表达，反映时间序列中所包含的动态依存关系较为精确，具有明显的优势和特色[6-8]。模型预测的优劣最终需看预测值与实际值的相符情况。本研究对RTI各项指标建模，预测结果显示与实际值较接近，说明建立的模型较好，可用于我国RTI发生情况的预测。

我国RTI特征包括机动车数量增加、道路交通事故数多、RTI死亡率高、道路交通管理与发展不协调、驾驶员安全意识不高等[9-11]。2000年我国RTI死亡上升，已成为第一位的伤害死因[12]，事故起数亦呈逐年上升趋势，虽在2003年后有所下降，但仍超过60万起，死亡人数居高不下。因此对RTI发生情况有必要进行预测。

ARIMA模型只适用于短期预测[13]，如本研究图5所示，外推时间延长，预测值的95%*CI*增大，说明预测精度降低。已有研究表明RTI的发生具有季节性和时间聚集性[14]，但此两特征并不像某些传染病那样明显，且由于本研究所采用数据的局限性，未能对其分析，若进行分析，模型的精度应有提高。

参 考 文 献

[1] World Health Organization. World report on traffic injury prevention. Geneva/New York: World Health Organization/the World Bank, 2004.

[2] Wang ZG. Traffic accident in 2001, in China. Chin J Traumatol, 2003, 19(11): 645-648. (in Chinese)
王正国. 我国2001年的交通事故. 中华创伤杂志, 2003, 19(11): 645-648.

[3] Zhang X, Xiang H, Jing R, et al. Road traffic injuries in the People's Republic of China, 1951-2008. Traffic Inj Prev, 2011, 12(6): 614-620.

[4] Chi GB, Wang SY. Pattern of road traffic injuries in China. Chin J Epidemiol, 2004, 25(7): 598-601. (in Chinese)
池桂波，王声湧. 中国道路交通伤害的模式. 中华流行病学杂志, 2004, 25(7): 598-601.

[5] Gao WW, Guo CY, Zhou YJ. Application of time-series analysis in China's public health fields. Chin J Soc Med, 2011, 28(2): 78-80. (in Chinese)
高围微，郭常义，周义军. 时间序列分析在我国公共卫生领域的应用. 中国社会医学杂志, 2011, 28(2): 78-80.

[6] Liu Q, Liu X, Jiang B, et al. Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model. BMC Infect Dis, 2011, 11: 218.

[7] Martinez EZ, Silva EA. Predicting the number of cases of dengue infection in Ribeirao Preto, Sao Paulo State, Brazil, using a SARIMA model. Cad Saude Publica, 2011, 27(9): 1809-1818.

[8] Fernandez-Gonzalez M, Rodriguez-Rajo FJ, Jato V, et al. Forecasting ARIMA models for atmospheric vineyard pathogens in Galicia and Northern Portugal: Botrytis cinerea spores. Ann Agric Environ Med, 2012, 19(2): 255-262.

[9] Zhang XJ, Jia J, Chen ZQ, et al. Epidemiological analysis on road traffic injury in China in 2004. Chin J Epidemiol, 2007, 28(2): 204-205. (in Chinese)
张徐军，贾佳，陈宗道，等. 中国2004年道路交通伤害的流行病学研究. 中华流行病学杂志, 2007, 28(2): 204-205.

[10] Huang KY, Yang L. Progress of epidemiological research of road traffic injury. Chin J Prev Control Chronic Non-Communicable Dis, 2012, 20(2): 217-220. (in Chinese)
黄开勇，杨莉. 道路交通伤害的流行病学研究进展. 中国慢性病预防与控制, 2012, 20(2): 217-220.

[11] Dong XM, Peng L, Wang SY. Progress of intervention study of road traffic injuries. Chin J Public Health, 2012, 28(5): 569-571. (in Chinese)
董晓梅，彭淋，王声湧. 道路交通伤害干预研究进展. 中国公共卫生, 2012, 28(5): 569-571.

[12] Chi GB, Wang SY. Study on the secular trend of road traffic injuries and its influencing factors in China. Chin J Epidemiol, 2007, 28(2): 148-153. (in Chinese)
池桂波，王声湧. 中国道路交通伤害长期趋势及其影响因素分析. 中华流行病学杂志, 2007, 28(2): 148-153.

[13] Wangdi K, Singhasivanon P, Silawan T, et al. Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: a case study in endemic districts of Bhutan. Malar J, 2010, 9: 251.

[14] Liu GF, Han S, Liang DH, et al. Epidemiological characteristics of road traffic accidents during 2001 in the city of Shenyang. Chin J Traumatol, 2003, 19(9): 524-526. (in Chinese)
刘改芬，韩松，梁多宏，等. 2001年沈阳市道路交通事故流行病学特点. 中华创伤杂志, 2003, 19(9): 524-526.

（收稿日期：2013-01-21）
（本文编辑：张林东）