文章编号：1006-3617(2011)01-0009-04　　中图分类号：R18　　文献标志码：A

【专栏】

广义相加模型在上海世博会园区医疗站就诊人数预测中的应用

陶芳芳，虞慧婷，林庆能，何懿，冯玮，董晨，郭翔，毛智盛，孙晓冬*

摘要：［目的］探讨应用广义相加模型进行中国2010年上海世博会园区医疗站就诊人数预测的可行性。［方法］采用时间序列的广义相加模型，在控制星期效应的基础上，对2010年5月1日至8月8日开园100 d间的每日世博园区就诊人数、入园人数和气象因素资料进行模型拟合。［结果］园区就诊人数存在周末效应，就诊人数随入园总人数的增加而增加；风速对就诊人数的影响趋势是随风速的增加，就诊人数先增加后减少；随着气温上升和日温差增加，就诊人数呈上升趋势（$P<0.01$）。利用模型对每日就诊人数进行预测，预测的平均相对误差为10.44%。［结论］广义相加模型能较好拟合世博园区医疗站就诊人数的趋势，可用于大型活动中的预测研究。

关键词：2010上海世博会；就诊人数；预测；广义相加模型

Application of Generalized Additive Model in Forecasting Cases of Medical Sites in Expo 2010 Shanghai China

TAO Fang-fang, YU Hui-ting, LIN Qing-neng, HE Yi, FENG Wei, DONG Chen, GUO Xiang, MAO Zhi-sheng, SUN Xiao-dong (Shanghai Maniciple Center for Disease Control and Prevention, Shanghai 200336, China). *Address correspondence to SUN Xiao-dong; E-mail: xdsun@scdc.sh.cn*

Abstract: [Objective] To explore application of generalized additive model in forecasting cases of medical sites in Expo 2010 Shanghai China. [Methods] By using a generalized additive model with time series adjustment for weekend, the daily cases of medical sites, the total visitors to the Expo and the meteorological factors from May 1st to Aug 8th were analyzed and the forecast was compared with the actual cases. [Results] The number of medical site cases showed a weekend effect. With the total visitors increasing, the number of cases increased first and then decreased. With the daily temperature increased, the number of cases increases. The average relative error of prediction was 10.44%. [Conclusion] A generalized additive model fit the number of cases of the Expo medical stations and can be used for similar forecast in mass gathering in the future.

Key Words: Expo 2010 Shanghai China; cases of medical sites; forecasting; generalized additive model

中国2010年上海世博会（简称“上海世博会”）是一次国际性大型集会活动，通常由于参加人数多、人群密度高，发生踩踏、传染病流行等突发事件的风险较大，历史上国际奥林匹克运动会、穆斯林麦加朝圣等大型集会活动均有类似的经验教训[1]。为了实时掌握世博会期间园区内游客及相关人员的就诊信息和满足各种可能的就诊需求，上海市卫生局在世博园区内和园区外分别设置了医疗站和定点医院，其中医疗站为各定点医院在园区内的前哨，为游客及园区内的工作人员和志愿者提供医疗服务。为做好就诊人员的就诊信息监测，预警异常情况并进行及时应对，本课题组开发了中国2010年上海世博会园区内就诊异常情况报告和预警系统（简称“就诊监测报告系统”），通过该系统，可以动态掌握每日园区就诊人群的基本情况，及早发现突发传染病、群体性食物中毒、化学中毒及其他群体性公共卫生事件。

[作者简介]陶芳芳（1980–），女，硕士，主管医师；研究方向：流行病与卫生统计学；E-mail：fftao@scdc.sh.cn

[*通信作者]孙晓冬副主任医师；E-mail：xdsun@scdc.sh.cn

[作者单位]上海市疾病预防控制中心，上海 200336

了解园区就诊趋势，尽早发现可能的就诊高峰，在就诊高峰到来之前及时提供准确的信息及预警预报，以便采取相应的应对措施，有助于提高园区公共卫生事件的反应能力，从而保证园区正常运行。预测是关注某类观察目标处于何种状况，对每日园区就诊人数进行预测，推断就诊人数的趋势，可为园区防控措施的制定提供科学依据。由于就诊人数可能受多种因素影响，如入园人数、气象因素、各种干预措施等社会经济因素影响，所以单一自变量时间序列模型如ARIMA模型不能满足需要，要考虑影响因素相互关联与制约的复杂关系，所以本研究采用广义相加模型（generalized additive model，GAM）来建立园区就诊人数的预测模型。与传统方法比，此模型对应变量的分布没有限制，适用于多种分布类型的数据，通过“加性”的假设，将一些与应变量存在复杂非线性关系的自变量以函数加和的形式拟合入模型，能探索到变量间的非单调、非线性关系，灵活性强。从1990年HASTIE和TIBSHIRANI系统地阐述了GAM的理论后，国外越来越多的医学研究者开始将该模型应用于医学数据的分析，取得了较为满意的效果[2-4]。

本研究拟应用时间序列资料，通过广义相加模型分析世博

园区入园人数、气象因素等对就诊人数的影响，探讨使用此模型进行园区就诊人数预测的可行性，为园区公共卫生安全保障工作提供科学依据。

1　材料与方法

1.1　资料来源

通过“中国2010年上海世界博览会园区内就诊异常情况报告和预警系统”收集2010年5月1日至2010年8月31日每日世博园区就诊人数资料，该资料由上海市疾病预防控制中心提供。

2010年5月1日至2010年8月31日世博园区气象因素资料由上海市气象局城市环境气象中心提供。包括日最高气温、日最低气温、日平均气温、日温差、日平均相对湿度、日平均气压、平均风速。

1.2　统计分析

1.2.1　*单因素描述分析*　绘制散点图和由统计指标初步分析就诊人数与气象资料、入园总人数等相关因素的关系和分布类型。

1.2.2　*GAM建模分析*　GAM为非参数回归方法，适用于多种分布类型、多种非线性关系的分析。GAM的基本形式为$g(\mu_i)=\beta_0+f_1(x_{1i})+f_2(x_{2i})+\cdots+f_n(x_{ni})+\varepsilon$。

本研究中，拟将开园天数、气象因素（气温、日温差、相对湿度、气压、风速等）、入园总人数纳入模型，考虑就诊人数可能受气温、气压、湿度、降雨量等气象因素及入园总人数的影响，而这些因素与就诊人数为非线性关系，因此，采用GAM建立预测模型。设t为开园时间，Yt为第t日的就诊人数。对于t，Yt服从总体均数为$E(Yt)$的Possion分布，GAM连接函数为对数函数，基本形式为：

$$\log(E(Yt))=\alpha+f(t)+s(T_t)+s(T_{1t})+s(RH_t)+s(AP_t)+s(WS_t)+s(SY_t)+\gamma(Dow_t)+\varepsilon_t$$

式中：α为截距；f为时间效应的系数；s为平滑样条函数；T_t为第t天日平均气温；T_{1t}为第t天日温差；RH_t为第t天日平均相对湿度；AP_t为第t天日平均气压；WS_t为第t天日平均风速；SY_t为第t天日入园总人数；γ为星期效应的回归模型系数；Dow_t为双休日效应的哑变量；ε_t为残差。

拟合平滑样条函数控制气温、气压、风速、相对湿度等因素的混杂，并纳入双休日哑变量控制短期趋势的作用。

选择2010年5月1日至8月8日开园100 d的数据进行模型拟合，并用拟合的模型进行短期预测，预测效果评价指标为相对误差。

本研究采用SAS 9.1软件进行统计分析。

2　结果

2.1　单因素描述分析

从图1可以看出世博园区就诊人数总体呈上升趋势，随着温度的升高和入园人数的增加，就诊人数也增加，入园人数为40~50万时，就诊人数达最高值。就诊人数随着开园时间呈现余弦函数的变化形式。日温差、气压和风速与就诊人数均存在相关关系，相对湿度与就诊人数的关系不明确，并非简单的线性关系。

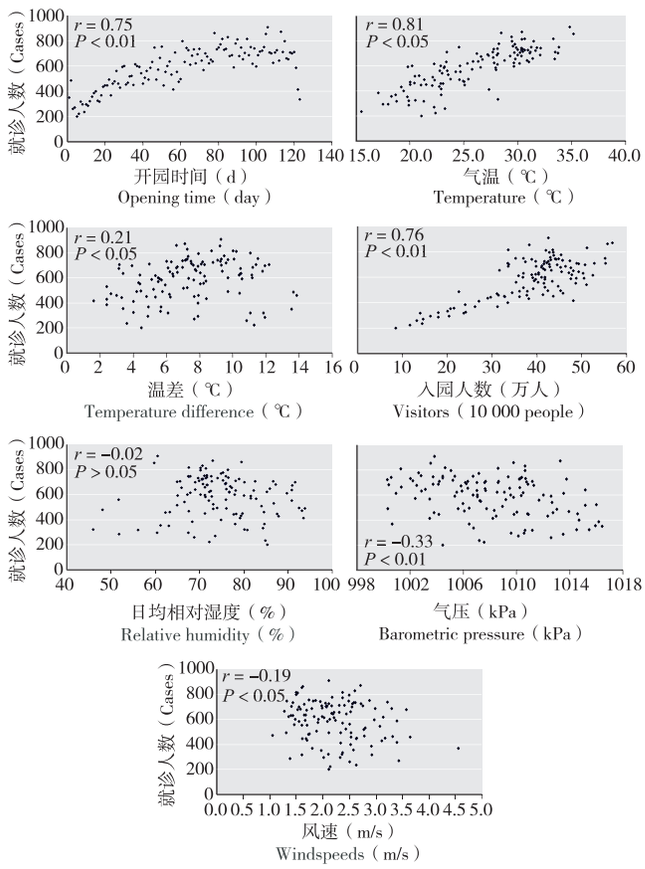

图1　就诊人数与开园时间、气温、温差、入园人数、日均相对湿度、气压、风速的关系

Figure 1　Correlation between medical visits and opening time, temperature, temperature difference, number of visitors, relative humidity, barometric pressure, windspeeds

2.2　模型相关指标的分布

2010年5月1日至2010年8月31日，世博园区每日入园人数在8.6万~56.8万人之间波动，中位数为41.2万人；每日园区医疗站就诊人数在201~908人之间，中位数为619人。气温在15.5~35.2℃之间，中位数为26.6℃；日温差在1.6~13.9℃之间，中位数为7.5℃；相对湿度在46.0%~93.8%之间，中位数为73.6%；气压在1 000.3~1 016.5 kPa之间，中位数为1 007.7 kPa；风速在1.1~4.6 m/s之间，中位数为2.1 m/s（表1）。

表1　2010年5月1日~8月31日园区就诊人数与气象因素、入园总人数的分布

Table 1　Distribution of medical visits and meteorological factors and number of visitors, May 1st-Aug 8th, 2010

指标（Factors）	范围（Range）	P_{25}	P_{50}	P_{75}
就诊人数（人） Vistis（Person）	201~908	462	619	709
气温（℃） Temperature（℃）	15.5~35.2	22.4	26.6	30.0
日温差（℃） Temperature difference（℃）	1.6~13.9	5.5	7.5	9.5
日平均相对湿度（%） Humidity（%）	46.0~93.8	69.6	73.6	80.0
日气压（kPa） Barometric pressure（kPa）	1 000.3~1 016.5	1 005.1	1 007.7	1 010.7
日风速（m/s） Windspeeds（m/s）	1.1~4.6	1.7	2.1	2.6
入园总人数（万人） Visitors（10 000 persons）	8.6~56.8	33.6	41.2	45.3

2.3　模型拟合结果

根据单因素分析结果，对于各引入的解释变量，根据序列图分析其分布类型，初步判断各变量之间的相关性。结果显示，日平均相对湿度、日平均气压未纳入模型。模型中通过设置星期哑变量来控制双休日效应，将开园天数的余弦形式纳入模型来控制时间效应，同时排除每日就诊人数的自相关。由于气温和温差有相关性，模型中将气温和温差用薄板样条函数（thin-plate spline，TPS）拟合，将其交互效应纳入模型。

在模型拟合中第一部分是模型的参数回归分析（表2），以线性参数形式进入模型的入园总人数有统计学意义（$t=32.55$，$P<0.0001$）；风速有统计学意义（$t=-5.34$，$P<0.0001$）。第二部分是光滑样条非参数分析的结果（表3），给出了在自由度为4的情况下，每一成分的光滑参数值、自由度、广义交叉确认（GCV）等。第三部分偏差分析中对模型中的每一个光滑效应给出F检验（表4），用于比较全模型和不含变量模型的偏差。偏差分析结果表明，入园总人数、风速和气温对就诊人数的效应有统计学意义（$P<0.01$）。

表2　模型参数部分的参数估计

Table 2　Parameter estimates of model fitting

变量 Parameter	参数估计值 Parameter Estimate	标准误 Standard Error	t	P
Intercept	5.56779	0.03200	174.01	<0.0001
Dow	0.03421	0.01205	2.84	0.0058
Cos（t）	0.02527	0.00922	2.74	0.0076
Linear（SY）	0.02155	0.00066	32.55	<0.0001
Linear（WS）	−0.03896	0.00730	−5.34	<0.0001

表3　模型非参数部分的光滑成分分析

Table 3　Summary of Smoothing components model fitting

光滑成分 Component	光滑参数 Smoothing parameter	自由度 v	广义交叉确认 GCV	变量取不同值的数目 Num unique obs
Spline（SY）	0.99942	5.06277	0.00715	98
Spline（WS）	0.76448	6.80392	0.00410	25
Spline2（T|T1）	35.56340	7.31673	3.58614	99

表4　模型非参数部分的离差分析

Table 4　Deviance analysis of the smoothing model

光滑成分 Source	自由度 v	平方和 Sum of Squares	χ^2 Chi-Square	P
Spline（SY）	5.06277	94.46514	94.4651	<0.0001
Spline（WS）	6.80391	63.60789	63.6079	<0.0001
Spline2（T|T1）	7.31673	933.94786	933.9479	<0.0001

预测模型如下：

$$Y=5.5678+0.0253\cos(0.0628t-0.7854)+0.0216SY+0.9994s(SY)-0.0390WS+0.7654s(WS)+35.5634s(T|T_1)+0.0342(Dow)$$

。式中t，SY，WS，T，Dow等含义同“1.2.1”中的解释。

从模型拟合结果可以看出，世博园区就诊人数存在周末效应，周六就诊人数高于周一至周五；就诊人数随入园总人数的增加而增加；风速对就诊人数的影响趋势是随风速的增加，就诊人数先增加后减少；随着气温上升和日温差增加，就诊人数呈上升趋势。

根据建立的模型，对2010年5月1日至8月8日开园100 d的就诊人数进行回代预测（组内回代），结果显示预测值与实际值平均相对误差为0.66%。利用模型对2010年8月9日至31日开园第101天至第123天每日就诊人数进行预测，结果见表5，其平均相对误差为10.44%。对预测值和实际值作图后，显示预测数据与实际数据吻合程度较高，其中预测在开园第105至106天和第113至114天时就诊人数将明显上升，出现就诊高峰，见图2。综上所述，模型拟合较好，可用于园区就诊人数的短期预测。

表5　拟合模型的预测值

Table 5　Prediction of the final model

开园天数（天） Opening time（day）	实际值 Observation values	预测值 Predicted values	相对误差（%） Relative error
101	687	682.52	−0.65
102	714	760.15	6.46
103	720	750.90	4.29
104	749	800.44	6.87
105	854	859.53	0.65
106	908	882.05	−2.86
107	708	748.30	5.69
108	701	797.61	13.78
109	624	734.81	17.76
110	706	742.51	5.17
111	726	775.24	6.78
112	642	803.17	25.10
113	871	862.53	−0.97
114	711	837.38	17.78
115	796	813.40	2.19
116	659	789.17	19.75
117	655	776.57	18.56
118	707	852.30	20.55
119	699	832.81	19.14
120	709	766.36	8.09
121	584	653.88	11.97
122	413	467.96	13.31
123	335	404.40	20.72

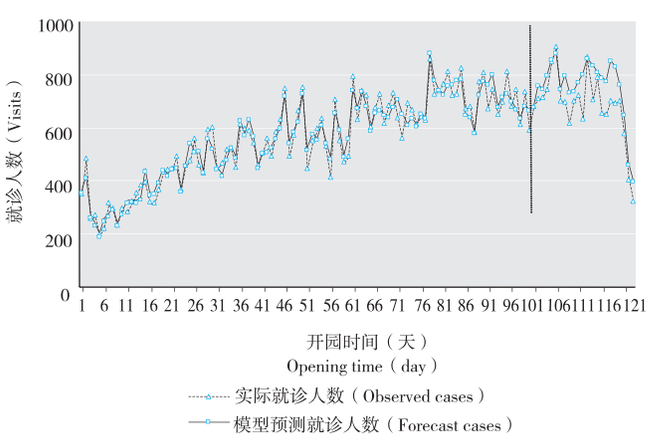

图2　模型预测就诊人数与实际就诊人数散点图

Figure 2　Scatter plot of cases observation and forecast values

3　讨论

广义相加模型是在广义线性模型和加性模型的基础上发展起来的，容许自变量函数为半参数形式的函数（样条函数），适用于处理应变量和众多解释变量间过度复杂非线性的关系[5]，近年在环境流行病学中的应用越来越多[5-9]。广义相加模型适用于多种分布资料的分析，模型中既可以包括参数拟合部分也可包括非参数拟合部分，甚至可以全部是非参数拟合。模型的构建灵活，并不拘泥于某一种形式的函数。当解释变量的个数较多或反应量与解释变量之间的关系不明确，反应变量的分布不易判定或不符合所要求的分布时均可考虑用广义相加模型。目前还未见在大型活动中开展预测研究的报道，本研究尝试对世博园区就诊人数进行预测，对大型活动公共卫生安全保障工作具有重要指导意义。

根据园区医疗站就诊监测的数据显示，园区就诊原因以肠道疾病、上呼吸道感染、中暑为主，慢性病急性发作也有发生。而气象因素是导致以上就诊原因较为直接的客观因素之一，入园人数也是一个重要因素，所以在模型拟合时选择入园人数、气象因素等纳入模型。以广义相加模型为基本统计模型，运用时间序列格式的资料，借助广义相加模型尽可能地控制混杂因素的影响，用哑变量的形式控制可能存在的双休日效应，在此基础上拟合气象因素、入园总人数与就诊人数的关系，预测世博园区内就诊人数。在GAM中，通常默认自由度为4，而本研究中计算出的各因素自由度均大于4，说明入园总人数、风速和气温对就诊人数的影响较为复杂。从模型拟合结果可以看出，世博园区就诊人数存在周末效应，周六就诊人数高于周一至周五工作时间的，这与周末入园人数相对较多有关；就诊人数随入园总人数的增加而增加；风速对就诊人数的影响趋势是随风速的增加，就诊人数先增加后减少；随着气温上升和日温差增加，就诊人数呈上升趋势，气温高中暑发生率增加；夏天人们广泛使用空调，空调空间内外的温差也易引发流感之类的呼吸系统疾病[10]，气温高，人们不舒适感增加，也将会导致各种疾病的发生；食物在高温下也会容易腐败变质，食物中毒和肠道疾病也较易发生。以上因素都会使就诊人数随之上升。

根据本研究模型验证结果，预测世博园区就诊人数和实际就诊人数较相近，平均相对误差为10.44%，上升与下降的变化趋势也基本一致，所以此模型可用于世博园区就诊人数的预测，广义相加模型作为大型公众活动预测研究是可行的。

参考文献：

[1] GESTELAND P H，GARDNER R M，TSUI F C，et al. Automated syndromic surveillance for the 2002 winter Olympics[J]. J Am Med Inform Assoc，2003，10(6)：547-554.

[2] HONG Y C，LEE J T. Effects of air pollutants on acute stroke mortality [J]. Environ Health Perspect，2002，110(2)：187.

[3] ABRAHAMOWICZ M，DU BERGER R，GROVER S A. Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality[J]. Am J Epidemiol，1997，145(8)：714.

[4] ROSSI G，VIGOTTIM A. Air pollution and cause-specific mortality in Milan，Italy，1980-1989[J]. Arch Environ Health，1999，54(3)：158.

[5] SIMON N，WOOD A，NICOLE H，et al. GAMs with integrated model selection using generalized regression splines and applications to environmental modeling[J]. Ecological Moedlling，2002，(157)：157-177.

[6] LAURENT F，ALAIN L T，ISABELLE B，et al. Difference in the relation between daily mortality and air pollution among elderly and all ages populations in south western France[J]. Environ Res，2004，(94)：249-253.

[7] FRANCESCA D，AIDAN M D，SCOTT L Z. On the use of GAM in Time Series studies of air pollution and health[J]. Am J Epidemiol，2002，156(3)：193-202.

[8] PATTENDEN S，NIKIFOROV B G，ARMSTRONG B. Mortality and temperature in Sofia and London[J]. J Epidemiol Community Health，2003，57(8)：628-633.

[9] BIBI H，NUTMAN A，SHOSEYOV D，et al. Prediction of emergency department visits for respiratory symptoms using an artificial neural network[J]. Chest，2002，122(5)：1627-1632.

[10] LIN W S，ZHENG S Y.The relationship between influenza peak and weather in Hong Kong[J]. J Environ Health，2004，21(6)：389-391.

（**收稿日期**：2010-11-24）

（英文编审：金克峙；编辑：洪琪；校对：丁瑾瑜）