·基础理论与方法·

采用广义可加模型分析预测上海市流感样病例发病情况

陈健　张磊　陆帅　姜晨彦　胡家瑜　姜庆五　吴凡

【导读】 探讨广义可加模型(GAM)在上海市流感样病例发病趋势分析和预测中应用，并通过已知的气象条件预测可能的流感样病例数量。研究中利用2006—2010年上海市每周气象数据以及流感样病例监测数据，按照GAM理论，建立气象数据与流感样病例间基于非线性回归的数学模型。通过初步的数据分析，构造多个候选模型，并通过AIC(Akaike information criterion)指标选取合适的模型进行数据分析及预测。基于周平均气温及周平均日温差(周相对湿度)的模型较好拟合了原始数据，且模型简洁、明确。对于原始数据的拟合残差及部分拟合残差基本符合上海市流感样病例发病的实际情况，并具有一定的预测能力。结果表明GAM能够较好拟合上海市流感样病例发病与气象因素的变化趋势，准确预测流感样病例的发病情况，适合应用于气象因素依赖的疾病发病预测和分析。

【关键词】 广义可加模型；流感样病例；预测

Prediction of influenza-like illness in Shanghai based on the generalized additive method

CHEN Jian[1,3], *ZHANG Lei*[2], *LU Shuai*[2], *JIANG Chen-yan*[1], *HU Jia-yu*[1], *JIANG Qing-wu*[3], *WU Fan*[1].
1 *Shanghai Municipal Center for Disease Control and Prevention*, *Shanghai* 200336, China; 2 *School of Mathematical Sciences*, 3 *School of Public Health*, *Key Laboratory of Public Health Safety in Ministry of Education*, *Fudan University*
Corresponding authors: *WU Fan*, Email: fwu@scdc.sh.cn; *JIANG Qing-wu*, Email: jiangqw@fudan.edu.cn

This work was supported by grants from the Shanghai Municipal Health Bureau (No. 2010188); *Shanghai Municipal Science and Technology Commission's Major Program of Biological Medicine* (No. 09DZ1906600); *the National Science Foundation of China* (No. 11101093); *Shanghai Pujiang Program* (No. 11PJ1400800) *and Shanghai Leading Talents Training Plan*, *Local Team 2010.*

【Introduction】 The aim of the current research topic was to test the generalized additive method (GAM), using data from the analysis and prediction on influenza-like illness (ILI) in Shanghai. Through collecting the meteorological data as well as the ILI from 2006 to 2010, we established several nonlinear regression candidate models based on the GAM. These models considered factors as: the nonlinear dependence on the meteorological data, i.e. weekly avcrage temperature and weekly average (maximum) temperature differences and the ILI. The AIC (Akaike information criterion) involved two simplified models which were implemented for further analysis and prediction. Finally, numerical examples showed that the proposed models could shed light on the connection between the meteorological data and the ILI. GAM could be used to fit the frequencies of ILI and meteorological factors in Shanghai. The proposed models were able to accurately analyze the onset of ILI, implying that GAM might be suitable for the prediction and analysis of those meteorological correlative diseases.

【Key words】 Generalized additive method; Influenza-like illness; Prediction

DOI: 10.3760/cma.j.issn.0254-6450.2013.04.022

基金项目：上海市卫生局课题(2010188)；上海市科委生物医药重大专项(09DZ1906600)；国家自然科学青年科学基金(11101093)；上海市浦江人才计划(11PJ1400800)；2010年上海领军人才"地方队"培养计划

作者单位：200336 上海市疾病预防控制中心(陈健、姜晨彦、胡家瑜、吴凡)；复旦大学数学科学学院(张磊、陆帅)，公共卫生学院公共卫生安全教育部重点实验室(陈健、姜庆五)

通信作者：吴凡，Email：fwu@scdc.sh.cn；姜庆五，Email：jiangqw@fudan.edu.cn

流感具有明显的季节性特征。分析上海市历年流感样病例(ILI)监测数据,发现在季节交替阶段,如冬春、夏秋时节,ILI数量明显增多,推测天气因素对其发生有重要影响。为此本研究对上海市连续5年(2006—2010)气象数据以及监测哨点ILI例数建立广义回归模型(GAM),分析气象因素与ILI发病间的联系,并试图找出气象条件中哪些因素对发病起到关键作用,以数学模型阐述ILI发病对天气状况的依赖,开展ILI的预测与分析。

基本原理

1. GAM的基本数学原理:在统计分析的回归模型中,线性回归模型是最简单也是最常用的方法之一。例如,设Y为响应变量,$X=(X_1,X_2,\cdots,X_p)$为选定的解释变量,故线性回归模型可表示为

$$E(Y\mid X_1,X_2,\cdots,X_p)=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p \quad (1)$$

式中$E(Y\mid X_1,X_2,\cdots,X_p)$可看成是反应变量$Y$依赖于解释变量$X_1,X_2,\cdots,X_p$的一个平均预测值。

而在很多实际问题中,Y的条件数学期望和X不是简单线性关系。此时作为线性回归模型的推广,广义线性回归模型(GLM)就用来拟合这类非线性的数据。在GLM中,引入了连接函数(link function),记为$g(\mu)$,其中$\mu=E(Y\mid X_1,X_2,\cdots,X_p)$。于是GLM就有以下形式

$$g(\mu)=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p \quad (2)$$

如将式(2)中线性形式的β_jX_j均替换成非线性函数$f_j(X_j)$,那么得到的即是GAM

$$g(\mu)=\alpha+f_1(X_1)+f_2(X_2)+\cdots+f_p(X_p) \quad (3)$$

式中$f_j(X_j)$为满足一定条件的特定(非线性)函数且$Ef_j(X_j)=0$。

2. GAM的建立及求解:如确定使用GAM进行数据分析,就应建立该模型的相关形式。一般来说,模型设定的过程中,首先需要根据反应变量Y的类型选择相应的连接函数$g(\mu)$。譬如,满足Poisson分布的反应变量一般选取对数函数来作为连接函数,即$g(\mu)=\ln(\mu)$。

选取一个候选模型后,即式(3)中p的值确定后,就需要基于给定的样本数据通过参数或者非参数方法来确定式(3)中解释变量X_j所满足的函数$f_j(X_j)$。通常采用局部加权平滑法(LOESS smoothers)或平滑样条法(smoothing splines)等非参数方法得到。在后种方法中,利用带惩罚项的最小二乘法估计函数$f_j(X_j)$,在数学上的表达式为

$$\min\Big(\sum_{i=1}^{n}\omega_i[g(y_i)-a-\sum_{j=1}^{p}f_j(x_{ij})]^2+\sum_{j=1}^{p}\lambda_j\int_a^b[f_j^{(2)}(t)]^2dt\Big) \quad (4)$$

式中$(X_1,X_2,\cdots,X_p)=(x_{i1},x_{i2},\cdots,x_{ip})_{i=1}^n$;$Y=(y_i)_{i=1}^n$为给定的样本数据,以$X_j=(x_{ij})_{i=1}^n$表示第$j$个解释变量的样本数据。式(4)中另外出现的参数$\omega_i$为权重系数,公式第二项一般称为惩罚项,其中$f_j^{(2)}$为$f_j$的二阶导数,$\lambda_j$为对应于$f_j$的平滑参数[1],用以控制所估计$f_j$的平滑程度。若$f_j^{(2)}$越小,则$f_j$越接近线性函数,从而更加平滑。

在数学和统计学上,通过local scoring算法[2,3]获得式(4)中的函数f_j,且理论上式(4)的解是惟一的,该解可以由一个以x_{ij}为节点的自然三次样条函数表示[4]。自然三次样条函数是分段多项式函数,满足整体二阶连续,且在边界点满足二阶导数为零。在拟合自然三次样条函数时,仅需要反演其在每个子区域的4个多项式系数即可。关于平滑参数λ_j的选择,可采用交叉校验(cross-validation)方法[4]。本文将利用统计软件S-PLUS 8.0以实现这些函数[5]。

3. GAM的拟合检验:对于多个解释变量的样本数据,可选取不同解释变量组合成多个GAM,即式(3)中选择不同的p值。这就涉及了等同模型的概念,也即最后可能出现一个或者多个可选的模型。通过设定多个可选模型可以避免单一模型出现拟合错误、数据处理失败的弊端,从而提高数据处理的精度和速度。

一般来说,实际问题中可以选择人工交互法、AIC指标(Akaike information criterion)、TURBO回归样条等方法来甄别候选模型的拟合优度。这些方法通常依赖于候选模型的残差讯息,即$\mathrm{Re}=\sum_{i=1}^{n}(y_i-\mu_i)^2$,其中$\mu_i$为基于已确定函数式(3)的拟合值。以AIC指标为例,该指标可描述回归模型对原始数据拟合的好坏,且是相对的。通过分别计算一系列候选模型的AIC值,并对候选模型排序,即AIC值较小的模型相对较稳定。通常AIC值可由以下公式计算

$$\mathrm{AIC}=2p+n\ln(\mathrm{Re}/n)$$

式中p为统计模型中的参数个数,Re为残差值。本文将计算候选模型的AIC值,并选择该值较小的模型作为最终GAM。

实例分析

1. 资料来源：用于预测ILI发病的资料来自2006年1月至2010年12月上海市流感监测网络每日ILI监测数据，是由各监测哨点医院根据《全国流感监测方案》每日登记门、急诊就诊的ILI例数并上报《中国流感监测信息系统》，所有数据均经市、区（县）和监测哨点专职负责人审核，并定期对监测数据的准确性、完整性和及时性进行质控。气象数据源自上海市中心气象台每日气象监测数据，主要包括日平均气温、日最高气温、日最低气温、日平均相对湿度和降水量等。

2. 数据处理：

（1）数据的预分析：为了解气象因素对ILI的影响，以上海地区各气象要素为解释变量，ILI例数为反应变量，通过GAM了解这些气象要素对ILI例数的影响并做出较为准确的预测。考虑到流感一般潜伏期为2～4 d，最长7 d，另外症状出现后至就诊也有1～2 d的时差，即某天的天气情况对当天的ILI例数几无影响，但可影响到若干天后的例数。同时考虑到数据的可及性，因此本研究将以周为单位进行统计分析，即找出周平均气象数据与周ILI总例数间的关系。

首先需要确定候选GAM的连接函数$g(\mu)$。由于回归模型中，响应变量Y是周ILI例数。事实上，Y是在某段时间内对某一事件的计数，因此可以假设Y服从Poisson随机分布[6]，因此建立的模型就是Poisson型的回归广义可加模型，即式（3）可变换为模型（m1）

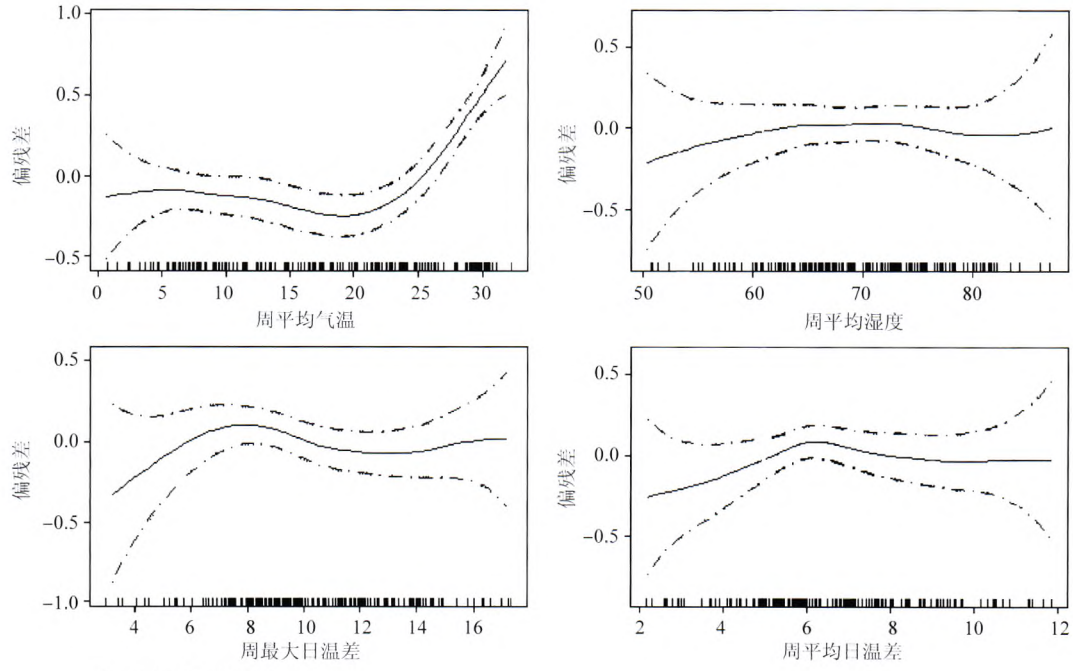
$$\ln\mu = a + \sum_{j=1}^{p} f_j(X_j)$$

由此可见GAM的连接函数为$\ln(\mu)$。模型（m1）解释变量X_j的选择主要基于给定的样本数据，包括日平均气温、日最高气温、日最低气温、日平均相对湿度、降水量等气象因素。本研究从中提出4个可能解释变量X_j分别为周一至周日的周平均气温、周平均相对湿度、周平均温差、周最大温差，并假设这些数据是独立的随机变量，并将这4个气象要素的解释变量分别记为X_1、X_2、X_3和X_4。

（2）候选模型的设定及评价：为进一步了解上述4个气象要素对ILI例数的影响，分别建立4个单解释变量的GAM，即模型（m2）

$$\ln\mu = a + f_j(X_j),\ j=1,2,3,4$$

通过观察Partial Residual拟合曲线（图1），发现周平均气温（X_1）对ILI例数的影响最为显著，而周平均相对湿度（X_2）、周平均日温差（X_3）和周最大日温差（X_4）3个要素对ILI例数的影响程度较为类似，由于（X_3, X_4）均为日温差衍生数据，可简化模型，即在设置候选模型中不同时引入（X_3, X_4）。为此可考虑模型（m3～m6），分别为$\ln\mu = a + f_1(X_1) + f_2(X_2) + f_3(X_3)$，$\ln\mu = a + f_1(X_1) + f_3(X_3)$，$\ln\mu = a + f_1(X_1) + f_2(X_2)$，$\ln\mu = a + f_1(X_1) + f_4(X_4)$。



注：虚线为置信区域

图1　模型（m2）的Partial Residuals拟合曲线

通过计算模型(m3～m6)的AIC值(分别为59 352、59 334、59 335和57 976),先选择其中AIC值最小的模型(m6),即理论上模型(m6)是4个模型中对原始数据拟合最好,其次模型(m4)和(m5)的AIC值几乎相等,由于模型(m4)的解释变量与已选择出的模型(m6)的解释变量[周平均日温差(X_3)与周最大日温差(X_4)]类似,故选择包括周平均相对湿度(X_2)作为解释变量的模型(m5)较为合适。至此已确定最终后的GAM为模型(m5)和(m6)。

(3)数据拟合及模型估计:首先给出模型(m5)的拟合残差(图2),可见该模型对数据的拟合效果较好,大量样本数据的残差点聚集在0值附近,其中第204～206点偏离整体残差点较远且残差值较大,怀疑其为异常点。而从模型(m5)的Partial Residual拟合曲线(图3)可见,当平均气温<10 ℃和>25 ℃时,ILI数量增长明显,其中后者比前者增长更快。该结果基本符合上海市冬春之交和夏秋之交ILI发病明显高于其他时间的现象。另外图3也显示,由于周平均相对湿度对ILI例数的影响,当相对湿度在50%～70%时,ILI例数较少。

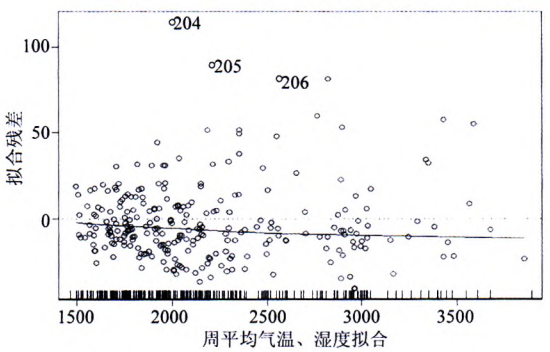


图2 模型(m5)的Deviance Residual

同样也给出模型(m6)的拟合残差(图4)和Partial Residual拟合曲线(图5)。图4中第29～31点偏离整体点较远,怀疑其为异常点。关于周平均气温对ILI例数的影响,模型(m6)的拟合结果基本上和模型(m5)相同。而关于周最大日温差的影响,发现当在8 ℃左右或>12 ℃时,ILI例数增长也很明显。上述的气温和温差数据恰好对应了上海市冬春之交和夏秋之交的气候情况。

建立了GAM,并采用统计软件S-PLUS 8.0得到其拟合结果,就可用所建立的模型根据未来一周的天气预测ILI数量。图6是根据模型(m6)的拟合结果在不同周平均气温和周最大日温差的情况下,预测周ILI数量。可见当周平均气温<25 ℃、且周最大日温差较小时,ILI例数较少;而在极端情况下,即

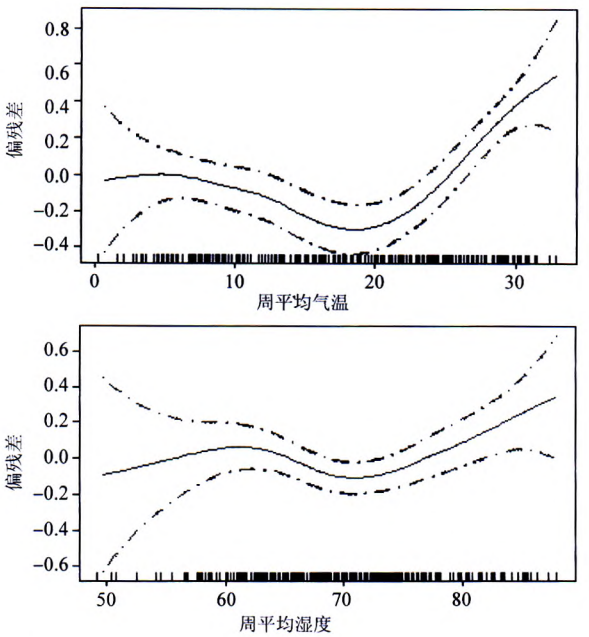


图3 模型(m5)的Partial Residuals拟合曲线

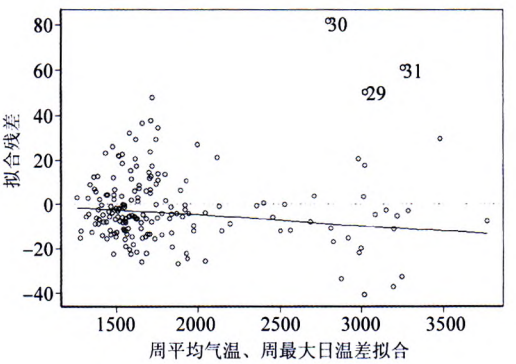


图4 模型(m6)的Deviance Residuals

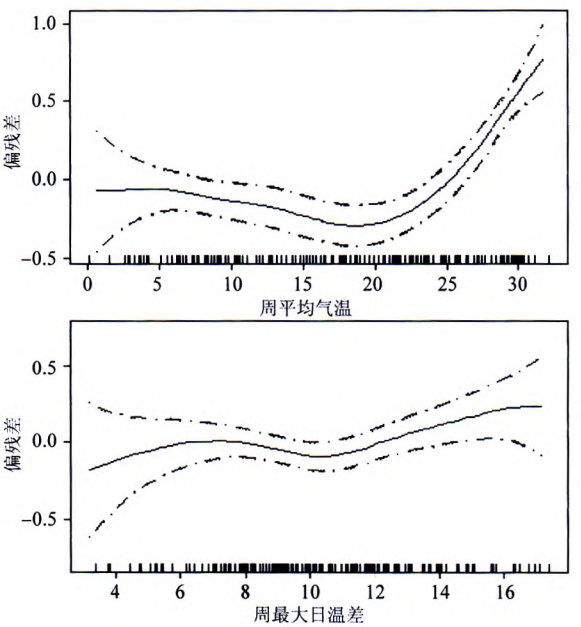


图5 模型(m6)的Partial Residuals拟合曲线

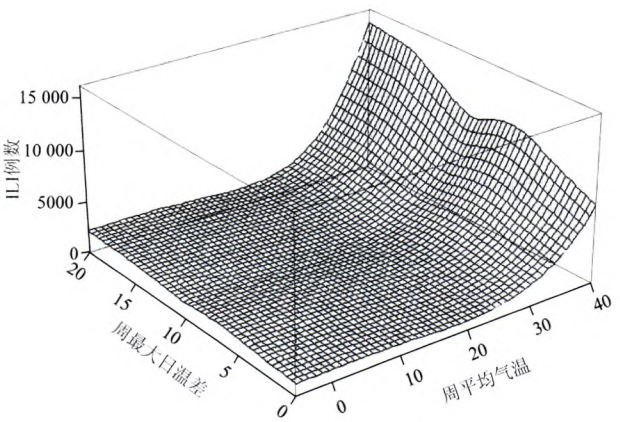


图6　基于模型(m6)预测ILI数量

周平均气温>35 ℃、且周最大日温差较大时，将出现大量的ILI。

讨　论

本研究通过设定若干气象要素为解释变量，建立预测ILI例数的数学模型，评估2006—2010年上海市气候及ILI数据。GAM可预测未来短期内ILI例数，一旦周平均气温和日平均温差超过相应的警戒值，卫生部门通过发布流感指数等方法进行及时的预警并采取针对性干预措施，对流感的防控具有参考价值。

GAM是一个简洁描述多元非线性回归的方法，适用于流感等有复杂传播过程的传染病研究。由于其需要拟合的非线性函数f_i是一个仅依赖于单变量的函数，一般而言需要对样本数据加以预分析，总结出一组线性无关的解释变量，从而构造候选模型并加以检验甄选。这样可简化统计模型和有效降低模型的方差。而关于单变量函数的拟合已有很多成熟的方法，无论是参数性还是非参数性的方法，均能很好的融合到模型中。

利用GAM开展对疾病的预测也存在局限性。流感的发生和流行还受到病毒活动强度、一般人群抗体水平、病毒变异程度、人群的生活和行为方式、空气污染(如$PM_{2.5}$[7])等多种因素的综合作用，本研究仅考虑气象因素，模型的精确性还有待进一步提高。

参 考 文 献

[1] Engl HW, Hanke M, Neubauer A. Regularization of inverse problems. Kluwer Academic Publishers, Dordrecht, 1996.

[2] Hastie TJ, Tibshirani RJ. Generalized additive models. Chapman and Hall, London, 1990.

[3] Hastie TJ, Tibshirani RJ. Generalized additive models: some applications. J Am Stat Assoc, 1987, 82(398): 371-386.

[4] Wahba G. Spline models for observational data. SIAM, Phildelphia, 1990.

[5] Feng GS, Chen JW. Generalized additive model and its SAS program. Chin J Health Stat, 2007, 24(1): 82-84. (in Chinese)
冯国双，陈景武. 广义可加模型及其SAS程序实现. 中国卫生统计，2007，24(1)：82-84.

[6] Venables WN, Ripley BD. Modern applied statistics with S. Springer, Berlin, 2002.

[7] Dominici F, McDermott A, Zeger SL, et al. On the use of generalized additive models in time-series studies of air pollution and health. Am J Epidemiol, 2002, 156(3): 193-203.

(收稿日期：2012-09-24)
(本文编辑：张林东)

读者·作者·编者

关于中华医学会系列杂志投稿网址的声明

为维护广大读者和作者的权益以及中华医学会系列杂志的声誉，防止非法网站假冒我方网站诱导作者投稿、并通过骗取相关费用非法获利，现将中华医学系列杂志稿件管理系统网址公布如下，请广大作者加以甄别。

1. “稿件远程管理系统”网址：中华医学会网站(http://www.cma.org.cn)首页的“业务中心”栏目、中华医学会杂志社网站(http://www.medline.org.cn)首页的“稿件远程管理系统”以及各中华医学会系列杂志官方网站接受投稿。作者可随时查阅到稿件处理情况。

2. 编辑部信息获取：登录中华医学会杂志社网站(http://www.medline.org.cn)首页，在《中华医学会系列杂志一览表》中可查阅系列杂志名称、编辑部地址、联系电话等信息。

3. 费用支付：中华医学会系列杂志视杂志具体情况，按照有关规定，酌情收取稿件处理费和版面费。稿件处理费作者在投稿时支付；版面费为该稿件通过专家审稿并决定刊用后才收取。

欢迎投稿，并与编辑部联系。特此声明。

中华医学会杂志社