·论著·方法学研究·

环境流行病学研究中广义可加模型在Python中的实现

李湘莹，李培政，王 静，罗晨曦，张清煜，马 露

武汉大学公共卫生学院（武汉 430071）

【摘要】目的 探索环境流行病学领域中常见的时间序列资料，利用 Python 以及其他常用统计软件实现广义可加模型（generalized additive models，GAM）的建模，比较各软件建模过程和结果的异同。**方法** 以研究某地 $PM_{2.5}$ 暴露与呼吸系统疾病入院人次之间的关系作为实例，分别利用 Python 软件 statsmodles 库、R 软件 mgcv 库和 SAS 软件的 proc gam 语法，构建 GAM 模型，比较各软件命令代码、参数设置以及参数估计的差异。**结果** 三种软件构建 GAM 模型的建模逻辑相似，但在内置函数拟合过程、命令代码以及可调用的样条函数等方面有所差别，各软件输出结果基本一致。**结论** Python 软件利用第三方库可实现广义可加模型的构建，为进一步拓展其在流行病学领域的应用提供了参考。

【关键词】Python；广义可加模型；时间序列数据；环境流行病学

Implementation of generalized additive models in environmental epidemiology research in Python

Xiang-Ying LI, Pei-Zheng LI, Jing WANG, Chen-Xi LUO, Qing-Yu ZHANG, Lu MA

School of Public Health, Wuhan University, Wuhan 430071, China
Corresponding author: Lu MA, Email: malu@whu.edu.cn

【Abstract】Objective To explore the common time series data in the field of environmental epidemiology, using Python and other statistical softwares to realize the modeling of generalized additive models (GAM), and to compare the similarities and differences of the modeling process and results of each software. **Method** A study of the relationship between $PM_{2.5}$ and the number of hospital admissions of respiratory diseases was taken as an example. Python software used statsmodles library, R software used mgcv library, SAS software used proc gam syntax to build GAM models, and the differences in codes, parameter settings, and parameter estimates were compared. **Results** The modeling logic of 3 programs is similar, but there are differences in the built-in function fitting process, code using and callable spline function. The outputs are basically consistent. **Conclusion** Python software can build GAM by using third-party libraries. It provides a reference for further expanding its application in the field of epidemiological scientific research.

【Keywords】Python; Generalized additive models; Time series data; Environmental epidemiology

DOI: 10.12173/j.issn.1004-4337.202302032
基金项目：湖北省卫生健康委 2021—2022 年度科研项目（WJ2021F103）
通信作者：马露，博士，副教授，硕士研究生导师，Email：malu@whu.edu.cn

环境流行病学领域研究工作中往往涉及大量时间序列数据，其影响因素的多样性和复杂性使得在研究时往往难以确定回归关系的基本形式，广义可加模型（generalized additive models，GAM）是解决这一问题的一种途径。GAM 是在广义线性模型（generalized linear mode，GLM）基础上发展而来的，它提供了更为灵活的建模框架，其反应变量可以为非正态的指数族分布，解释变量与反应变量间可存在复杂的非线性关系，由指定的协变量平滑函数加上线性项的常规参数分量之和得出。伴随环境流行病学领域研究资料的数字化高速发展，多源数据融合、智能化管理和分析已成为必然趋势。Python 作为一个拥有强大第三方库的开源软件，其规范且相对简洁的编程语言，在同应用程序整合、数据源连接与读取、调用其他语言，以及实现人工智能等方面展现出了显著优势[1]。然而，在环境流行病学领域中，使用 Python 进行 GAM 分析的研究尚较少。本文利用环境流行病学研究中的一组时间序列资料，分别进行 Python、R 和 SAS 的 GAM 建模，比较它们在运算逻辑、参数、结果方面的差异，以拓宽 Python 软件在环境流行病学领域的应用场景。

1 资料与方法

1.1 资料来源

因呼吸系统疾病入院人次数据主要来源于某地区卫生信息中心 2017 年 1 月 1 日至 2018 年 12 月 31 日记录的该地医院病案首页。每日大气中 $PM_{2.5}$ 浓度、日均温度、平均湿度等数据来源于当地环境监测部门官方网站。$PM_{2.5}$ 浓度和研究对象日入院人次逐日变化图如图 1 所示。

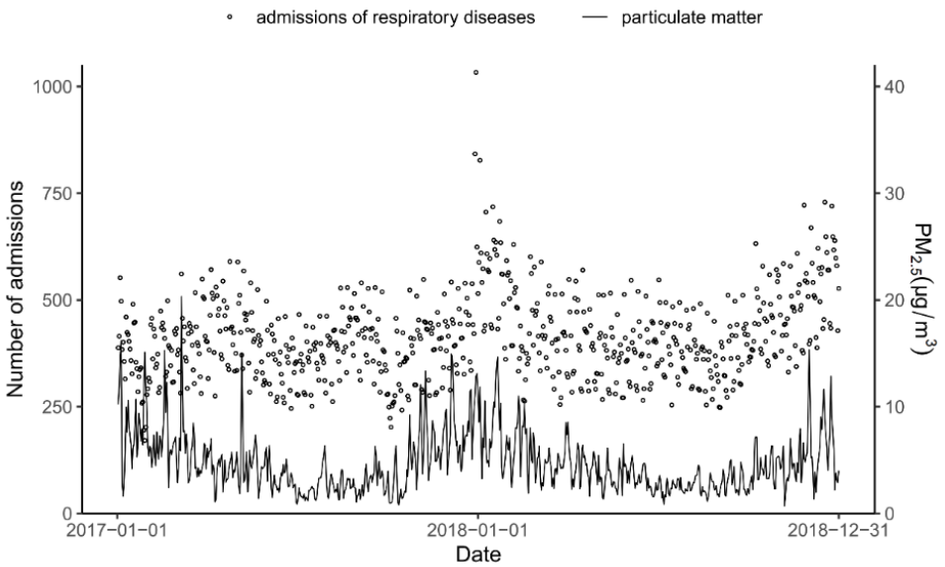

图1 2017—2018年某地$PM_{2.5}$浓度和研究对象日入院人次逐日变化图

Figure 1. Daily variation of $PM_{2.5}$ concentration and the number of hospital admissions of research objects from 2017 to 2018

1.2 模型构建

为了评价大气污染物与因呼吸系统疾病入院人次的关系，需要对长期趋势和季节趋势、星期效应、平均温度、相对湿度等产生的影响进行控制。此处长期趋势指连续数年观察中入院人数呈现的某种总的变动趋势；季节趋势指入院人数在每年随季节产生的周期变化。这里，我们对数据库进行了调整，为了控制长期趋势和季节趋势，将事件发生的日期调整为这一事件发生的先后次序，引入时间序列变量（命名为 time），记作 1，2，3…。为了控制星期效应的影响引入星期变量（命名为 dow），工作日（星期一至星期五）入院为 dow=0、周末（周六至周日）入院为 dow=1。平均温度的单位为℃，相对湿度的单位为 %，$PM_{2.5}$ 浓度单位为 μg/m^3，数据形式如表 1（仅展示部分数据）。

本研究首先在入院前三天的单日和移动平均暴露的多个滞后结构下估计了空气污染物对全因住院的短期影响 (lag0–3, lag01–03)。然后根据模型的广义交叉验证（generalized cross-validation，

表1 颗粒物、气象及研究对象数据资料

Table 1. Data of particulate matter, meteorology and research objects

Date	时间序列	入院人数	星期变量	平均温度	相对湿度	$PM_{2.5}$
2017/1/1	1	388	1	9.5	76	10.225
2017/1/2	2	415	0	9.2	79	11.934
2017/1/3	3	552	0	9.8	81	14.838
2017/1/4	4	497	0	10.9	93	16.185
2017/1/5	5	384	0	8.8	98	2.848
2017/1/6	6	401	0	7.7	98	1.596
2017/1/7	7	356	1	6.6	89	2.690
2017/1/8	8	315	1	5.9	88	5.828
2017/1/9	9	460	0	5.1	86	9.975
2017/1/10	10	424	0	6.5	75	7.606

GCV）和 R^2，最终选择以 $PM_{2.5}$ 单污染物滞后 0 天的暴露模型作为先验滞后结构，利用 Python 中的 statsmodels 库构建 GAM 模型，并与 R 软件、SAS 软件的输出结果进行比较，验证 Python 结果。

GAM 的基本形式为：

$$g(\mu) = s_1(X_1) + s_2(X_2) + \cdots + s_p(X_p) + \varepsilon$$

$$\mu = E(Y|X_1, X_2, \cdots, X_p)$$

其中，$g(\mu)$ 为连接函数，$g(\mu)$ 的选择由反应变量的分布形式决定：①当反应变量服从正态分布时，等同于一般可加模型，$g(\mu)=\mu$。②当反应变量服从二项分布时，选用 logit 连接。③当反应变量服从 Gamma 分布时，选用 Identity 连接[2]。④当反应变量服从 Poisson 分布时，选用 Log 连接[3]。$Si(\cdot)$ 为非参数光滑样条函数，ε 为误差项。模型构建过程如下：

（1）调用相关参数并载入数据

加载 statsmodles 库与 pandas 库，从 statmodels 库中加载 GLMGam 语句，CyclicCubicSplines 语句，以便于读取数据和拟合模型。使用 pd.read 语句载入数据保存在 base_data 中。命令如下：

```
import statsmodels.api as sm
import pandas as pd
from statsmodels.gam.api import GLMGam, CyclicCubicSplines
base_data=pd.read_csv(r"D:data.csv")
```

（2）对解释变量进行样条变换

在 Python statsmodles 库中，使用 GLMGam 函数进行拟合。分类变量被视为线性项，对非线性解释变量通过样条函数进行拟合。GLMGam 函数中主要选择的平滑基函数有两类，分别是 B 样条函数与循环立方样条函数。在本例中我们选用循环立方样条函数对长期趋势、温度、湿度这三个变量进行样条变换。命令如下：

```
x_spline=base_data[['time','wendu','shidu']]
bs=CyclicCubicSplines(x_spline, df)
```

（3）确定模型自由度

自由度很大程度上会影响模型拟合和预测的准确性，目前主要有四种确定模型自由度的方法：①根据既往学者的经验设置固定的自由度；②根据最小化模型残差自相关绝对值最小来选择；③广义交叉验证（generalized cross-validation，GCV）依据 GCV 值最小选择；④赤池信息准则（Akaike information criterion，AIC），依据 AIC 值最小选择[4]。在实际操作中对于最优自由度需要进行多次评估，本例依照既往学者的经验并结合 AIC 值、GCV 值，设置固定自由度，其中长期趋势自由度设为 15/ 年，平均温度自由度设为 4，相对湿度自由度设为 4。命令如下：

```
bs=CyclicCubicSplines(x_spline, df=[15*2,4,4])
```

（4）构建基础模型

由于因呼吸疾病系统入院人次服从 Poisson 分布，故连接函数选择 Log 连接，根据 GAM 函数定义，构建如下模型：

$$Log(\mu_t)=\beta_1 PM_{2.5}+\beta_2 dow+s(time,df=30)+s(wendu,df=4)+s(shidu,df=4)$$

其中 β_1、β_2 分别为 $PM_{2.5}$ 和短期波动（dow）的系数，s 为非参数平滑函数，ε 为残差项。

使用 GLMgam 函数构建模型，对三个非线性解释变量使用循环立方样条函数拟合，定义反应变量为泊松分布，使用 fit 语句拟合函数，summary 语句查看结果。具体命令如下：

```
gam_bs=GLMGam.from_formula(
'y ~ p2.5+ dow', data=base_data,smoother=bs,
family=sm.families.Poisson())
res_bs=gam_bs.fit()
print(res_bs.summary())
```

1.3 模型验证

采用 R 3.6.1 软件的 mgcv 包与 SAS 9.2 中的 proc gam 语句构建 GAM 模型作为参照，由于 Python 与 SAS 软件中可选择样条函数的限制，为保证结果的可比较性，本文选用 R、Python 软件中的循环立方样条（cyclic cubic splines）函数及 SAS 软件中的三次平滑样条函数对非参数项进行拟合，并设置统一且固定的自由度，验证输出结果是否正确。R、SAS 相关代码见框 1。

R	SAS
library(mgcv)	proc gam
base_data <– read.csv（“D:data.csv”）	data= base_data;
model<–gam(y~ p2.5+s(time,	model y=
df=30)+as.factor(dow)+	param(p0)
s(wendu,df=4)+s(shidu,df=4),	param(dow)
family=poisson,data=base_data)	spline(time,df=30)
summary(model)	spline(wendu,df=4)
	spline(shidu,df=4) /dist=POISSON; run;

框1 R、SAS中的GAM代码

Box 1. GAM codes in R and SAS

2 结果

2.1 GAM建模及参数估计结果

Python、R、SAS 三种软件输出结果与指数转换结果如表 2。Python 与 R 输出的结果主要包括：偏回归系数、标准误、*Z* 值、*P* 值等。SAS 除上述结果外还会输出迭代汇总、平滑模型分析及图例。

表2 Python、R、SAS参数估计结果

Table 2. Results of parameter estimation in Python, R and SAS

软件	变量	估计值	标准误	OR值（95%CI）	*Z*/*t*值	*P*值
Python						
	截距	3.937	0.003	–	1298.86[a]	<0.01
	$PM_{2.5}$	0.087	0.009	1.091（1.072, 1.110）	9.29[a]	<0.01
	星期变量	–0.299	0.004	0.742（0.736, 0.747）	–68.63[a]	<0.01
R						
	截距	6.047	0.004	–	1273.85[a]	<0.01
	$PM_{2.5}$	0.088	0.009	1.091（1.073, 1.111）	9.62[a]	<0.01
	星期变量	–0.298	0.004	0.742（0.736, 0.748）	–68.63[a]	<0.01
SAS						
	截距	5.889	0.017	–	348.00[b]	<0.01
	$PM_{2.5}$	0.087	0.007	1.091（1.076, 1.106）	11.70[b]	<0.01
	星期变量	–0.301	0.004	0.740（0.734, 0.746）	–69.27[b]	<0.01

注：[a]*Z*值；[b]*t*值

Python 输出的截距、$PM_{2.5}$、星期变量参数估计结果分别为：ε=3.937，α=0.003，Z=1298.86，P < 0.01；β_1=0.087，α=0.009，Z=9.29，P < 0.01；β_2=–0.299，α=0.004，Z=–68.63，P < 0.01。

指数转换后的结果说明，在控制了长期趋势和季节趋势、星期效应、平均温度、相对湿度的影响后，每日因呼吸系统疾病入院人次与 $PM_{2.5}$ 浓度变化有关联。三种软件输出的污染物及星期变量的估计值和标准误结果有所不同，但差异较小。Python 参数估计结果显示，$PM_{2.5}$ 每升高 100 μg/m^3，入院人次增加 9.1%[95%CI（7.2%，11.0%）]。R 输出结果为 $PM_{2.5}$ 每升高 100 μg/m^3，入院人次增加 9.1%[95%CI（7.3%，11.1%）]，SAS 输出结果为 $PM_{2.5}$ 每升高 100 μg/m^3，入院人

次增加 9.1%[95%CI（7.6%，10.6%）]。

在模型拟合结果中，R 软件输出 R^2=0.659，Python 输出 Pseudo R^2=1.000，但 SAS 中仅针对各非参数项输出 GCV 值，GCV 值越小表明拟合效果越好，本例中各非参数项 GCV 值均较小。

3 讨论

本文结合环境流行病学研究实例，简要介绍了 GAM 模型的基本形式，并比较 Python、R 和 SAS 在建模和分析结果上的异同。结果发现三种软件的输出结果基本一致，Python 输出结果可信。

在 GAM 模型构建的参数设置上，三种软件存在差异。第一，三者内置函数拟合过程不同。Python 软件与 R 软件是通过惩罚似然最大化来拟合平滑参数，通过对每个平滑参数添加惩罚项来避免拟合过度或拟合不良[5]，而 SAS 软件是通过双重迭代方法估计平滑参数[6]。第二，三者构建模型命令代码不同。Python 软件先统一对非线性参数项使用同一样条函数，再将非线性项与线性项通过 GLMGam 函数构建模型进行拟合。R 软件与 SAS 软件均通过相应函数直接将线性项与非线性项放在同一语句内构建模型。第三，三者可选用的样条函数不同。Python 软件目前已得到验证的样条函数仅有 B 样条函数与循环立方样条函数，SAS 软件中只可以使用三次平滑样条函数、局部回归与薄板样条函数[7]，R 软件则能在这两种软件的基础上提供如三次自然样条函数等更多可供选择的平滑函数[8]。这些在参数设置和选择上的不同可能是其结果出现差异的主要原因。

尽管在本研究中，运用 Python 进行 GAM 建模无论是从程序语言上还是参数设置上，都没有展现出其在统计分析方面的优势，但 Python 作为当前主流的计算机语言之一，其在网络爬虫、数据可视化、机器学习等方面拥有丰富的第三方库资源支持，能够高效实施多源数据连接与管理，并最终使实现数据收集与处理的环境统一成为可能。不断拓展 Python 在环境流行病学领域的应用，必将为促进计算机与医学相关领域交叉融合发展提供借鉴和参考。

参考文献

1 Mckinney W. Data structures for statistical computing in Python[J]. Proc of the 9th Python in science Conf, 2010, 56–61. https://www.researchgate.net/publication/265001241_Data_Structures_for_Statistical_Computing_in_Python.

2 Wood SN. mgcv: GAMs and generalized ridge regression for R. R news, 2001,1(2), 20–25. https://xueshu.baidu.com/usercenter/paper/show?paperid=f73e92306ef7bae5280d10b61de86d8f&site=xueshu_se&hitarticle=1.

3 李莉莉，张璇，杜梅慧．基于广义可加模型的 $PM_{2.5}$ 预测研究 [J]. 数理统计与管理，2020, 39(5): 811–823. [Li LL, Zhang X, Du HH. Research on $PM_{2.5}$ prediction based on generalized additive model[J]. Jouranl of Applied Statistics and Management, 2020, 39(5): 811–823.] DOI: 10.13860/j.cnki.sltj.20200818–005.

4 张云权，朱耀辉，李存禄，等．广义相加模型在 R 软件中的实现 [J]. 中国卫生统计，2015, 32(6): 1073–1075. [Zhang YQ, Zhu YH, Li CL, et al. Generalized additional model in R[J]. Chinese Journal of Health Statistics, 2015, 32(6): 1073–1075.] DOI: CNKI:SUN:ZGWT.0.2015–06–053.

5 White LF, Jiang W, Ma Y, et al. Tutorial in Biostatistics: the use of generalized additive models to evaluate alcohol consumption as an exposure variable[J]. Drug Alcohol Depend, 2020, 209: 107944. DOI: 10.1016/j.drugalcdep.2020.107944.

6 Cai W, Inc SI, Cary NC. Fitting generalized additive models with the GAM procedure in SAS 9.2[J]. 1287–1312 World Scientific Review, 2008. DOI: 10.1080/00401706.1997.10485461.

7 Xiang D. Fitting generalized additive models with the GAM procedure[J]. Sugi Proceedings, 2001, 256–326. https://www.researchgate.net/publication/228390997_Fitting_generalized_additive_models_with_the_GAM_procedure.

8 Wood SN. Stable and efficient multiple smoothing parameter estimation for generalized additive models[J]. Journal of the American Statistical Association, 2004, 99(467): 673–686. DOI: 10.1198/016214504000000980.

收稿日期：2023 年 02 月 08 日 修回日期：2023 年 02 月 26 日
本文编辑：李 阳 黄 笛

引用本文：李湘莹, 李培政, 王静, 等. 环境流行病学研究中广义可加模型在Python中的实现[J]. 数理医药学杂志, 2023, 36(4): 241–245. DOI: 10.12173/j.issn.1004–4337.202302032
Li XY, Li PZ, Wang J, et al. Implementation of generalized additive models in environmental epidemiology research in Python[J]. Journal of Mathematical Medicine, 2023, 36(3): 241–245. DOI: 10.12173/j.issn.1004–4337.202302032