DOI:10.13878/j.cnki.jnuist.2023.01.004

王静[1]　曹春正[1]

基于贝叶斯分层自回归时空模型的北京 $PM_{2.5}$ 预测

摘要

为解决 $PM_{2.5}$ 的多站点同步预测问题,提出一种贝叶斯框架下的分层自回归时空模型.将 $PM_{2.5}$ 日均浓度真实值视为潜在时空过程,利用一阶自回归过程刻画时间相关性,并基于 Matérn 过程捕获空间相关性,极大程度地提高了降维和同步预测的效率.此外,还将日最高温度、相对湿度和风速等气象因素作为解释变量,用于提升 $PM_{2.5}$ 的预测效果.借助模型的分层结构,通过贝叶斯方法结合马尔可夫链蒙特卡罗(MCMC)算法实现参数估计和预测过程.对北京市日均 $PM_{2.5}$ 浓度的实证分析表明,模型在空间和时间维度上均有良好的插值或预测效果.

关键词

贝叶斯;分层模型;自回归;时空模型;$PM_{2.5}$ 预测;马尔可夫链蒙特卡罗(MCMC)

中图分类号 X513

文献标志码 A

0　引言

$PM_{2.5}$ 作为主要空气污染物之一,由于粒径小,可以直接被人体吸入,且在大气中的停留时间长、输送距离远,因而对人体健康和大气环境质量的影响很大.医学研究表明,过高浓度的 $PM_{2.5}$ 不仅会导致心肺疾病发病率和死亡率升高[1],而且还会对人体的心血管系统、神经系统和免疫系统造成影响[2-3],甚至会对染色体和 DNA 等不同水平的遗传物质产生毒性作用,引起癌症和出生缺陷的发生[4-5].

$PM_{2.5}$ 的研究工作包括数据采集方法、机理成因和影响因素等[6-7].从统计学的角度,一个地区一段时间的 $PM_{2.5}$ 浓度作为典型的时空数据集,相关研究重点是空间插值以及时间上的短期或长期预测等.$PM_{2.5}$ 的空间插值较流行的是时空克里金方法[8-9],该方法能够基于时空数据时空位置关系与时空变异特征,实现对未观测位置的线性无偏最优估计,而 $PM_{2.5}$ 在时间维度上的预测可使用机理分析或统计建模方法等.机理分析方法主要对大气污染物的产生、转换和扩散的物理化学过程进行建模,如 CMAQ 模型[10]等;统计建模方法即通过数据捕捉特征,得到污染物浓度变化规律,包括多元线性回归(Multivariable Linear Regression,MLR)[11]、广义加性模型(Generalized Additive Model,GAM)[12-13],以及统计学习模型,如 BP 神经网络[14]和长短期记忆网络模型(Long Short-Term Memory,LSTM)[15-16]等的各种扩展模型.相比机理分析方法,统计方法较少依赖于污染源数据、传输模式和物理机理,更侧重数据本身的规律,定量分析的精确度更具优势,是处理复杂数据的强有力工具.

近年来,许多研究关注于 $PM_{2.5}$ 浓度的时空特征和统计推断.例如:Cheam 等[17]将 EM 算法应用于参数时空混合模型的推断中,用于对空气质量数据进行聚类;Clifford 等[18]在半参数时空模型的基础上,利用高斯马尔可夫随机场对空间随机效应和非参数时间趋势进行近似,对大气颗粒物浓度进行预测等贝叶斯推断.这些研究更多关注模型和计算的灵活性,没有考虑在触发空气污染方面起重要作用的气象变量.也有一些研究开发了包含气象变量的时空模型,并将其用于时空预测[19-20],包括 Wan 等[21]通过建立精细的参数统计模型对北京市 $PM_{2.5}$ 浓度进行综合研究,分析 $PM_{2.5}$ 浓度的时空依赖结构并进行了预测.但是在应对大规模数据,特别是多站点同步预测时,此类时空模型将面临过高的计算复杂度难题.

收稿日期 2021-11-18

资助项目 江苏省自然科学基金(BK20191394);国家社会科学基金重大项目(16ZDA047)

作者简介

王静,女,硕士生,研究方向为 $PM_{2.5}$ 等时空数据建模及分析.jwang_jane@163.com

曹春正(通信作者),男,教授,主要研究方向为函数型数据分析、气象中的复杂数据建模及分析.caochunzheng@163.com

1　南京信息工程大学　数学与统计学院,南京,210044

本文在贝叶斯框架下,结合分层模型理论,同时考虑气象因素,对北京市35个空气质量监测点的日均$PM_{2.5}$浓度整体建立了贝叶斯分层自回归(Bayesian Hierarchical Autoregression,BHAR)时空模型.模型有三个优点:一是利用分层结构描述出清晰的变量关联性和时空结构关系;二是利用贝叶斯方法可同时达到参数估计和多站点同步预测的目的;三是可以对除现有空气质量监测点之外,有气象信息的地点进行预测,解决部分地区空气质量监测点分布稀疏的问题.BHAR时空模型在潜在时空过程中同时对时间和空间相关性进行拟合,实现降维,解决了传统时空模型计算复杂度较高的问题.进一步借助R软件中的spTimer包[22],使用马尔可夫链蒙特卡罗(Markov Chain Monte Carlo,MCMC)算法对模型进行参数估计和预测.

1 数据初步分析

本文的研究区域北京位于115.7°~117.4°E,39.4°~41.6°N,地势西北高、东南低.西部、北部和东北部三面环山,东南部平原逐渐向渤海倾斜,见图1.

北京市原有27个空气质量监测点,2012年新增了8个$PM_{2.5}$监测点.由于从2021年1月23日起,北京市生态环境监测中心的空气质量发布平台(http://zx.bjmemc.com.cn)更新了北京市的空气质量监测点名称,按照城六区、东南部、东北部、西南部、西北部五个区域对监测点重新进行了划分(表1),同时考虑到$PM_{2.5}$污染多发生在冬季[23],因此本文收集了北京市35个监测点2021年2月1日至2021年3月31日每天24 h的$PM_{2.5}$质量浓度(μg/m^3)数据用于后续分析.本文将基于采集数据对$PM_{2.5}$每日24 h的平均质量浓度(日均质量浓度)进行建模和预测.

为初步探索北京市$PM_{2.5}$浓度的时空分布特征,首先分别按小时计算所有站点一天中每一小时的$PM_{2.5}$平均质量浓度,按天计算一周中每一天的平均质量浓度,然后绘制箱线图.由图2a可以观察到,$PM_{2.5}$质量浓度在一天中呈现出先下降再上升的趋势,自凌晨起逐渐下降,直至14时开始出现明显上升趋势,这可能与地形和冬季的气候条件有关,这些条件导致了热逆温和早晨、晚上的弱风,阻止了污染物扩散[24];同时,还可以发现在早高峰和晚高峰时间段内,$PM_{2.5}$质量浓度有小幅度上升,可能是受到汽车尾气的影响.图2b显示$PM_{2.5}$在一周中的变化也具有规律性,随着一周的人类活动水平不断提高,$PM_{2.5}$质量浓度也逐渐升高,而周一和休息日,$PM_{2.5}$质量浓度明显较低.

接下来分析北京$PM_{2.5}$浓度的空间分布特征.首先绘制2021年2月和3月北京市的平均$PM_{2.5}$质量浓度空间分布图.由图3可以观察到北部山区的$PM_{2.5}$质量浓度较低,其中延庆县的浓度最低,可能是受到了山谷地形的影响,而中心城区的$PM_{2.5}$质量浓度较高显然也是合理的.

然后借助全局莫兰指数分析$PM_{2.5}$质量浓度分布的空间自相关性.全局莫兰指数公式为

$$I=\frac{n\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\right)\sum_{i=1}^{n}(y_i-\bar{y})^2},\tag{1}$$

式中w_{ij}是距离权重,代表站点i和j之间的加权距离,y_i代表$PM_{2.5}$质量浓度值.经过计算得到$PM_{2.5}$质量浓度的全局莫兰指数$I=0.2$,为正值,且显著性检验p值等于3.9×10^{-6},说明北京市35个空气质量监测点2—3月$PM_{2.5}$质量浓度分布存在显著的正相关性及空间聚集性.

进一步收集北京地区延庆、密云、北京三个气象台站2021年2月1日到2021年3月31日每日的风速(m/s)、相对湿度(%)和最高温度(℃)数据,数据来自中国气象数据网(http://data.cma.cn).将三个气象变量分别简记为WS、RH和MT,详细汇总信息见表2.为了分析各气象变量和$PM_{2.5}$质量浓度的相

表1 北京市空气质量监测站点

Table 1 Air quality monitoring stations in Beijing

区域	监测点
城六区	东城东四、东城天坛、西城官园、西城万寿西宫、朝阳奥体中心、朝阳农展馆、海淀万柳、海淀四季青、丰台小屯、丰台云岗、石景山古城、石景山老山
东南部	通州永顺、通州东关、大兴黄村、大兴旧宫、亦庄开发区、京东南区域点
东北部	怀柔镇、怀柔新城、密云镇、密云新城、平谷镇、平谷新城、顺义新城、顺义北小营
西北部	昌平镇、昌平南邵、定陵(对照点)、延庆夏都、延庆石河营
西南部	门头沟双峪、门头沟三家店、房山良乡、房山燕山

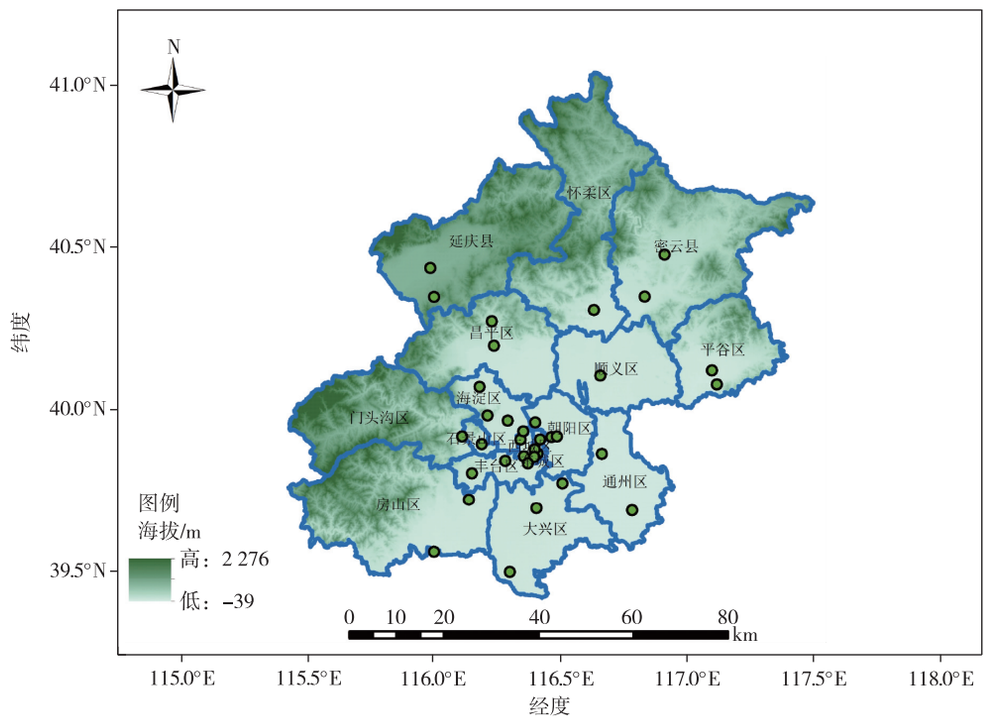

图 1　北京市地形及空气质量监测站点分布(绿点)

Fig. 1　Map of Beijing's topography and its air quality monitoring stations (green dots)

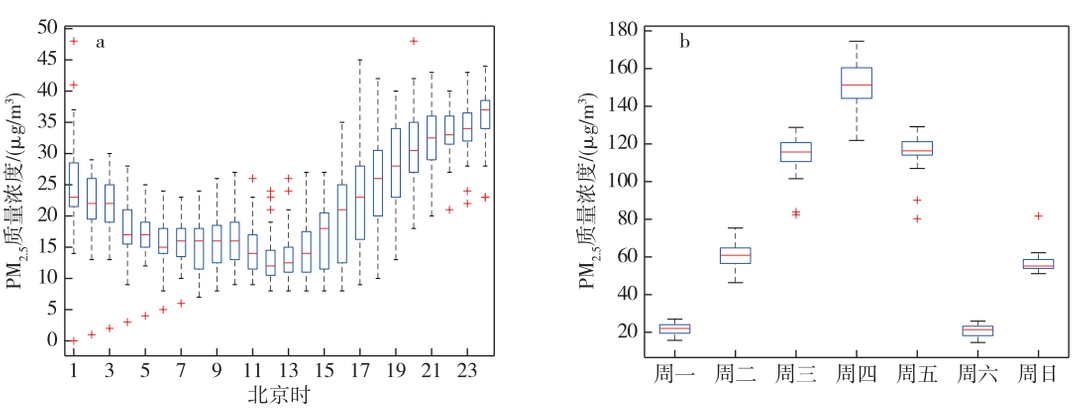

图 2　2021 年 2—3 月北京市 35 个空气质量监测站点一天(a)和一周(b)内的平均 $PM_{2.5}$ 质量浓度分布

Fig. 2　Diurnal (a) and weekly (b) variation of $PM_{2.5}$ concentration averaged by 35 air quality monitoring stations in Beijing from February to March 2021

关性,首先为三个气象站匹配了距离最近的空气质量监测点.与延庆、密云和北京三个气象站相匹配的空气质量监测点分别为延庆夏都、怀柔新城以及大兴黄村三个监测点.分别计算 $PM_{2.5}$ 质量浓度与三个气象变量的 Spearman 相关系数,结果见表 2 最后一列.结果表明,相对湿度与 $PM_{2.5}$ 呈现出较强的正相关.2—3 月北京的相对湿度比较低,此时大气污染物的化学聚合作用使得 $PM_{2.5}$ 增加的速率,高于沉降作用使得 $PM_{2.5}$ 降低的速率.相对湿度对 $PM_{2.5}$ 影响呈正向,即 $PM_{2.5}$ 质量浓度随着相对湿度的上升而增加.风速与 $PM_{2.5}$ 呈现负相关,而温度与 $PM_{2.5}$ 相关性较弱.

表 2　日均 $PM_{2.5}$ 质量浓度和气象变量的信息汇总及相关系数

Table 2　Summary of and correlation coefficients between daily $PM_{2.5}$ concentrations and meteorological variables

变量	均值	最小值	最大值	Spearman 相关系数
$\rho(PM_{2.5})/(\mu g/m^3)$	74.20	3.25	296.42	
WS/(m/s)	1.85	0.40	4.80	-0.44^{***}
RH/%	50.69	15.30	91.00	0.68^{***}
MT/℃	12.05	-0.20	25.60	0.26^{***}

注:*** 表示 $p<0.001$.

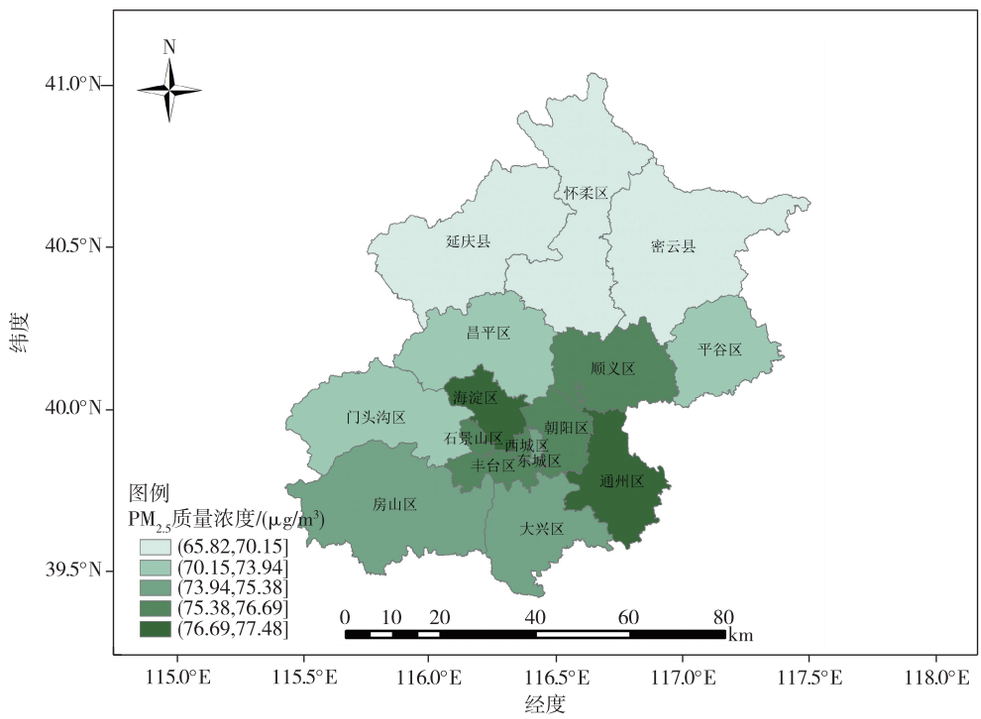

图 3　2021 年 2—3 月北京市平均 $PM_{2.5}$ 质量浓度分布

Fig. 3　Distribution of average $PM_{2.5}$ concentration in Beijing from February to March 2021

2　BHAR 时空模型

2.1　模型建立

假设 $\boldsymbol{Z}(s,t)$ 代表站点 s 在时间 t 的 $PM_{2.5}$ 质量浓度实际观测值,其对应的真实浓度值通过潜在的随机过程 $\boldsymbol{Y}(s,t)$ 刻画,二者满足如下测量误差模型:

$$\boldsymbol{Z}(s,t)=\boldsymbol{X}^{\mathrm{T}}(s,t)\boldsymbol{\beta}+\boldsymbol{Y}(s,t)+\boldsymbol{\varepsilon}(s,t), \tag{2}$$

其中:$s=s_1,s_2,\cdots,s_n$ 为 n 个站点的地理位置;$t=1,2,\cdots,T$ 为时间(d);$\boldsymbol{X}(s,t)$ 代表 p 维气象变量,即 $\boldsymbol{X}(s,t)=(x_1(s,t),x_2(s,t),\cdots,x_p(s,t))^{\mathrm{T}}$;$\boldsymbol{\beta}$ 为回归系数;$\boldsymbol{\varepsilon}(s,t)$ 为误差项,通常假定是白噪声过程,即 $\boldsymbol{\varepsilon}(s,t)\sim \mathrm{GP}(0,\sigma_\varepsilon^2)$.在空间统计中 σ_ε^2 通常被称为块金值,当采样点的距离为 0 时,半方差函数值也应为 0,但由于存在测量误差和空间变异,使得两采样点非常接近时,对应半方差函数值不为 0,即存在块金值.

对潜在的污染物排放水平 $\boldsymbol{Y}(s,t)$ 建立一阶自回归模型[22]:

$$\boldsymbol{Y}(s,t)=\rho\boldsymbol{Y}(s,t-1)+\boldsymbol{\eta}(s,t), \tag{3}$$

其中 $\boldsymbol{\eta}(s,t)$ 为残差随机项,用来刻画潜在污染物排放水平的时空随机效应.假定 $\boldsymbol{\eta}(s,t)$ 在时间上独立而在空间上满足高斯过程 $\mathrm{GP}(0,\boldsymbol{\Sigma}_\eta)$,其中 $\boldsymbol{\Sigma}_\eta=\sigma_\eta^2\boldsymbol{S}_\eta$,$\sigma_\eta^2$ 是不随空间变化的方差,$\boldsymbol{S}_\eta$ 代表与空间相关的协方差矩阵,通常用 Matérn 族相关函数来刻画[25].此时 $\boldsymbol{\eta}(s,t)$ 的协方差矩阵是 $n\times n$ 维的,而不是 $nT\times nT$ 维的,实现了降维,简化了计算.Matérn 族相关函数的一般形式为

$$\kappa(u;\phi,\nu)=\frac{1}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}u\phi)^{\nu}K_\nu(2\sqrt{\nu}u\phi),\quad \phi>0,\nu>0, \tag{4}$$

其中:$u=\|s_i-s_j\|$ 表示监测点 s_i 和 s_j 之间的距离,在这里选择的是欧氏距离;ϕ 用来控制空间相关性的衰减速度,即距离 u 越大,衰减速度越快;ν 为控制平滑程度的参数;K_ν 为 ν 阶的第二类贝塞尔函数.当 $\nu=0.5$ 时,Matérn 族相关函数退化为指数相关函数,即 $\kappa(u;\phi)=\exp(-\phi u)$;当 $\nu=3/2$ 时,$\kappa(u;\phi)=(1+\phi u)\exp(-\phi u)$;当 $\nu\to\infty$ 时,Matérn 族相关函数退化为高斯过程函数,即 $\kappa(u;\phi)=\exp(-\phi^2u^2)$.

综上,对于实测数据,BHAR 时空模型结构如下:

$$\boldsymbol{Z}_t=\boldsymbol{X}_t\boldsymbol{\beta}+\boldsymbol{Y}_t+\boldsymbol{\varepsilon}_t, \tag{5}$$

$$\boldsymbol{Y}_t=\rho\boldsymbol{Y}_{t-1}+\boldsymbol{\eta}_t, \tag{6}$$

其中 $\boldsymbol{Z}_t=(Z(s_1,t),\cdots,Z(s_n,t))^{\mathrm{T}}$,$\boldsymbol{Y}_t=(Y(s_1,t),\cdots,Y(s_n,t))^{\mathrm{T}}$,$\boldsymbol{\varepsilon}_t=(\varepsilon(s_1,t),\cdots,\varepsilon(s_n,t))^{\mathrm{T}}$,$\boldsymbol{\eta}_t=(\eta(s_1,t),\cdots,\eta(s_n,t))^{\mathrm{T}}$,且 $\boldsymbol{\varepsilon}_t\sim N(0,\sigma_\varepsilon^2\boldsymbol{I}_n)$,$\boldsymbol{\eta}_t\sim N(0,\sigma_\eta^2\boldsymbol{S}_\eta)$.根据分层模型结构,BHAR 时空模型可

分为三层:第一层表示在给定潜在时空过程和参数的条件下原始数据的分布;第二层表示参数给定的条件下潜在过程的分布 $\boldsymbol{Y}_t \mid \boldsymbol{\Theta}$;第三层代表引入的参数或超参数的先验分布.其中第二层中的过程可以添加不同水平的解释[26],第一水平表示真正的潜在过程,第二水平刻画了潜在过程的时空随机效应.

2.2 参数估计与预测

BHAR 时空模型中的待估参数为 $\boldsymbol{\Theta}=\{\boldsymbol{\beta},\rho,\sigma_\varepsilon^2,\sigma_\eta^2,\phi,\nu\}$,利用 MCMC 方法进行参数估计.除了 ϕ 和 ν 以外的其他参数都存在共轭先验分布,即在给定先验分布后 $\boldsymbol{\beta},\rho,\sigma_\varepsilon^2,\sigma_\eta^2$ 都可以求出标准的满条件后验分布,进一步采用 Gibbs 抽样法进行参数估计.固定 $\nu=0.5$,对于参数 ϕ 采用 Metropolis-Hastings (MH)算法进行估计.

对 $\boldsymbol{Z}(s,t)$ 的预测可以分为三类:一是预测未知监测点 s_0 在已知时间 t 的值;二是预测已知监测点 s 在未知时间点 t_0 的值;三是预测未知监测点 s_0 在未知时间点 t_0 的值.将第一类预测为空间插值,第二、三类预测为在时间上的预测.

首先介绍空间插值,在未知监测点 s_0,由方程(1)可以得到:

$$\boldsymbol{Z}(s_0,t)=\boldsymbol{X}^{\mathrm{T}}(s_0,t)\boldsymbol{\beta}+\boldsymbol{Y}(s_0,t)+\boldsymbol{\varepsilon}(s_0,t), \quad (7)$$

其中 $\boldsymbol{Y}(s_0,t)=\rho\boldsymbol{Y}(s_0,t-1)+\boldsymbol{\eta}(s_0,t)$.显然,$\boldsymbol{Y}(s_0,t)$ 只能由 t 之前所有时间点的 $\boldsymbol{Y}(s_0,\cdot)$ 顺序确定,且包括 $\boldsymbol{Y}(s_0,0)$.可以根据初始条件 Y_0 的先验分布来计算 $\boldsymbol{Y}(s_0,0)$.当然如果指定 Y_0 为一固定常数,那么 $\boldsymbol{Y}(s_0,0)$ 也可以被认为是同样的常数[19],因此为了简便,Y_0 通常选择固定值.

$\boldsymbol{Z}(s_0,t)$ 的预测一般基于其后验分布 $\pi(\boldsymbol{Z}(s_0,t)\mid\boldsymbol{Z})$,该后验分布可以通过对联合后验分布积分得到:

$$\begin{aligned}\pi(\boldsymbol{Z}(s_0,t)\mid\boldsymbol{Z})=&\int\pi(\boldsymbol{Z}(s_0,t)\mid\boldsymbol{Y}(s_0,t),\sigma_\varepsilon^2)\times\\&\pi(\boldsymbol{Y}(s_0,t)\mid\boldsymbol{\Theta},\boldsymbol{Y})\times\\&\pi(\boldsymbol{\Theta},\boldsymbol{Y}\mid\boldsymbol{Z})\mathrm{d}\boldsymbol{Y}(s_0,t)\mathrm{d}\boldsymbol{Y}\mathrm{d}\boldsymbol{\Theta},\end{aligned} \quad (8)$$

上式中 $\boldsymbol{Z},\boldsymbol{Y}$ 分别代表已知时间 t 和监测点 s 的值.预测值 $\boldsymbol{Z}(s_0,t)$ 的估计通过 MCMC 按成分抽样法获得,具体抽样方法如下:

1) 从后验分布 $\pi(\boldsymbol{\Theta},\boldsymbol{Y}\mid\boldsymbol{Z})$ 中抽取随机样本 $\boldsymbol{\Theta}^{(j)},\boldsymbol{Y}^{(j)}$;

2) 从后验分布 $\pi(\boldsymbol{Y}(s_0,t)\mid\boldsymbol{\Theta}^{(j)},\boldsymbol{Y}^{(j)})$ 中抽取样本 $\boldsymbol{Y}^{(j)}(s_0,t)$;

3) 从后验分布 $\pi(\boldsymbol{Z}(s_0,t)\mid\boldsymbol{Y}^{(j)}(s_0,t),\sigma_\varepsilon^{2(j)})$ 中抽取样本 $\boldsymbol{Z}^{(j)}(s_0,t)$.

在时间维度上的预测过程与空间插值过程类似,对某一站点 s(包含现有监测点或任意指定位置作为监测点),做向前一步时间预测,同样可以基于 $\boldsymbol{Z}(s,T+1)$ 的后验分布根据类似空间插值的 MCMC 抽样方法实现.与空间插值的不同主要体现在时间维度上的预测需要从具有零均值、方差为 $\sigma_\eta^2\boldsymbol{S}_\eta$ 的边缘分布出发来模拟 $\boldsymbol{Y}(s,T+1)$,而不是条件分布.因为从 0 时刻一直到时刻 T,观测点的信息已经全部被用来获得 $\boldsymbol{Y}(s,T)$,在未来的 $T+1$ 时刻,除了回归项 $\boldsymbol{X}(s,T+1)$ 的新值外,已经没有可用于条件分布的新信息.

3 实例分析

结合北京市 35 个空气质量监测点的空间分布情况,选取其中 9 个监测点作为空间验证集,同时选取 2021 年 3 月 30 日和 3 月 31 日两天作为时间验证集,剩下的 26 个监测点数据作为训练集用来拟合模型.利用 R 软件包 spTimer 同时实现以下参数估计和预测的计算过程.由参数估计表(表 3)可以看到,WS 和 MT 两个变量的回归系数 β_1 和 β_3 的估计的 95% 置信区间包含零点,因此是不显著的.气象变量中只有相对湿度 RH 的回归系数 β_2 显著且为正值,这与数据初步分析的结果一致,进一步说明了 2—3 月北京的相对湿度 RH 对 $PM_{2.5}$ 有显著的正向影响.

表 3 BHAR 模型的 MCMC 参数估计

Table 3 MCMC parameter estimation of BHAR model

参数	均值	中位数	标准差	95%置信区间
β_0	3.557 1	3.546 8	0.679 0	[2.234 2,4.900 9]
β_1	0.108 4	0.108 9	0.075 9	[-0.039 9,0.254 2]
β_2	0.017 4	0.017 4	0.007 8	[0.001 6,0.032 9]
β_3	-0.000 3	-0.000 7	0.025 3	[-0.050 3,0.049 0]
ρ	0.414 4	0.414 6	0.023 4	[0.367 8,0.459 6]
σ_ε^2	0.006 3	0.006 3	0.000 3	[0.005 8,0.006 9]
σ_η^2	5.949 4	5.852 2	0.858 1	[4.585 4,7.754 7]
ϕ	0.002 0	0.002 0	0.000 3	[0.001 5,0.002 6]

为了评估 BHAR 时空模型的预测性能,采用均方根误差 RMSE 和平均绝对误差 MAE 两个测量指标对预测数据和原始数据进行误差对比分析.RMSE 和 MAE 的公式如下:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Z}_i-Z_i)^2}, \quad (9)$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|\hat{Z}_i-Z_i|. \tag{10}$$

首先对空间验证集的 9 个监测点进行空间插值,BHAR 时空模型可以实现对 9 个监测点的同步预测.为了进行对比,进一步利用 R 软件的 mgcv 包中的 gam 函数对训练集数据进行 GAM 模型拟合,预测空间验证集中各站点的 $PM_{2.5}$ 质量浓度,同时计算以上两个测量指标.Sorek-Hamer 等[12]利用 GAM 能够拟合解释变量与被解释变量间非线性关系的优势,提高了 $PM_{2.5}$ 质量浓度的预测效果.对比结果见表 4,可以看到 BHAR 时空模型的 RMSE 和 MAE 都约为 GAM 的 1/3,证明本文提出的 BHAR 时空模型的空间插值效果一致优于 GAM.

表 4 空间插值效果对比

Table 4 Comparison of spatial interpolation performance

μg/m³

模型	RMSE	MAE
BHAR	12.45	8.16
GAM	34.80	24.97

进一步,分别对空间验证集和训练集中的监测点做同步时间预测,预测 3 月 30 日和 31 日两天的日均 $PM_{2.5}$ 质量浓度值.选取 LSTM 作为主要对比模型,选取常规的 ARIMA 模型作为基准对比模型.将得到的测量指标列在表 5 中,可以看到 ARIMA 模型的预测效果最差,其次是 LSTM 模型,BHAR 模型的时间预测效果最好.本文提出的 BHAR 时空模型对所有监测点进行整体建模,充分考虑了空间和时间相关性,因此得到了准确度较高的预测结果.

表 5 在时间维度上的预测效果对比

Table 5 Comparison of prediction performance in time dimension

μg/m³

模型	RMSE	MAE
BHAR	10.12	8.68
LSTM	12.22	11.48
ARIMA	26.81	24.85

4 总结

本文建立的 BHAR 时空模型是将一个区域的 $PM_{2.5}$ 数据看作时间序列的空间过程,从整体上拟合 $PM_{2.5}$ 质量浓度的时间和空间相关性特征,实现了对特定站点的 $PM_{2.5}$ 空间插值和时间上的短期预测功能,预测效果优于 GAM 和 LSTM.该模型不仅适用于 $PM_{2.5}$ 质量浓度预测,还可以推广到其他空气质量地面监测数据,例如 PM_{10} 和 O_3 浓度等.在贝叶斯框架下建模,对模型的不确定性更具有包容性,而且提前给定先验分布可以融合专家知识,提高预测精度.进一步,建立分层模型可以更加清晰地刻画出数据内部的潜在时空过程,增强模型的可解释性.同时,模型的分层结构也使得参数估计和预测等推断过程更加便捷.

本文选取了风速、湿度和温度三个气象因素作为解释变量,用以提高模型的实际预测效果.为了简化模型,本文采用的是固定系数,但气象变量对 $PM_{2.5}$ 的影响也可能会随着时间和空间的变化而存在差异,因此在后续研究中将考虑拟合变系数模型.本文采用一阶自回归来刻画时间维度上的相关性,达到了较好的数值效果,同时也降低了计算复杂度.在实际应用中,可以针对数据特点,结合模型选择方法选取合适的自回归阶数.此外,本文模型中用于刻画空间相关性的 Matérn 核函数采用了齐次的欧氏距离,在有更多的地理细节特征下,可以考虑非齐次的欧氏距离或其他非欧距离,捕捉更为真实的空间相关性.

参考文献

References

[1] Pope C A III, Burnett R T, Thun M J, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution [J]. JAMA, 2002, 287(9): 1132-1141

[2] Guo Y M, Jia Y P, Pan X C, et al. The association between fine particulate air pollution and hospital emergency room visits for cardiovascular diseases in Beijing, China[J]. Science of the Total Environment, 2009, 407(17): 4826-4830

[3] 郭玉明,刘利群,陈建民,等.大气可吸入颗粒物与心脑血管疾病急诊关系的病例交叉研究[J].中华流行病学杂志,2008,29(11):1064-1068
GUO Yuming, LIU Liqun, CHEN Jianmin, et al. Association between the concentration of particulate matters and the hospital emergency room visits for circulatory diseases: a case-crossover study[J]. Chinese Journal of Epidemiology, 2008, 29(11): 1064-1068

[4] Valavanidis A, Fiotakis K, Vlachogianni T. Airborne particulate matter and human health: toxicological assessment and importance of size and composition of particles for oxidative damage and carcinogenic mechanisms[J]. Journal of Environmental Science and Health, Part C: Environmental Carcinogenesis & Ecotoxicology Reviews, 2008, 26(4): 339-362

[5] Samet J M, DeMarini D M, Malling H V. Do airborne particles induce heritable mutations? [J]. Science, 2004,

304(5673):971-972
[6] Guo S,Hu M,Zamora M L,et al.Elucidating severe urban haze formation in China[J].Proceedings of the National Academy of Sciences of the United States of America,2014,111(49):17373-17378
[7] Sun Y L,Wang Z F,Du W,et al.Long-term real-time measurements of aerosol particle composition in Beijing,China:seasonal variations,meteorological effects,and source analysis[J].Atmospheric Chemistry and Physics,2015,15(17):10149-10165
[8] 卢月明,王亮,仇阿根,等.局部加权线性回归模型的 $PM_{2.5}$ 空间插值方法[J].测绘科学,2018,43(11):79-84,91
LU Yueming,WANG Liang,QIU Agen,et al.$PM_{2.5}$ spatial interpolation method based on local weighted linear regression model[J].Science of Surveying and Mapping,2018,43(11):79-84,91
[9] Gething P,Atkinson P,Noor A,et al.A local space-time Kriging approach applied to a national outpatient malaria dataset[J].Computers & Geosciences,2007,33(10):1337-1350
[10] Byun D,Schere K L.Review of the governing equations,computational algorithms,and other components of the models-3 community multiscale air quality (CMAQ) modeling system[J].Applied Mechanics Reviews,2006,59(2):51-77
[11] 李颖若,汪君霞,韩婷婷,等.利用多元线性回归方法评估气象条件和控制措施对 APEC 期间北京空气质量的影响[J].环境科学,2019,40(3):1024-1034
LI Yingruo,WANG Junxia,HAN Tingting,et al.Using multiple linear regression method to evaluate the impact of meteorological conditions and control measures on air quality in Beijing during APEC 2014[J].Environmental Science,2019,40(3):1024-1034
[12] Sorek-Hamer M,Strawa A W,Chatfield R B,et al.Improved retrieval of $PM_{2.5}$ from satellite data products using non-linear methods[J].Environmental Pollution,2013,182C:417-423
[13] Yu S,Wang G N,Wang L,et al.Estimation and inference for generalized geoadditive models[J].Journal of the American Statistical Association,2020,115(530):761-774
[14] Bai Y,Li Y,Wang X X,et al.Air pollutants concentrations forecasting using back propagation[J].Atmospheric Pollution Research,2016,7(3):557-566
[15] Zhou Q P,Jiang H Y,Wang J Z,et al.A hybrid model for $PM_{2.5}$ forecasting based on ensemble empirical mode decomposition and a general regression neural network[J].Science of the Total Environment,2014,496:264-274
[16] 白盛楠,申晓留.基于 LSTM 循环神经网络的 $PM_{2.5}$ 预测[J].计算机应用与软件,2019,36(1):67-70,104
BAI Shengnan,SHEN Xiaoliu.$PM_{2.5}$ prediction based on LSTM recurrent neural network[J].Computer Applications and Software,2019,36(1):67-70,104
[17] Cheam A S M,Marbac M,McNicholas P D.Model-based clustering for spatiotemporal data on air quality monitoring[J].Environmetrics,2017,28(3):e2437
[18] Clifford S,Low-Choy S,Mazaheri M,et al.A Bayesian spatiotemporal model of panel design data:airborne particle number concentration in Brisbane,Australia[J].Environmetrics,2019,30(7):e2597
[19] Nicolis O,Díaz M,Sahu S K,et al.Bayesian spatiotemporal modeling for estimating short-term exposure to air pollution in Santiago de Chile[J].Environmetrics,2019,30(7):e2574
[20] Padilla L,Lagos-Álvarez B,Mateu J,et al.Space-time autoregressive estimation and prediction with missing data based on Kalman filtering[J].Environmetrics,2020,31(7):e2627
[21] Wan Y T,Xu M Y,Huang H,et al.A spatio-temporal model for the analysis and prediction of fine particulate matter concentration in Beijing[J].Environmetrics,2021,32(1):e2648
[22] Bakar K S,Sahu S K.spTimer:spatio-temporal Bayesian modeling using R[J].Journal of Statistical Software,2015,63(15):1-32
[23] 梁丽思,靖娟利,王安娜,等.2014—2019 年冬季京津冀地区 $PM_{2.5}$ 质量浓度时空分布特征[J].桂林理工大学学报,2020,40(4):788-797
LIANG Lisi,JING Juanli,WANG Anna,et al.Spatial-temporal distribution characteristics of $PM_{2.5}$ concentrations in Beijing-Tianjin-Hebei region in winter from 2014 to 2019[J].Journal of Guilin University of Technology,2020,40(4):788-797
[24] 王晨,时悦,景悦,等.基于遥感数据的京津冀地区 $PM_{2.5}$ 时空分布特征[J].环境监测管理与技术,2020(1):37-41
WANG Chen,SHI Yue,JING Yue,et al.Spatial and temporal distribution characteristics of $PM_{2.5}$ in Beijing-Tianjin-Hebei region based on remote sensing data[J].The Administration and Technique of Environmental Monitoring,2020(1):37-41
[25] Handcock M S,Stein M L.A Bayesian analysis of Kriging[J].Technometrics,1993,35(4):403-410
[26] Gelfand A E.Hierarchical modeling for spatial data problems[J].Spatial Statistics,2012,1:30-39

Prediction of $PM_{2.5}$ concentration in Beijing based on Bayesian hierarchical autoregressive spatio-temporal model

WANG Jing[1] CAO Chunzheng[1]
1 School of Mathematics and Statistics,Nanjing University of Information Science & Technology,Nanjing 210044

Abstract Here,a hierarchical autoregressive spatio-temporal model under the Bayesian framework is proposed to address the simultaneous multi-site $PM_{2.5}$ prediction.The true daily average concentration of $PM_{2.5}$ is regarded as a potential spatio-temporal process,then the temporal correlation is described by the first-order autoregressive process and the spatial correlation is captured based on the Matérn process,which greatly improves the efficiency in dimension reduction and synchronous prediction.In addition,meteorological factors such as daily maximum temperature, relative humidity and wind speed are used as explanatory variables to improve the prediction accuracy.The combination of Bayesian method and MCMC can realize parameter estimation and prediction process due to the model's hierarchical structure.The empirical analysis of daily $PM_{2.5}$ concentration in Beijing shows that the proposed model has good interpolation or prediction performance in both spatial and temporal dimensions.

Key words Bayesian method;hierarchical model;autoregressive;spatio-temporal model;$PM_{2.5}$ prediction;Markov Chain Monte Carlo (MCMC)