汾渭平原空气质量数据的函数型主成分分析*

李妍琳，石小平，胡锡健†
(新疆大学 数学与系统科学学院，新疆 乌鲁木齐 830046)

摘　要： 汾渭平原作为我国的四大平原之一, 由于该地区大气污染事件频发, 已成为我国大气污染最严重的区域之一, 引起了社会各界普遍的关注. 本文基于汾渭平原地区11个城市的空气质量数据, 研究了2019年汾渭平原地区的$PM_{2.5}$浓度日数据. 根据空气质量数据的函数型特性, 采用函数型数据分析方法对$PM_{2.5}$浓度数据连续化, 从图像上能直观精确地看出2019年汾渭平原$PM_{2.5}$浓度变化动态, 进而对函数化的数据进行函数型主成分分析. 结果表明: 气候温度是影响汾渭平原地区空气质量的因素之一, 冬季采暖期各市的$PM_{2.5}$浓度普遍偏高; 地理位置也影响$PM_{2.5}$浓度, 河谷平原的$PM_{2.5}$浓度明显高于两侧山地，且呈现出向两侧山地递减趋势.

关键词： 汾渭平原; $PM_{2.5}$；函数型数据；函数型主成分分析
DOI： 10.13568/j.cnki.651094.651316. 2020.10.07.0001
中图分类号： O212.7　**文献标识码：** A　**文章编号：** 2096-7675(2021)06-0675-06
引文格式： 李妍琳, 石小平, 胡锡健. 汾渭平原空气质量数据的函数型主成分分析[J].新疆大学学报(自然科学版)(中英文), 2021, 38(6): 675-680.
英文引文格式： LI Y L, SHI X P, HU X J. Functional principal component analysis of air quality data in Fenwei Plain[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English), 2021, 38(6): 675-680.

Functional Principal Component Analysis of Air Quality Data in Fenwei Plain

LI Yanlin, SHI Xiaoping, HU Xijian

(*School of Mathematics and System Sciences, Xinjiang University, Urumqi Xinjiang 830046, China*)

Abstract： As one of the four great plains in China, Fenwei Plain has become one of the most serious air pollution areas in China due to its frequent air pollution events, which has aroused widespread concern from all walks of life. This paper studies the daily data of $PM_{2.5}$ concentration in Fenwei Plain in 2019. According to the functional characteristics of air quality data, the $PM_{2.5}$ concentration data is continuous by using the functional data analysis method, and the dynamic change of $PM_{2.5}$ concentration in Fenwei Plain in 2019 can be seen more intuitively and accurately from the image. The results show that the temperature is the main factor affecting the air quality in Fenwei Plain, and the $PM_{2.5}$ concentration of each city is generally higher in winter heating period; the concentration of $PM_{2.5}$ in the valley plain is significantly higher than that in the mountains on both sides.
Key words： Fenwei Plain; $PM_{2.5}$; functional data; functional principal component analysis

0　引 言

随着我国城镇化和工业化的快速推进、能源消耗量的持续增加, 大气污染问题已成为社会各界普遍关注的热点. 大气污染防治面临着严峻考验, 尤其是对汾渭平原的大气污染防治, 已成为当地环境质量改善工作的重点和难点. 汾渭平原的能源结构以煤为主, 煤炭在能源消费中占约90%, 远高于全国的平均水平(60%). 从地理位置来看, 汾渭平原北起山西省代县, 南抵陕西省秦岭山脉, 西至陕西省宝鸡市, 呈东北-西南方向分布, 受山脉阻挡和背风坡气流下沉作用的影响, 该地区容易形成反气旋式的气流停滞区, 在污染阶段地面辐合形式明显, 污染物辐合后被困, 不易扩散. 近年来汾渭平原的大气污染事件频发, 已经引起国家和社会的高度重视，但众多学者对空气质量状况的研究, 主要集中在中国东部地区，尤其是京津冀、长三角和珠三角等传统的空气污染重点

* **收稿日期：** 2020-10-07
基金项目： 国家自然科学基金(11961065；U1703237)；新疆高校科研计划项目(XJEDU2017M001).
作者简介： 李妍琳(1995-)，女，硕士生，从事空间统计的研究，E-mail: 1057653947@qq.com.
† **通讯作者：** 胡锡健(1964-)，男，博士，副教授，从事空间统计的研究，E-mail: xijianhu@126.com.

治理区域, 对西部地区的研究相对较少. 汾渭平原的生态环境有恶化趋势, 大气污染防治压力骤增. 2018年7月, 国务院印发《打赢蓝天保卫战三年行动计划》, 汾渭平原被纳入环境污染三大重点防控区域之一[1].

现今, 基于各个地方的空气质量情况, 我国建立了空气污染指数(API)、空气质量指数(AQI)及各类污染物指标数据的监测发布平台. 由于京津冀地区的地理位置原因, 其空气质量问题一直是政府关注的重点. 汾渭平原紧邻京津冀地区, 是京津冀地区的南部屏障, 各级政府非常重视本地的环保工作. $PM_{2.5}$ 浓度偏高对环境和人体健康有着不可忽视的影响, 楚德见等[2]分析了$PM_{2.5}$对高层建筑中人们生活环境的影响. 因此, 需要对汾渭平原空气质量数据做更为科学及系统化的分析, 以期对汾渭平原的空气质量改善提供更好的科学依据.

由于空气质量数据在时间尺度上有明显的函数特征, 而且累计数据是从2013年至今, 已达到上亿条, 这对分析空气质量精细化奠定了坚实的基础, 面对如此庞大的数据集, 常采用插值或平滑方法将离散的空气质量数据拟合成曲线, 运用函数型数据分析(Functional Data Analysis)方法分析. Ramsay[3]于1982年率先提出这种全新的数据分析思路. Ramsay和Sliverman[4−5]对函数型数据做了进一步详细的描述并讲述了诸多关于FDA的应用. 与传统方法相比, 函数型数据分析方法不仅在处理高维观测数据上能给出更加合理的直观解释, 而且在分析数据时能保留更多的数据信息, 从而得到更精确的分析结果. 函数型主成分分析作为函数型数据分析的有力工具, 得到了众多学者的广泛应用[6], 与传统多元主成分分析相比, 函数型主成分分析展现出了更大的优越性, 并且能够提取更多的重要数据信息. 目前, 国内很多学者对该方法都进行了研究, 吴京旺等[7]将该方法应用到了金融领域中; 唐裔等[8]运用函数型主成分方法分析了我国城市人口的变化差异. 在对空气质量数据的研究中, 梁银双等[9]运用函数型主成分分析方法对京津冀地区$PM_{2.5}$污染特征进行了分析, 并且得到了较好的结果. 目前还没有学者利用函数型数据方法对汾渭平原地区的空气质量进行研究. 本文以汾渭平原11个城市的空气质量问题为研究重点, 采用傅里叶基函数, 选取各城市2019年1月1日至2019年12月31日的$PM_{2.5}$浓度数据作为研究数据, 将汾渭平原地区11个城市2019年的$PM_{2.5}$浓度离散数据转化为连续的函数型数据, 应用函数型主成分分析寻找主成分指标, 刻画各城市$PM_{2.5}$浓度随时间的变化规律.

1 数据说明与方法

1.1 研究数据

本文选取汾渭平原（包括河南的洛阳、三门峡, 陕西的西安、咸阳、宝鸡、铜川、渭南, 但不含杨凌, 山西的吕梁、晋中、临汾、运城）11个城市作为研究对象, 整理了2019年汾渭平原地区11个城市的7项空气质量数据($PM_{2.5}$、$PM_{10.0}$、SO_2、NO_2、O_3 和CO的监测数据及空气质量指数(AQI)), 数据来自中国空气质量在线监测分析平台. 图1为2019年汾渭平原地区11个城市$PM_{2.5}$浓度变化折线图.

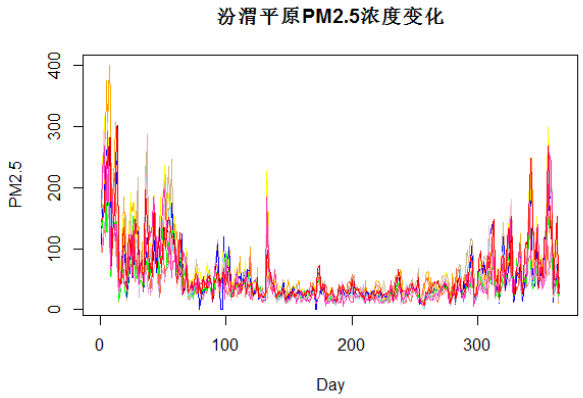


图 1 2019年汾渭平原地区11个城市$PM_{2.5}$浓度变化折线图

Fig 1 Line graph of $PM_{2.5}$ concentration changes in 11 cities in the Fenwei Plain in 2019

1.2 函数型主成分分析

作为传统主成分分析的一种推广, 函数型主成分分析将多元主成分分析技术与函数型数据分析相结合, 在面临更大的“维数灾难”时, 可以得到更加精确的分析结果.

1.2.1 曲线拟合

假定有n个观测样本, 每个样本有N对数据序列, 第i个样本的数据序列为$(t_1,x_{i1}),(t_2,x_{i2}),\cdots,(t_N,x_{iN})$, 将离散点对拟合成函数形式$x_i(t)$, 此时的$x_i(t)$满足$x_i(t_j)=x_{ij}+\epsilon_i$. 采用基函数方法拟合数据序列, 利用傅里叶基函数展开, 选择一组基函数$\Phi(t)=\{\phi_1(t),\phi_2(t),\cdots,\phi_K(t)\}$的线性组合来估计函数$x_i(t)$的值：

$$x_i(t)=\sum_{k=1}^{K}c_{ik}\phi_k(t) \tag{1}$$

其中：$x_i(t)$为第i个样本的曲线拟合, $\phi_k(t)$为第k个基函数, c_{ik}为对应的系数. 通过最小二乘法得到系数的估计

值，从而拟合曲线$x_i(t)$.

1.2.2　函数型主成分分析原理

假设已经得到拟合曲线$x_i(t)(t\in T,i=1,\cdots,n)$. 各个函数曲线的主成分得分为:

$$\xi_i=\int\beta(t)x_i(t)\,\mathrm{d}t,i=1,\cdots,n \tag{2}$$

其中：$\beta(t)$为权重函数.

从而, 求解第一主成分就变成了求解如下带有约束条件的优化问题:

$$\begin{cases}\max\dfrac{1}{n}\sum\limits_{i=1}^{n}\xi_i^2=\max\dfrac{1}{n}\sum\limits_{i=1}^{n}(\int\beta(t)x_i(t)\,\mathrm{d}t)^2\\ s.t.\ \|\beta\|^2=\int\beta(t)\beta(t)\,\mathrm{d}t=1\end{cases} \tag{3}$$

通过求解这个优化问题, 就得到了第一主成分$\beta_1(t)$.

同理, 可求得第k个主成分, 在满足前$k-1$个主成分权重函数相互垂直的基础上, 求解上述优化问题, 即

$$\begin{cases}\max\dfrac{1}{n}\sum\limits_{i=1}^{n}{\xi_i}^2=\max\dfrac{1}{n}\sum\limits_{i=1}^{n}(\int\beta(t)x_i(t)\,\mathrm{d}t)^2\\ s.t.\ \|\beta\|^2=\int\beta(t)\beta(t)\,\mathrm{d}t=1\\ \int\beta(t)\beta_l(t)\,\mathrm{d}t=0,l=1,\cdots,k-1\end{cases} \tag{4}$$

这个优化问题的求解与传统的主成分分析的方法类似, 通过拟合曲线的协方差函数矩阵, 求解函数型主成分的权函数$\beta(t)$. 记协方差函数为:

$$\nu(s,t)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i(s)-\bar{x}(s))(x_i(t)-\bar{x}(t)) \tag{5}$$

那么权重矩阵$\beta(t)$满足特征方程:

$$\int\nu(s,t)\beta(t)\,\mathrm{d}t=\lambda\beta(t) \tag{6}$$

定义积分变换:

$$V\beta(t)=\int\nu(s,t)\beta(t)\,\mathrm{d}t \tag{7}$$

这里的V为协方差算子, 它将权重函数以协方差函数$\nu(s,t)$为内核做积分变换, 则

$$V\beta(t)=\lambda\beta(t) \tag{8}$$

类比传统的主成分分析, 同样使用特征值的累计贡献率来衡量主成分的占比:

$$FVE=\sum_{i=1}^{K}\lambda_i/\sum_{i=1}^{n-1}\lambda_i$$

一般累计贡献率要求不小于85%.

1.2.3　函数型主成分分析原理

设函数$x_i(t)$的基函数展开式如(1)式, 令函数向量$X(t)=(x_1(t),x_2(t),\cdots,x_n(t))',\Phi(t)=(\phi_1(t),\phi_2(t),\cdots,\phi_K(t))'$, 则所有曲线的基函数展开式为$X=C\Phi$, 协方差函数的矩阵形式为

$$\nu(s,t)=\frac{1}{n-1}X'X=\frac{1}{n-1}\Phi(s)'C'C\Phi(t) \tag{9}$$

现假定特征函数$\beta(t)$的基函数展开式为:

$$\beta(t)=\sum_{k=1}^{K}b_k\phi_k(t) \tag{10}$$

其中：$b=(b_1,b_2,\cdots,b_k)'$, 则上式可写成矩阵形式$\beta(t)=\Phi(s)'b$, 从而得

$$\int\nu(s,t)\beta(t)\,\mathrm{d}t=\int\frac{1}{n-1}\Phi(s)'C'C\Phi(t)\Phi(s)'b\,\mathrm{d}t=\frac{1}{n-1}\Phi(s)'C'CWb \tag{11}$$

令$W=\int\Phi\Phi'\,\mathrm{d}t$, 其中$W$为$K$阶对称矩阵. 将(9)(10) 式代入(5)式可得

$$\frac{1}{n-1}\Phi(t)'C'CWb=\lambda\Phi(t)'b \tag{12}$$

对于$\forall t\in T$, 式(12)均成立，故$\frac{1}{n-1}C'CWb=\lambda b$, 求解该矩阵方程的特征问题即可得到权重函数$\beta(t)$.

2 汾渭平原空气质量数据的函数型主成分分析

2.1 曲线拟合

本文选取汾渭平原11个城市2019年的空气质量监测数据, 采用傅里叶样条基函数拟合，并基于R语言编程[10]绘制出函数化的$PM_{2.5}$浓度变化曲线, 如图2所示.

从图2可以看出, 原始数据经过傅里叶样条基函数处理后得到了光滑函数曲线, 反映2019年汾渭平原11个城市的$PM_{2.5}$浓度变化趋势. 图像显示$PM_{2.5}$浓度有一定的季节性和周期性变化特征. 总体上$PM_{2.5}$浓度表现为冬季浓度高于另外三个季节, 1 月、2月、12月的$PM_{2.5}$浓度值都较大, 达到最高峰值; 夏季浓度最低, 6—7月的$PM_{2.5}$浓度值均在0~50 $\mu g/m^3$, 属于良好的空气状态. $PM_{2.5}$浓度有这样的变化动态主要原因是冬季处于采暖期, 随着气温的回升和雨季的到来, 大气污染物排放量逐渐减少, 大气对$PM_{2.5}$的稀释和湿沉降能力增强, $PM_{2.5}$浓度逐渐下降.

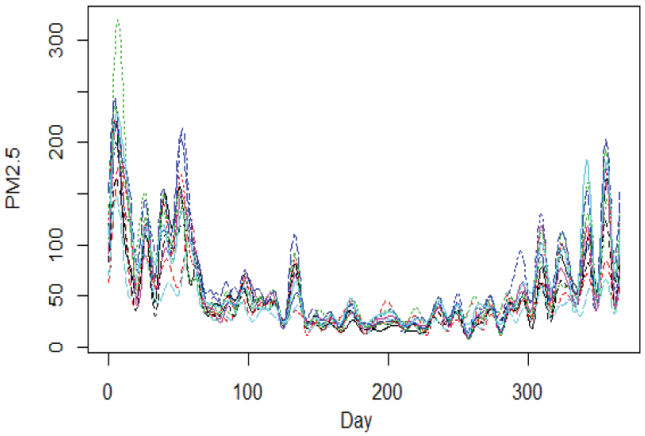
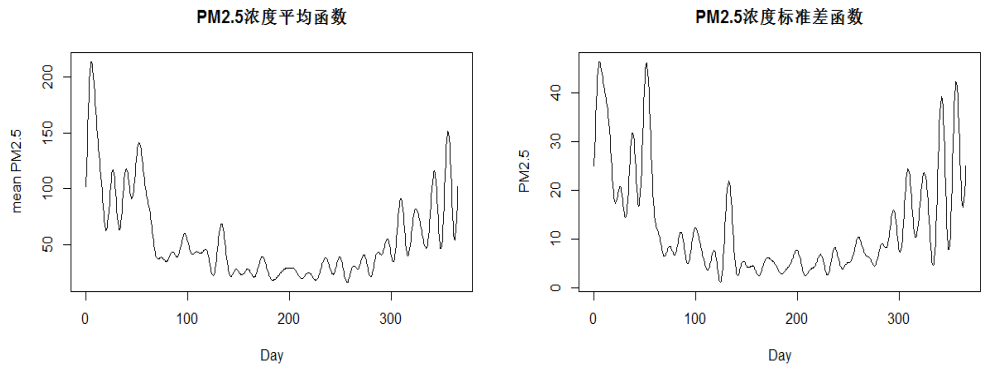


图 2 2019年汾渭平原地区11个城市$PM_{2.5}$浓度变化曲线

Fig 2 Concentration curve of $PM_{2.5}$ in 11 cities in the Fenwei Plain in 2019

2.2 相关描述性分析

采用傅里叶基函数拟合得到2019年汾渭平原11个城市的$PM_{2.5}$浓度均值曲线以及标准差曲线图, 如图3所示. 由均值曲线可以看出2019年汾渭平原的$PM_{2.5}$浓度大约在1月份处于最高水平, 最高峰在220 $\mu g/m^3$左右. 5—10月的$PM_{2.5}$浓度达到良好状态. 2—3月, 11—12月$PM_{2.5}$浓度处于轻中度污染. 由标准差曲线可以看出1—2月的$PM_{2.5}$浓度变化差异最大, 紧接着是4月、11—12月、5—10月$PM_{2.5}$浓度变化差异最小.



(a)$PM_{2.5}$浓度均值曲线 (b)$PM_{2.5}$浓度标准差曲线

图 3 2019年汾渭平原11个城市$PM_{2.5}$浓度均值曲线和标准差曲线

Fig 3 The mean curve and standard deviation curve of $PM_{2.5}$ concentration in 11 cities in Fenwei Plain in 2019

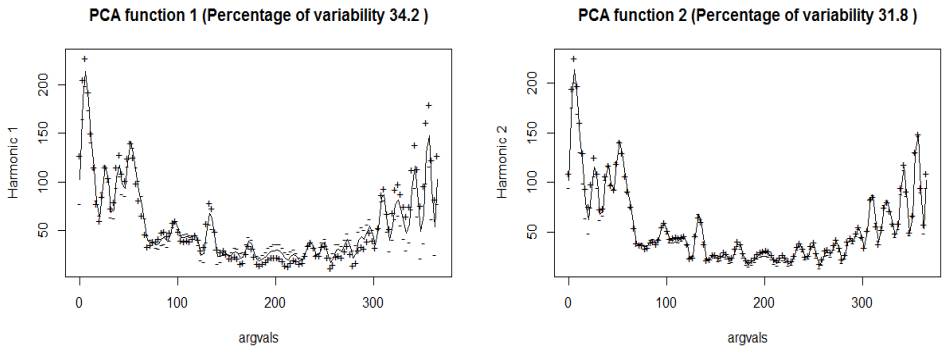
2.3　函数型主成分分析

利用前面所述的函数型主成分分析方法, 实现汾渭平原11个城市的$PM_{2.5}$浓度变化的实证分析. 根据表1的结果显示, 前四个主成分的方差累计贡献率达到96.4%, 对全部数据已经达到相当全面的解释效果, 因此在这里选用前四个主成分来分析汾渭平原$PM_{2.5}$浓度的整体变化模式. 图4为前四个主成分偏离均值的效果图. 实线为11个城市的$PM_{2.5}$浓度变化的均值函数, 图中“+”“−”表示在均值的基础上加、减主成分的常数倍数.

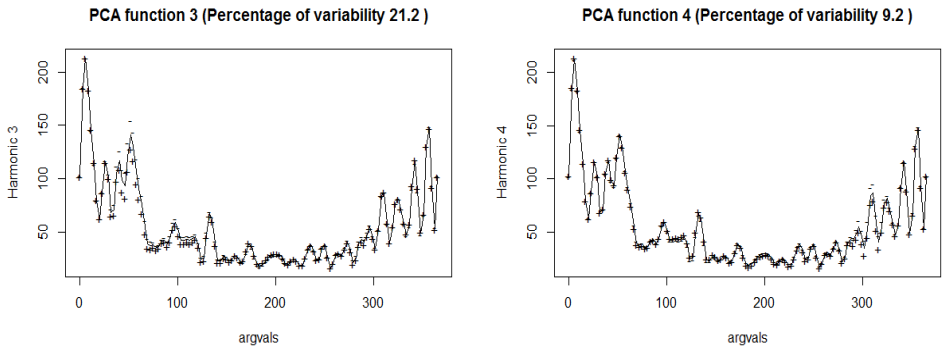
表 1　函数型主成分分析的贡献率和累计贡献率

Tab 1　Contribution rate and cumulative contribution rate of functional principal component analysis

函数型主成分	FPC1	FPC2	FPC3	FPC4
贡献率	0.342	0.318	0.212	0.092
累计贡献率	0.342	0.66	0.872	0.964



(a)第一主成分　　(b)第二主成分



(c)第三主成分　　(d)第四主成分

图 4　前四个函数型主成分权重函数

Fig 4　The first four functional principal component weight functions

从图4可以看出, 第一个函数型主成分显示2019年$PM_{2.5}$浓度曲线在1—2 月、11—12月份的变化显著, 主要描述汾渭平原地区气候温度对$PM_{2.5}$浓度变化的影响. 从汾渭平原采暖期与非采暖期的大气污染状况来看, 由于1月, 11—12 月为采暖期, $PM_{2.5}$浓度明显偏离均值, 故第二主成分主要描述汾渭平原地区采暖期与非采暖期对$PM_{2.5}$浓度变化的影响. 第三个函数型主成分显示2019年$PM_{2.5}$浓度曲线在3—4月份的变化显著, 主要描述汾渭平原地区湿度对$PM_{2.5}$浓度变化的影响. 第四函数型主成分显示2019年$PM_{2.5}$浓度曲线在10月份前后的变化显著, 主要描述汾渭平原地区南北部$PM_{2.5}$浓度变化差异.

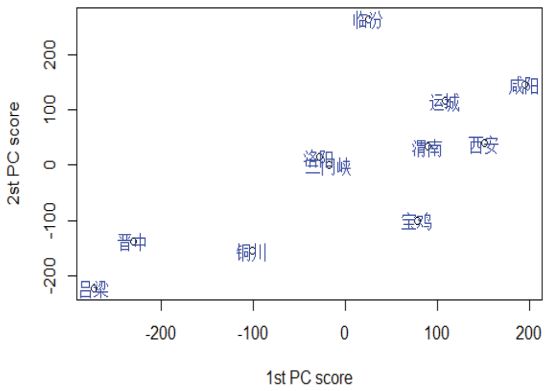


图 5　第一和第二主成分得分图

Fig 5　First and second principal component score plot

图5是函数型主成分分析的第一和第二主成分得分图, 图中左下角区域的第一、第二主成分得分都为负, 而且值都很小, 表明这些地区的$PM_{2.5}$浓度全年相对较低, 是汾渭平原地区空气质量最好的地方; 右上角区域的第一、第二主成分得分都为正, 尤其是咸阳市, 第一、第二主成分得分都很大, 这表明该市的$PM_{2.5}$浓度全年都较高, 是汾渭平原地区空气质量最差的城市, 其次是临汾市, 该市$PM_{2.5}$浓度在采暖期波动最大; 中间区域的第一、第二主成分得分均接近0, 表明这些地方的$PM_{2.5}$浓度接近汾渭平原地区的平均水平, 空气质量居中. 因此, 对汾渭平原地区$PM_{2.5}$浓度曲线的函数型主成分分析, 可以很好地解释$PM_{2.5}$浓度的变化形式, 根据第一、第二主成分得分图, 可将11个城市按照空气质量的不同状况, 由好到差分为三类: 吕梁、晋中、铜川为一类, 洛阳、三门峡、宝鸡、渭南为一类, 运城、西安、临汾、咸阳为一类. 综上所述, 汾渭平原地区的$PM_{2.5}$浓度变化有明显的气候特征和地域特征.

3 结 论

本文对汾渭平原大气污染进行分析, 根据数据高维性、复杂性的特征, 结合函数型分析方法, 采用傅里叶基函数生成$PM_{2.5}$浓度曲线, 运用函数型主成分分析方法, 对汾渭平原地区$PM_{2.5}$浓度数据进行分析, 结果表明:

(1)汾渭平原地区11个城市的$PM_{2.5}$浓度受季节、气候条件影响较大. 冬季的空气质量相对较差, $PM_{2.5}$浓度偏高, 尤其是临汾市最高$PM_{2.5}$浓度达到400 $\mu g/m^3$左右. 故气候温度是导致$PM_{2.5}$浓度差异的首要因素.

(2)汾渭平原地区11个城市的$PM_{2.5}$浓度在采暖期和非采暖期有较大差别. 采暖期的$PM_{2.5}$浓度明显高于非采暖期, 尤其是咸阳市的空气质量, 受供暖影响较大.

(3)由于汾渭平原地理位置的复杂性, 河谷平原的$PM_{2.5}$浓度明显高于两侧山地, 且呈现出向两侧山地递减趋势. 其中运城市和渭南市平原地区的$PM_{2.5}$污染严重, 这样极易形成连片的高污染区域.

(4)大数据时代面对实时监测数据这样庞大的数据集, 函数型数据分析已成为行之有效的数据处理工具. 将空气质量数据函数化, 可以直观展现数据本身的变化, 避免重要信息的丢失, 从而使分析更加全面准确.

参考文献:

[1] 张义学. 汾渭平原坚决打赢蓝天保卫战[J]. 西部大开发, 2019(4): 97-99.
ZHANG Y X. The Fenwei Plain resolutely wins the battle against the blue sky[J]. Western Development, 2019(4): 97-99. (in Chinese)

[2] 楚德见, 金阿芳, 沈广旭. 高层建筑室外颗粒污染物扩散的数值模拟研究[J]. 新疆大学学报(自然科学版), 2018, 35(2): 126-130.
CHU D J, JIN A F, SHENG G X. Numerical simulation of outdoor particulate pollutant dispersion in high-rise buildings[J]. Journal of Xinjiang University(Natural Science Edition), 2018, 35(2): 126-130. (in Chinese)

[3] RAMSAY J O. When the data are functions[J]. Psychometrika, 1982, 47(4): 379-396.

[4] RAMSAY J O, SILVEMAN B. Functional data analysis[M]. New York: Springer, 1997.

[5] RAMSAY J O, BERNARD W, SILVEMAN B. Applied functional data analysis: methods and case studies[M]. New York: Springer, 2002.

[6] 严明义. 函数性数据的统计分析: 思想、方法和应用[J]. 统计研究, 2007(2): 87-94.
YAN M Y. Statistical analysis of functional data: ideas, methods and applications[J]. Statistical Research, 2007(2): 87-94. (in Chinese)

[7] 吴金旺, 顾洲一. 长三角地区数字普惠金融一体化实证分析: 基于函数型主成分分析方法[J]. 武汉金融, 2019(11): 23-28.
WU J W, GU Z Y. An empirical analysis of the integration of digital inclusive finance in the Yangtze River Delta Region: based on the functional principal component analys is method[J]. Wuhan Finance, 2019(11): 23-28. (in Chinese)

[8] 唐裔, 冯长焕. 基于函数型主成分分析的我国城市人口研究[J]. 伊犁师范学院学报（自然科学版）, 2019, 13(3): 9-16.
TANG Y, FENG C H. Research on my country's urban population based on functional principal component analysis[J]. Journal of Yili Normal University(Natural Science Editio), 2019, 13(3): 9-16. (in Chinese)

[9] 梁银双, 刘黎明. 京津冀地区$PM_{2.5}$污染特征的研究: 基于函数型数据分析的视角[J]. 运筹学报, 2018, 22(2): 105-114.
LIANG Y S, LIU L M. Research on the characteristics of $PM_{2.5}$ pollution in the Beijing Tianjin and Hebei Region: based on the perspective of functional data analysis[J]. Journal of Operations Research, 2018, 22(2): 105-114. (in Chinese)

[10] RAMSAY J O, HOOKER G, GRAVES S. Functional data analysis with R and Matlab[M]. New York: Springer, 2009.

责任编辑: 赵新科