Statistics and Application 统计学与应用, 2017, 6(3), 311-319
Published Online August 2017 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2017.63036

Hans汉斯

Multivariate Statistical Analysis and Prediction of PM2.5 Concentration in Xiamen

Xiaofen Xu, Qitong Ou

School of Applied Mathematics, Xiamen University of Technology, Xiamen Fujian
Email: 519051110@qq.com, ouqitong@xmut.edu.cn

Received: Jul. 20th, 2017; accepted: Aug. 5th, 2017; published: Aug. 9th, 2017

Abstract

In this paper, we use the statistical principle to analyze the factors influencing the concentration of PM2.5 in Xiamen daily in 2016, and use statistical software to test whether it has correlation at first. Then we use the method of multiple linear regression analysis to build reasonable mathematical model, to forecast the concentration of PM2.5 in Xiamen during January and February 2017. Finally, reasonable suggestions are put forward according to the prediction results.

Keywords

PM2.5, Multiple Stepwise Regression Analysis, Air Quality Index AQI, SPSS

厦门市PM2.5浓度的多元统计分析与预测

许晓芬，欧启通

厦门理工学院应用数学学院，福建 厦门
Email: 519051110@qq.com, ouqitong@xmut.edu.cn

收稿日期：2017年7月20日；录用日期：2017年8月5日；发布日期：2017年8月9日

摘　要

本文首先运用统计学原理对厦门市2016年全年每日影响PM2.5浓度的因素进行整体分析，并且利用统计软件检验其是否具有相关性，然后采用多元线性回归分析的方法构建合理的数学模型，预测厦门市2017年1月和2月的PM2.5的浓度，最后结合预测结果提出了合理建议。

文章引用: 许晓芬, 欧启通. 厦门市 PM2.5 浓度的多元统计分析与预测[J]. 统计学与应用, 2017, 6(3): 311-319.
DOI: 10.12677/sa.2017.63036

关键词

PM2.5，多元线性回归分析，空气质量指数AQI，SPASS

1. 引言

我国在经济迅速发展的今天，许多环境问题相继而生，因此，进行环境治理势在必行。肖建能[1]结合厦门市土地利用分类专题图和主要重工业企业分布图进行厦门市环境空气质量状况污染源的分析。赵晨曦[2]和吴建南[3]通过对 PM2.5 的研究，分析 PM2.5 与植被和雾霾的影响。本课题对厦门市往年空气中的 PM2.5 浓度进行研究和预测，以便于提前采取有效措施控制 PM2.5 的浓度，这可以为城市规划师和环保局治理空气中的污染成分提供依据。

厦门市未来 PM2.5 的浓度与其它空气质量检测指标 PM10、NO_2、SO_2、CO、O_3 等因素相关，对以上指标做出合理的分析能够帮助我们准确预测 PM2.5 的浓度，更好地对厦门的空气质量进行改善。本文主要解决了以下几个问题：

1) 分析厦门市 2016 年 AQI 里的 6 个基本监测指标的相关性；

2) PM2.5 的浓度变化受其他 5 个指标影响的显著性，建立回归模型，求出 PM2.5 与各变量的相关系数；

3) 利用搜集到的数据，以及对李子奈[4]、何晓群[5]和杨云[6]相关著作的研究，建立回归模型，对 2017 年前两个月 PM2.5 的浓度做线性回归进而完成预测。

2. 数据的收集和处理

为了分析和预测出厦门市 PM2.5 的浓度，我们从天气后报网和厦门市环保局收集到 2016 年厦门市空气质量指数和影响 PM2.5 浓度的其他指标的浓度日均值。利用日均值数据求出月均值如表 1。

为了让结果更加直观的呈现出来，我们对 PM2.5 的月均值数据做了折线图，如图 1。以便于进行整体上的初步分析，观察是否能寻找数据中隐藏的规律。由以下的图 1 容易发现的是其浓度在月份之间的变化范围很大，有明显的季节性变化。

厦门 2016 年 PM2.5 浓度的变化为 5~10 月浓度较低，3 月登顶为 40 μg/m^3，几近年均值的 1.5 倍，3~6 月急剧下降，6 月达最低 16 μg/m^3，低于年均线。6~10 月浓度均值相对较低且相差幅度小，全年 PM2.5 浓度没有明显变化势头，而是受季节影响，夏秋低，春冬高。

3. PM2.5 浓度影响因素的相关性分析

从图 1 中可知 3 月 PM2.5 的浓度全年最高 6 月最低，我们就对 3 月和 6 月的数据进行分析和研究。为了检验这些因素是否能与 PM2.5 建立线性回归模型，根据相关分析方法判断这些因素与 PM2.5 的浓度值是否具有相关性。首先采用皮尔逊相关系数法来检验变量的相关性，运用软件 SPSS 容易获得 3 月份 6 个监测指标的相关系数的数据如表 2。

由表 2 可知，3 月 PM2.5 与空气质量指数 AQI 具有很强的相关性，且各指标中 PM2.5 乃威胁空气质量的重大因素，其相关系数达 0.996，PM2.5 与 PM10、NO_2、SO_2 之间明显有着强相关性，而对于臭氧

Table 1. 2016 Air Quality Index and Monthly Mean of 6 Indicators (μg/m^3)
表 1. 2016 年空气质量指数及 6 项指标的月均值(μg/m^3)

年月	AQI	PM2.5	PM10	NO_2	SO_2	CO	O_3
2016/1	48.35	31.55	47.71	29.42	7.35	0.721	35.97
2016/2	48.07	30.86	46.86	20.62	5.31	0.5952	45.34
2016/3	58.84	40.42	55.61	31.9	8.87	0.7477	52.71
2016/4	52.47	33.53	53.67	31.03	8.9	0.68	49.4
2016/5	44.94	25.58	43.68	20.71	7.52	0.5452	54.39
2016/6	34.17	16.1	32.37	15.07	6.97	0.433	36.17
2016/7	38.65	18.23	37.1	17.39	8.1	0.4161	43.52
2016/8	43.06	22.19	40.65	26	10.16	0.5374	59.65
2016/9	42.4	20.67	37.5	24.13	7.8	0.5937	69.53
2016/10	41.26	19.97	37.84	24.19	7.81	0.5245	54.77
2016/11	52.97	31.3	55.23	30.5	9.77	0.592	52.67
2016/12	57.9	36.52	63	35.61	11.71	0.6571	54.39

Table 2. Pearson correlation coefficients for each indicator in March
表 2. 3 月份各指标的 Pearson 相关系数

	AQI 指数	PM2.5	PM10	NO_2	SO_2	CO	O_3
AQI 指数	1	0.996	0.975	0.567	0.809	0.185	0.173
PM2.5	0.996	1	0.969	0.597	0.81	0.229	0.137
PM10	0.975	0.969	1	0.527	0.835	0.049	0.259
NO_2	0.567	0.597	0.527	1	0.712	0.559	−0.546
SO_2	0.809	0.81	0.835	0.712	1	0.198	−0.021
CO	0.185	0.229	0.049	0.559	0.198	1	−0.607
O_3	0.173	0.137	0.259	−0.546	−0.21	−0.607	1

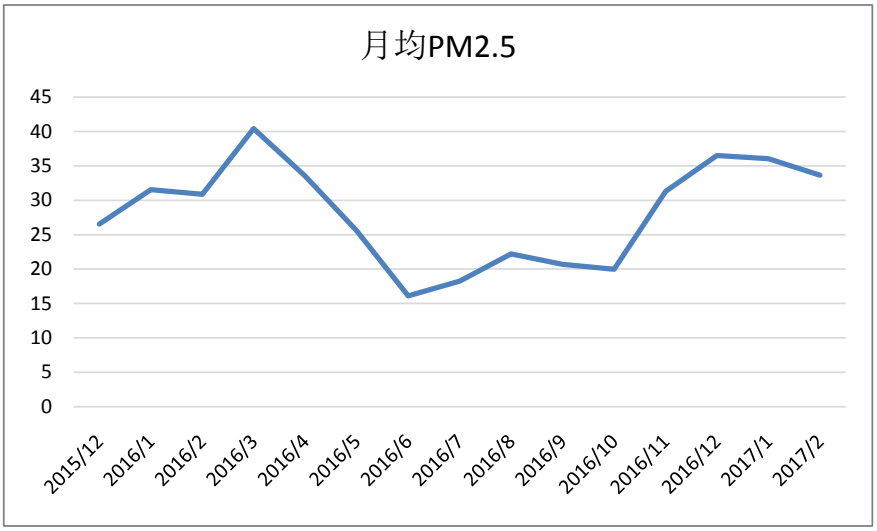

Figure 1. The monthly mean change of PM2.5 concentration
图 1. PM2.5 浓度的月均值变化线图(μg/m^3)

和 CO 则相关性非常弱，不同的污染物浓度之间的相关性分析可以作为构建 PM2.5 浓度的预测模型提供有力的理论支撑。

基于 2016 年 6 月份的污染物浓度监测数据进行分析，运用软件 SPSS 容易获得 6 个监测指标的相关系数的数据如表 3。

由表 3 可以看出，6 月份 PM10 成为 AQI 的主要影响成分，PM2.5 与每个指标都具有强相关性，与表 2 所分析出的内容产生了一定的矛盾。由此，受季节性变化和诸多未知因素的共同影响，PM2.5 浓度在月份之间的波动范围甚大，单纯对某一月份进行分析会造成很大的误差。

为了减少外界因素对结果造成的误差，我们不采用某些月份来研究，而是基于 2016 年 366 天的污染物浓度监测数据进行分析，用软件 SPSS 得 6 个监测指标的相关系数如表 4。

由表 4 可以看出，除了 O_3 与 PM2.5 的相关性不算太强以外，PM2.5 与其他指标都显示有强相关性。

为了提高相关分析结果的准确性，本文接下来运用另一种分析方法——方差分析。从观测变量的方差着手，研讨各种控制变量中多少变量是对观测变量起到明显作用的变量。运用 SPSS 软件可以得出如图 2 所示的几个单因素方差分析表的缩略图。

由图 2 可以得出以下结论：PM2.5 与 PM10、NO_2、SO_2、CO 和 O_3 方差分析的 P 值均小于系统设定 0.05 的显著性水平，表明以上因素对厦门市大气中 PM2.5 浓度值的变化存在显著性，所研究变量的相关性检验通过。

4. 多元线性回归模型

综合以上相关分析的结果，对于厦门市 2016 年影响 PM2.5 浓度的要素分析，主要考虑以下 5 个因素 PM10、NO_2、SO_2、CO、O_3 的影响，且已满足多元线性回归模型的条件，只要数据准确无误，便可以着手构建模型。

多元线性回归模型，通常形式为：

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \mu_i, i = 1,2,\cdots,n$$

其中 y_i 即 PM2.5，x_1 为自变量 PM10，x_2 为 NO_2，x_3 是 SO_2，x_4 为 CO，x_5 为 O_3，β_i $(i = 1,2,\cdots,n)$称为回归系数，可称上式为总体回归函数的随机表达式。

将搜集到的数据代入回归模型计算变量的回归系数，运行 SPSS 软件得到如下结果：

表 5 中 B 代表各项自变量的系数，系数 B 的置信区间中，变量 x_2 (NO_2)过了 0 点，说明其系数取值在正负之间摆动，没达到预测效果，且表中 NO_2 的 P 值大于显著性水平 0.05，故该变量不适合作回归分析。在满足构建模型的情况下，造成以上结果很可能源于数据存在异常值或者异方差性。

首先我们进行的是检验数据是否存在异常值。回归分析中，搜集的数据常出现一些异常或极端的值，极大地影响拟合效果，一元线性回归可用简单的残差图或散点图来识别，本文的多元线性回归采用库克距离(D_i)来识别，判别标准为：若 D_i < 0.5 时，是非异常值点；D_i > 1 便是异常值点。用软件计算结果如下图：

由图 3 可以看出第 53 行数据的库克距离(表中 COO_1 列)大于 1，我们将这行异常数据剔除后再计算一次库克距离后的结果显示库克距离全部小于 1，那么异常值已经剔除了，可以进行拟合，拟合结果如表 6。

由表 6 可知变量 x_2 (NO_2)系数的 P 值为 0.747，剔除异常值后依然没有通过显著性检验，所以对回归模型计算出错的原因不是因为数据存在异常值。

接下来，本文改斯皮尔曼等级相关系数法来检验数据的异方差性，其应用更广泛，可用于线性也可

Table 3. Pearson correlation coefficients for each index in June
表 3. 6 月份各指标的 Pearson 相关系数

	AQI 指数	PM2.5	PM10	NO_2	SO_2	CO	O_3
AQI 指数	1	0.957	0.962	0.468	0.307	0.479	0.618
PM2.5	0.957	1	0.918	0.417	0.17	0.459	0.663
PM10	0.962	0.918	1	0.383	0.325	0.438	0.448
NO_2	0.468	0.417	0.383	1	0.602	0.614	0.355
SO_2	0.307	0.17	0.325	0.602	1	0.459	−0.136
CO	0.479	0.459	0.438	0.614	0.459	1	0.188
O_3	0.618	0.663	0.448	0.355	−0.136	0.188	1

Table 4. Pearson correlation coefficients for each indicator in 2016
表 4. 2016 年各指标的 Pearson 相关系数

	PM2.5	PM10	NO_2	SO_2	CO	O_3	AQI
PM2.5	1	0.934	0.585	0.602	0.623	0.26	0.959
PM10	0.934	1	0.539	0.673	0.475	0.34	0.972
NO_2	0.585	0.539	1	0.682	0.655	−0.04	0.546
SO_2	0.602	0.673	0.682	1	0.376	0.212	0.648
CO	0.603	0.475	0.655	0.376	1	0.07	0.546
O3	0.26	0.34	−0.04	0.212	0.07	1	0.404
AQI 指数	0.959	0.972	0.546	0.648	0.546	0.404	1

AQI指数与PM2.5的单因素方差分析

	平方和	df	均方	F	显著性
组间	107778.3	58	1858.25	69.783	0
组内	8175.038	307	26.629		
总数	115953.3	365			

PM2.5与CO的单因素方差分析

	平方和	df	均方	F	显著性
组间	4.886	58	0.084	5.522	0
组内	4.683	307	0.015		
总数	9.569	365			

PM10与PM2.5的单因素方差分析

	平方和	df	均方	F	显著性
组间	152817.4	58	2634.78	47.302	0
组内	17100.41	307	55.702		
总数	169917.8	365			

PM2.5与O_3的单因素方差分析

	平方和	df	均方	F	显著性
组间	33265.84	58	573.55	1.865	0
组内	94395.29	307	307.48		
总数	127661.1	365			

PM2.5与NO_2的单因素方差分析

	平方和	df	均方	F	显著性
组间	21342.79	58	367.979	5.442	0
组内	20758.41	307	67.617		
总数	42101.2	365			

PM2.5与CO的单因素方差分析

	平方和	df	均方	F	显著性
组间	4.886	58	0.084	5.522	0
组内	4.683	307	0.015		
总数	9.569	365			

Figure 2. Single factor analysis of variance
图 2. 单因素方差分析结果

Table 5. Calculation of regression model coefficients
表 5. 回归模型系数计算

	B	标准误差	标准系数	t	Sig.
(常量)	−9.446	1.031		−9.158	0
PM10	0.615	0.016	0.888	39.399	0
NO_2	0.015	0.037	0.01	0.391	0.696
SO_2	−0.306	0.096	−0.08	−3.204	0.001
CO	20.899	1.951	0.226	10.713	0
O_3	−0.033	0.014	−0.041	−2.422	0.016

Table 6. Calculate the regression model coefficients after rejecting the outliers
表 6. 剔除异常值后的回归模型系数计算

	B	标准误差	标准系数	t	Sig.
(常量)	−9.11	0.942		−9.674	0
PM10	0.653	0.015	0.927	43.804	0
NO_2	−0.011	0.034	−0.008	−0.322	0.747
SO_2	−0.394	0.088	−0.103	−4.49	0
CO	20.44	1.78	0.222	11.482	0
O_3	−0.039	0.012	−0.049	−3.148	0.002

	A	B	C	D	E	F	G	H	I
1	日期	质量等级	PM2.5	PM10	No2	So2	Co	O3	COO_1
50	2016/2/18	优	22	31	18	4	0.44	61	0.00356
51	2016/2/19	良	36	42	32	6	0.62	34	0.00783
52	2016/2/20	良	39	47	22	5	0.67	49	0.00518
53	2016/2/21	良	44	126	13	7	0.58	69	1.10092
54	2016/2/22	良	44	99	32	7	0.64	38	0.16436
55	2016/2/23	优	31	34	18	2	0.63	37	0.00811

Figure 3. Cook distance
图 3. 库克距离

用于非线性的情况。使用软件对已有数据进行 Spearman 测验，结果如表 7。

由表 7 可知变量 PM10、O_3 与残差绝对值(abs)都出现显著相关现象，表明数据存在异方差。

当一个回归问题有了异方差性时，普通最小二乘估计的有效性已被破坏。对于异方差的修正我们用加权最小二乘估计法，再次利用斯皮尔曼相关系数法检验出与 PM2.5 具有强相关性的自变量为 PM10，我们令 PM10 做权变量，利用加权最小二乘估计，结果如表 8。

从表 8 能够直观地看出各个变量的 P 值都是 0，小于显著性水平 0.05，表明能够继续对数据进行拟合。综上可得线性回归模型：

$$y_i = -9.579 + 0.617x_1 + 0.016x_2 - 0.316x_3 + 20.974x_4 - 0.031x_5.$$

最后，为了进一步确认上述加权回归模型的可靠性，我们还需要对模型进行显著性测验，借助软件可以得到结果如表 9，表 10。

由表 9，表 10 的显示可得出以下结论：

Table 7. Spearman rank correlation coefficients
表 7. Spearman 等级相关系数

		PM10	NO_2	SO_2	CO	O_3	残差绝对值
PM10	相关系数	1.000	0.511**	0.677**	0.493**	0.423**	0.212**
	Sig.		0.000	0.000	0.000	0.000	0.000
NO_2	相关系数	0.511**	1.000	0.637**	0.667**	0.029	−0.008
	Sig.	0.000		0.000	0.000	0.574	0.881
SO_2	相关系数	0.677**	0.637**	1.000	0.397**	0.251**	−0.001
	Sig.	0.000	0.000		0.000	0.000	0.987
CO	相关系数	0.493**	0.667**	0.397**	1.000	0.132*	0.070
	Sig.	0.000	0.000	0.000		0.012	0.182
O_3	相关系数	0.423**	0.029	0.251**	0.132*	1.000	0.148**
	Sig.	0.000	0.574	0.000	0.012		0.004
残差绝对值	相关系数	0.212**	−0.008	−0.001	0.070	0.148**	1.000
	Sig.	0.000	0.881	0.987	0.182	0.004	

Table 8. Weighted least squares estimation regression coefficients
表 8. 加权最小二乘估计回归系数

	B	标准误差	t	Sig.
(常数)	−9.579	0.154	−62.312	0
PM10	0.617	0.002	315.369	0
NO_2	0.016	0.004	3.656	0
SO_2	−0.316	0.009	−36.939	0
CO	20.974	0.204	103.04	0
O3	−0.031	0.002	−13.79	0

Table 9. The test of R^2
表 9. R^2 检验

复相关系数	1.000
R方	0.999
估计的标准误	4.320

Table 10. The variance homogeneity of F test
表 10. 方差齐性 F 检验

	平方和	df	均方	F	Sig.
回归	8260575.188	5	1652115.038	88540.663	0.000
残差	6717.382	360	18.659		
总计	8267292.569	365			

1) 在拟合前，由统计软件可计算得 $R^2 = 0.918$，表 9 显示拟合之后 $R^2 = 0.999$，相差甚大，可见自变量与因变量之间的相关性极高，说明模型拟合比原模型好；

2) 方差齐性分析中的计算概率为 0，小于显著性水平 0.05，说明上述加权多元回归模型有效。

Table 11. The comparison of PM2.5 actual value and the predicted value
表 11. PM2.5 实际值与预测值的对比

日期	PM2.5 的实测值	PM2.5 预测值
2017/1/1	24	27
2017/1/2	53	53
……	……	……
2017/2/28	60	50

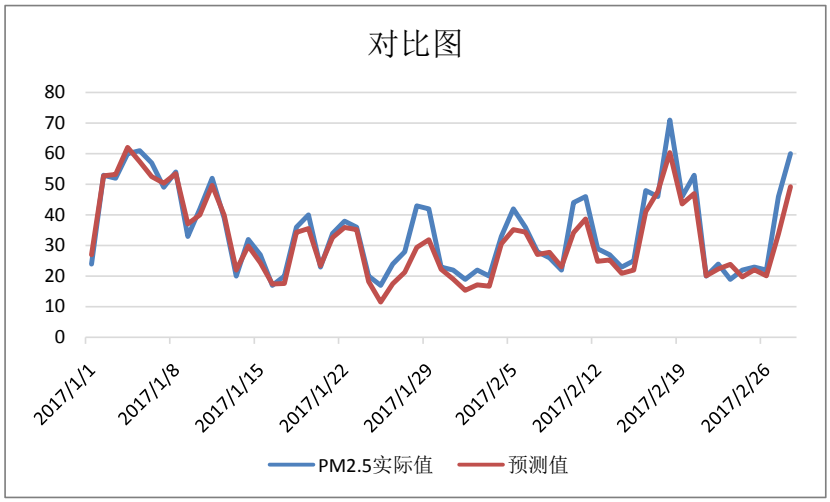

Figure 4. Comparison of actual and predicted values
图 4. 实际值与预测值对比图

5. PM2.5 浓度的预测

本论文最重要的目的就是预测 PM2.5 的浓度，利用已有数据和已经建立的模型，预测 2017 年 1 月~2 月的 PM2.5 浓度，结果如下：

由表 11 可知 PM2.5 的预测值与真实值，可能有几个稍有偏离的数据，整体上不会有太大区别，为保证模型和方法的合理性，并且更直观地对两者进行比较，我们通过线性拟合来判读计算出的数据是否正确、合理。

由图 4 可见拟合效果还算不错，预测值与实测值的数据相对统一，预测值为峰值时实际值也几乎达到峰值，说明预测结果与实际值的变化趋势相类似，故也又深入一层检验本文运用模型的准确性，对现今大气中 PM2.5 的预测具备极好的实践意义。

6. 结论

从 2017 年 1 月和 2 月厦门市 PM2.5 浓度的预测值中可以看出其浓度并未呈现稳定的上升还是降低的趋势，一样随着季节的变化在波动，可知 PM2.5 浓度的变化是一个非平稳的具有周期变化的时间序列。但是厦门的地理位置优越，空气质量的级别都属优良，不会影响到人类的正常户外活动。但 PM2.5 是严重影响空气质量的成分，对 PM2.5 的预测对于空气治理和人类的健康生活都具有很大的促进作用。PM2.5 的治理除了控制相关影响因素的浓度以外，最大的治理方式就是控制其污染源，而 PM2.5 的来源主要是机动车的直接和间接排放、煤炭的污染、工业的喷涂、城市的扬尘污染、农村的秸秆焚烧、还有燃料的燃烧等人为活动。对于其治理有如下几个意见：

1) 积极响应国家“去产能”的号召，淘汰掉落后的产能。
2) 推动各产业的结构化调整和技术创新，对扬尘和工业污染进行综合治理。
3) 严格施行单双号车限行制度，增加大排气量车购置税，控制机动车尾气排放。
4) 对厦门不同区域的植被结构进行不同设置，合理高效地利用生态绿化消除 PM2.5.
5) 根据不同阶段的气候特征对 PM2.5 采取不同的治理措施。

参考文献 (References)

[1] 肖建能, 杜国明, 施益强, 等. 厦门市环境空气污染时空特征及其与气象因素相关分析[J]. 环境科学学报, 2016(9): 3363-3371.

[2] 赵晨曦, 王玉杰, 王云琦, 等. 细颗粒物(PM_(2.5))与植被关系的研究综述[J]. 生态学杂志, 2013(8): 2203-2210.

[3] 吴建南, 秦朝, 张攀. 雾霾污染的影响因素: 基于中国监测城市 PM2.5 浓度的实证研究[J]. 行政论坛, 2016(1): 62-66.

[4] 李子奈, 潘文卿. 计量经济学(第三版) [M]. 北京: 高等教育出版社, 2006.

[5] 何晓群, 刘文卿.应用回归分析(第四版) [M]. 北京: 中国人民大学出版社, 2015.

[6] 杨云, 付彦丽. 关于空气中 PM2.5 质量浓度预测研究[J]. 计算机仿真, 2016(3): 413-418.