

基于多元线性回归模型的西安市空气质量研究

张馨心

西安建筑科技大学信息与控制工程学院

DOI:10.12238/eep.v6i4.1824

[摘 要] 本文主要通过对陕西省西安市的空气质量六项分指标进行相关与独立性分析、建立PM2.5时空分布模型,来对陕西省西安市的空气质量进行研究。对于问题一: 本文使用皮尔逊相关性分析,对空气质量各项分指标的相关性进行研究,得出各分指标两两之间的相关系数,再采用了独立性检验来探究各项指标之间的独立性。其次建立了多元线性回归模型对PM2.5与其他五项分指标及其对应污染物之间的相关性及其关系进行深入研究。得到PM2.5与其他五项指标的得回归平面方程: $y = -32.9559 + 0.1733x_1 + 0.4417x_2 + 0.8820x_3 + 2.2071x_4 - 0.3339x_5$ 再通过残差分析,得出本模型能够较为完整的反应原始数据。对于问题二: 首先,根据陕西省西安市监测点的数据建立PM2.5时空分布模型,绘出时空分布仿真图,从图中,本文发现陕西省西安市的PM2.5分布呈现“冬多春少,北少南多”的规律。其次基于多元线性回归模型,建立小寨和阎良区的回归平面方程,通过未来十天其他五项指标的数据预测出小寨和阎良区PM2.5指数。

[关键词] 空气质量; 皮尔逊; 多元线性回归模型; PM2.5

中图分类号: R122.1 文献标识码: A

Study on Air Quality in Xi'an City Based on Multiple Linear Regression Model

Xinxin Zhang

School of Information and Control Engineering, Xi'an University of Architecture and Technology

[Abstract] This paper mainly studies the air quality of Xi'an City, Shaanxi Province through correlation and independence analysis of six sub-indexes of air quality and establishment of the spatial and temporal distribution model of PM2.5. As for question one, Pearson correlation analysis is used in this paper to study the correlation of each sub-index of air quality, and the correlation coefficient between each sub-index is obtained. Then independence test is adopted to explore the independence of each index. Secondly, a multiple linear regression model was established to further study the correlation and relationship between PM2.5 and other five sub-indexes and their corresponding pollutants. The regression plane equation of PM2.5 and other five indicators is obtained: $y = -32.9559 + 0.1733x_1 + 0.4417x_2 + 0.8820x_3 + 2.2071x_4 - 0.3339x_5$ Through the residual analysis, it is concluded that this model can reflect the original data more completely. As for question two: First, the spatial and temporal distribution model of PM2.5 is established according to the data of the monitoring points in Xi'an, Shaanxi Province, and the simulation diagram of spatial and temporal distribution is drawn. From the diagram, this paper finds that the distribution of PM2.5 in Xi'an City, Shaanxi Province presents a rule of "more in winter and less in spring, less in north and more in south". Secondly, based on the multiple linear regression model, the regression plane equation of Xiaozhai and Yanliang District is established, and the PM2.5 index of Xiaozhai and Yanliang District is predicted through the data of other five indicators in the next ten days.

[Key words] air quality; Pearson; multiple linear regression model; PM2.5

1 背景与问题

1.1 研究背景

大气为地球上的生命繁衍和人类的发展提供了理想的环境,成为了人类生产、生活以及生存密不可分的一环。但随着全球

社会的工业化发展,空气问题成为了政府、环境保护部门乃至全国人民关注的热点问题。良好的空气质量是人类美好生活的前提,了解空气质量背后的成因、研究空气质量相关的各项因素成为了解决空气质量问题的重中之重。

1.2问题的提出

问题一:根据陕西省西安市空气质量监测点的数据,建立适当的数学模型,对AQI中6个基本监测指标的相关与独立性进行定量分析,其中对PM2.5(含量)与其它5项分指标及其对应污染物(含量)之间的相关性及其关系进行分析。

问题二:根据陕西省西安市空气质量的监测数据,描述西安地区内PM2.5的时空分布及其规律。并且利用建立的模型,结合预测方法,预测未来十天小寨和阎良区PM2.5指数(即2021年4月17日-2021年4月26日)。

2 问题分析

2.1问题一的分析

问题一要求选取合适的数学模型,并且结合陕西省西安市空气质量的监测数据,定量的分析AQI中6个基本监测指标的相关与独立性,特别地,就PM2.5与其它5项分指标及其对应污染物之间的相关性及其关系进行分析。本文的求解思路如下:首先本文通过对数据的预处理,证明了各项指标间呈现线性关系且正态分布。其次为了研究各项指标间的相关性,采用了皮尔逊相关性分析。然后采用了独立性检验来探究各项指标之间是否相互独立。最后考虑到线性回归模型可以准确衡量各因素间的相关程度与拟合程度的高低,并参考文献^[1],所以选择多元线性回归模型作为此问题的研究模型,对PM2.5与其他五项分指标及其对应污染物之间的相关性及其关系进行了深入探究。

2.2问题二的分析

问题二要求根据陕西省西安市监测点的数据,对西安地区内PM2.5的时空分布及其规律进行描述。再利用建立的模型,结合相应的预测方法,预测未来十天小寨和阎良区PM2.5指数(2021年4月17日-2021年4月26日)。最后要求计算纺织城的AQI的数值。本文的求解思路如下:首先就陕西省西安市监测点的数据建立了时间和空间分布两个模型,绘出时空分布仿真图,阐述其规律。然后运用问题一所建立的多元线性回归模型,通过未来十天其他五项指标的数据预测出小寨和阎良区PM2.5指数。

3 模型假设

(1)假设所有数据真实可靠(2)问题一中PM2.5含量只与附件一其余五项检测指标有关系。(3)各项空气污染物之间不会相互作用。(4)六个监测点的PM2.5浓度值平均值可以代表整个陕西省西安市的PM2.5浓度值。

4 模型的建立和求解

4.1问题一的模型建立和求解

4.1.1数据预处理

(1)首先对拿到的数据进行检查发现有空缺数据,利用Matlab对空缺数据进行线性插值补齐,对离群数据进行删除。(b)其次,将更新后数据转化为散点图。

4.1.2各项指标之间的相关性

为了对这六项指标进行相关性分析,绘出如图所示六项指标两两分布散点图,并且本文发现各项指标均成正态分布。

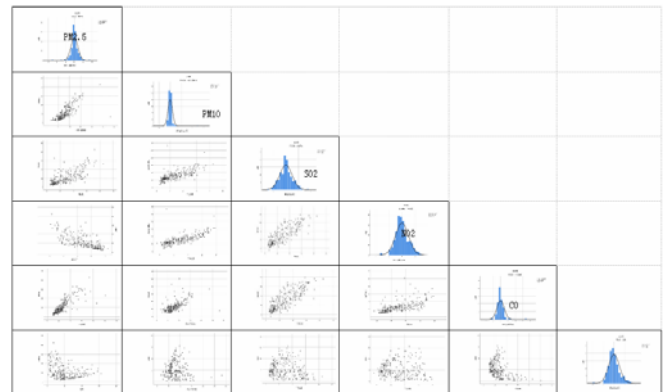


图1 各项指标散点图及正态分布图

本文采用皮尔逊相关性分析(公式如下):

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (1)$$

得出了六项指标两两之间的相关系数,如下表所示:

表1 各分指标两两相关系数表

相关系数	二氧化硫	PM10	二氧化氮	一氧化碳	臭氧	PM2.5
二氧化硫	1.000	0.685	0.783	0.647	-0.341	0.734
PM10	0.685	1.000	0.766	0.627	-0.191	0.834
二氧化氮	0.783	0.766	1.000	0.597	-0.187	0.731
一氧化碳	0.647	0.627	0.597	1.000	-0.471	0.820
臭氧	-0.341	-0.191	-0.187	-0.471	1.000	-0.462
PM2.5	0.734	0.834	0.731	0.820	-0.462	1.000

表2 独立性检验表

监测指标	Prob
SO ₂ *NO ₂	0.000
SO ₂ *PM10	0.000
SO ₂ *CO	0.000
SO ₂ *O ₃	0.006
SO ₂ *PM2.5	0.000
NO ₂ *PM10	0.000
NO ₂ *CO	0.000
NO ₂ *O ₃	0.336
NO ₂ *PM2.5	0.000
PM10*CO	0.000
PM10*O ₃	0.308
PM10*PM2.5	0.000
CO*O ₃	0.000
CO*PM2.5	0.000
O ₃ *PM2.5	0.000

从该图中可以发现PM2.5与其它各指标具有较高的相关性,其中以PM10的相关性最高,其次是CO。且本文发现SO₂、PM10、NO₂和CO均与PM2.5浓度呈正相关,O₃与PM2.5呈负相关性。

4.1.3 独立性检验

利用SPSS软件对该市空气质量监测数据进行独立性检验,检验结果如表2。若显著性概率(Prob)小于0.05,则落入拒绝域,接受备择假设,即两个因素之间不是相互独立的,而是具有一定相关关系的。

由表2结果分析可知,所选指标中的NO₂与O₃、PM10和O₃的显著性概率未落入拒绝域内,说明NO₂与O₃、PM10和O₃不相关、相互独立,而其他的两两相关且不独立。

4.1.4 基于多元线性回归模型

本文通过Matlab对PM2.5和其它五项指标(数据预处理后)刻画散点图(如图(1)),证明PM2.5和其他五项呈线性相关关系。

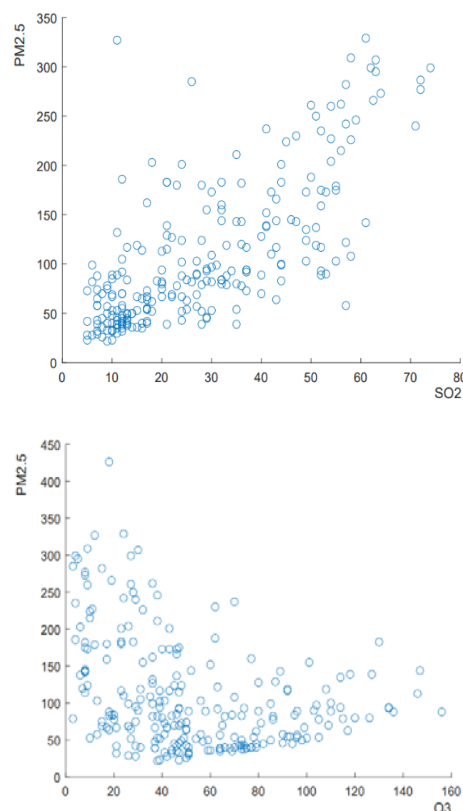
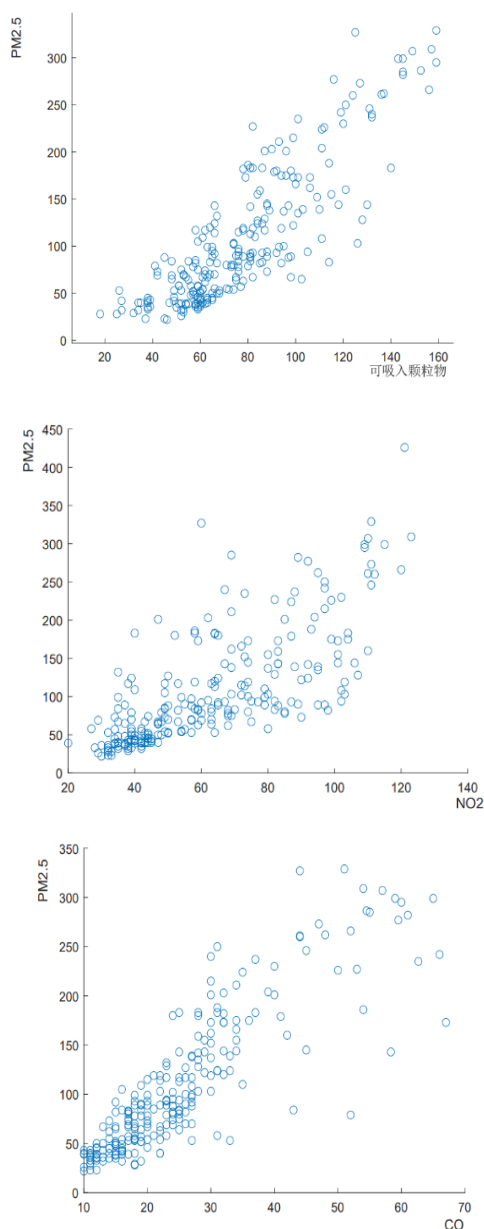


图2 PM2.5和其他五项分指标散点图

所以本文假设它们之间的线性关系式如式(1)所示:

$$y = b_0 + b_1x_1 + \dots + b_kx_k + \varepsilon_i, i = 1, 2, \dots, n \quad (2)$$

其中,本文定义 y 为因变量, k 为解释变量的数目, b 为回归系数, x_i 为非随机的自变量, ε_i 为随机变量, $\varepsilon_1 \dots \varepsilon_n$ 互不相关且均是同分布的随机变量,可用矩阵表示为:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \ddots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, b = \begin{bmatrix} b_0 \\ \vdots \\ b_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (3)$$

于是可将(1)式转化为:

$$y = XB + \varepsilon_i \quad (4)$$

其中 X 为已知的 $n \times (k+1)$ 阶回归设计矩阵, b 为 $k+1$ 维位置的列向量,本文对(1)取期望,可得回归平面方程如下:

$$y = b_0 + b_1x_1 + \dots + b_kx_k \quad (5)$$

4.1.5 PM2.5与其它5项指标间的相关性分析

基于上述所建模型,使用式(4),用Matlab求解模型,可得回归平面方程如下:

$$y = -32.9559 + 0.1733x_1 + 0.4417x_2 + 0.8820x_3 + 2.2071x_4 - 0.3339x_5 \quad (6)$$

其中, y 代表PM2.5的含量, x_i 代表另外五项指标。

此方程的实际意义指: PM2.5的含量跟其他5项指标的含量的关系。

本文对式(5)进行了残差分析, 其中-32.9559, 0.1733, 0.4417, 0.8820, 2.2071, 0.3339为点估计值, 对置信区间为: [-47.0906, -18.8213]、[-0.2300, 0.5765]、[0.1369, 0.7466]、[0.7005, 1.0634]、[1.7865, 2.6278]、[-0.4642, -0.2035], 相关系数为0.8366, 说明回归显著, $F=235.5429$, $p=0.0000$ 小于显著性水平0.05, 可知回归方程成立。

利用Matlab本文绘出回归残差图(如图3所示), 从中发现所有的残差点离零点都较为接近, 而且零点也在置信区间内, 说明本模型能够较为完整地反映原始数据。

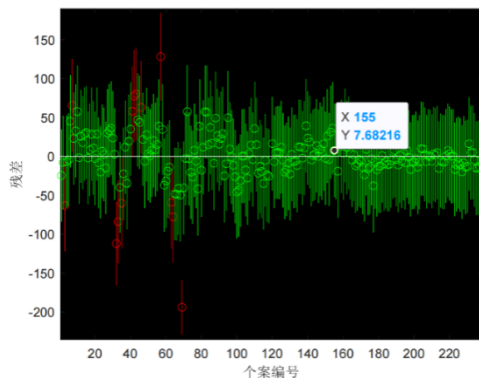


图3 回归残差图

4.2 问题二的模型建立和求解

4.2.1 数据预处理

本问中的完全缺失的数据进行剔除, 缺失部分的数据本文利用Matlab选取保形三次插值法(PCHIP)补全适当的数据。

4.2.2 预测小寨和阎良区PM2.5指数——基于多元线性回归模型

对于小寨和阎良区的预测算法研究中, 本文采取第一问建立的多元线性回归模型中的回归平面方程(4), 利用Matlab求解模型, 可以得到小寨和阎良区各自的回归平面方程如下:

$$y_1 = -65.8660 + 1.3825x_{11} - 0.6067x_{12} + 0.5321x_{13} + 1.0045x_{14} - 2.2707x_{15} + 2.0822x_{16}$$

$$y_2 = -16.5305 + 0.6141x_{21} - 0.2919x_{22} + 0.6822x_{23} + 1.3021x_{24} - 0.9426x_{25} + 0.4547x_{26}$$

其中 y_1 表示小寨的PM2.5的浓度、 y_2 代表阎良区的PM2.5

的浓度, x_{1i} 、 x_{2i} 分别代表了小寨和阎良区的其他六项指标的浓

度值。其中本文分别表示了臭氧在1小时平均和8小时滑动平均的含量。

表3 小寨PM2.5浓度值和首要污染物

点位名称	SO ₂	NO ₂	PM ₁₀	CO	O ₃ 1小时	O ₃ 8小时	PM _{2.5}	首要污染物	时间
小寨	40	117	151	42	31	41	56	中度污染	2021-4-26
小寨	38	125	169	45	31	40	59	中度污染	2021-4-25
小寨	31	108	152	60	23	30	63	中度污染	2021-4-24
小寨	30	80	132	77	24	28	78	轻度污染	2021-4-23
小寨	28	52	142	84	9	11	104	轻度污染	2021-4-22

运用此模型, 预测出了未来十天(即4月17日至4月26日)小寨和阎良区的PM2.5的浓度值和首要污染物。

表3 小寨PM2.5浓度值和首要污染物

点位名称	SO ₂	NO ₂	PM ₁₀	CO	O ₃ 1小时	O ₃ 8小时	PM _{2.5}	首要污染物	时间
小寨	44	55	113	76	18	23	105	轻度污染	2021-4-21
小寨	28	40	81	76	13	18	76	良	2021-4-20
小寨	18	37	112	74	16	22	80	轻度污染	2021-4-19
小寨	36	45	198	71	26	36	149	中度污染	2021-4-18
小寨	48	78	149	84	40	54	138	轻度污染	2021-4-17

表4 阎良区PM2.5浓度值和首要污染物

点位名称	SO ₂	NO ₂	PM ₁₀	CO	O ₃ 1小时	O ₃ 8小时	PM _{2.5}	首要污染物	时间
阎良区	51	54	90	35	21	30	100	良	2021-4-26
阎良区	41	52	87	34	18	25	91	良	2021-4-25
阎良区	36	50	142	50	21	32	148	轻度污染	2021-4-24
阎良区	NA	43	163	51	36	NA	150	中度污染	2021-4-23
阎良区	NA	23	155	40	NA	NA	135	中度污染	2021-4-22
阎良区	17	22	105	40	31	48	103	轻度污染	2021-4-21
阎良区	NA	20	82	28	25	37	74	良	2021-4-20
阎良区	NA	19	NA	28	21	32	129	中度污染	2021-4-19
阎良区	22	24	233	29	22	32	181	重度污染	2021-4-18
阎良区	41	33	181	34	31	43	157	中度污染	2021-4-17

为了描述首要污染物, 本文摘取了附件4中适当的内容自制列表, 方便观看和查找, 如下表所示:

表5 AQI空气指数类别表

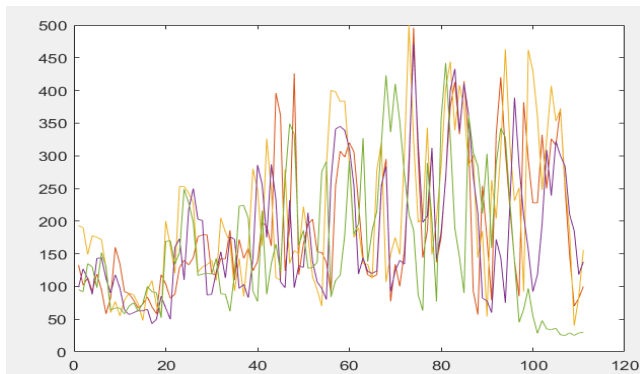
空气质量指数	0~50	51~100	101~150	151~200	201~300	>300
空气质量指数类别	优	良	轻度污染	中度污染	重度污染	严重污染

4.2.3 时间分布模型建立

本题要描述PM_{2.5}的时间分布, 本文将更新过的陕西省西安市监测点数据就PM_{2.5}和时间一栏, 利用Matlab绘出2021年1月1日至2021年4月26日, 纺织城、小寨、高新西区、长安区、阎良区以及临潼区的PM_{2.5}浓度变化曲线图如图4所示。

分析图4, 从这6个地区的111天的PM_{2.5}浓度变化可以发现, 处于一月份至二月中旬的PM_{2.5}浓度较高, 而处于三月末至四月的PM_{2.5}浓度显著降低。

由于气象条件和污染来源等的季节性差异, 陕西省西安市PM_{2.5}浓度整体呈“冬多春少”的季节性变化规律。

图4 2021年4月26日至2021年1月1日PM_{2.5}浓度变化

4.2.4 空间分布模型建立

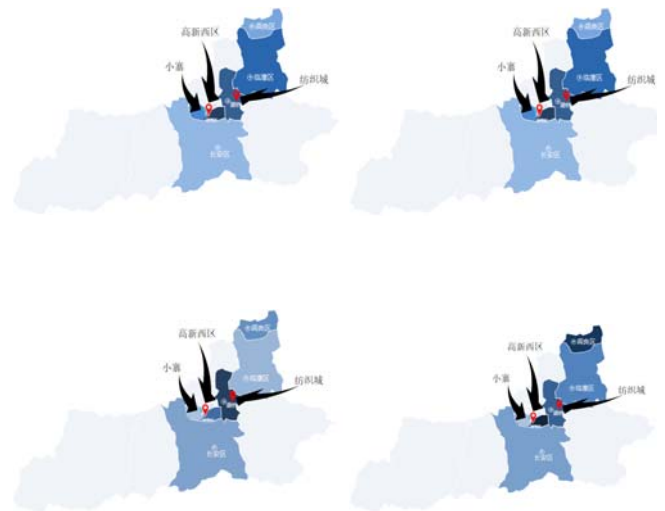
参考文献^[2-3], 本文将给出的陕西省西安市六个地点的监测数据进行整理归纳, 并且给出了这六个地点及其所属区域的经纬度表, 如表6所示。

表6 监测点经纬度分布

站点	所属区	经纬度	
纺织城	灞桥区	109.06 E	34.26° N
小寨	雁塔区	108.94 E	34.23° N
高新西区	雁塔区	108.88 E	34.23° N
长安区	长安区	108.91 E	34.16° N
阎良区	阎良区	109.2° E	34.66° N
临潼区	临潼区	109.22 E	34.37° N

为了更直观的表现空间分布规律, 按照所查的经纬度, 本文将这六个监测点的位置标注在西安地图上, 以每个月(一月份到四月份)的情况作图, PM_{2.5}的浓度以颜色深浅作为标定, 如图5所示。

通过对图5的分析本文可以看出每个月的小寨、高新西区、纺织城、临潼区、阎良区的PM_{2.5}的浓度都显著高于长安区, 说明西安的PM_{2.5}浓度呈现“北少南多”的规律。

图5 PM_{2.5}浓度1月到4月在六个监测点的分布图

5 模型的评价

5.1 模型的优点

多元线性回归模型是多元回归模型中最基础的, 可以研究多个自变量对一个因变量的影响, 其思想简单, 容易实现, 较为典型, 可操作性强, 有直观的理解和解释, 对异常值较为敏感。

5.2 模型的缺点

(1) 建模过程中, 只考虑了六种检测指标对空气质量带来的影响, 导致对于空气质量的研究并不精确。(2) 多元线性回归模型需要严格的假设, 对于情况较为负载, 外界因素较多的情况并不适用。

6 模型的推广和应用

多元线性回归模型可用于可以用来进行建模和预测, 在参加经济、生态、环境等实际课题的过程中, 感到多元回归分析是一个最广泛应用的统计方法, 在数据处理过程中能发挥有效的作用^[4]。

[参考文献]

- [1] 范天星, 刘亚曦. 多元线性回归模型对于空气质量的校准应用研究[J]. 科学技术创新, 2022, (2): 57-60.
- [2] 南国卫, 孙虎, 朱一梅. 陕西省PM_{2.5}时空分布规律及其影响因素[J]. 环境科学研究, 2018, 31(5): 834-843.
- [3] 孟丽红, 蔡子颖, 李英华, 等. 天津市PM_{2.5}浓度时空分布特征及重污染过程来源模拟分析[J]. 环境科学研究, 2020, 33(1): 9-17.
- [4] 项静恬, 郭世琪. 多元回归模型在实际应用中的几种推广[J]. 数理统计与管理, 1994, (04): 48-53.

作者简介:

张馨心(2003--), 女, 汉族, 湖南衡阳人, 在读本科, 研究方向: 自动化。