基于多元线性回归的雾霾预测方法研究

付倩娆
(西北工业大学理学院 西安 710072)

摘 要 提出了一种在线样本更新的多元线性回归分析的雾霾预测方法。首先搜集了北京市天气状况,包括平均气温、湿度、风级等气象数据以及 PM2.5、CO、NO_2、SO_2 等大气成分浓度数据,然后通过散点图对这些因素进行主要影响因素分析,筛选出对雾霾影响比较明显的因素作为雾霾预测的依据。通过在线样本更新的多元线性回归建立了 PM2.5 含量预测模型,并将气象要素作为雾霾的判断标准。最后给出实际例子,利用多元线性回归对北京未来一天、三天及一周的 PM2.5 含量进行较为精确的预测。

关键词 多元线性回归,主要影响因素分析,在线更新,PM2.5,雾霾预测
中图法分类号 O21 文献标识码 A

Research on Haze Prediction Based on Multivariate Linear Regression

FU Qian-rao
(School of Science, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract This paper presented a method to predict the haze based on multiple linear regression analysis, whose sample is online update. First, the data of Beijing weather conditions are collected, including average temperature, humidity, wind and other meteorological data level and PM2.5, CO, NO_2, SO_2 and other atmospheric concentration data. Then, the main influencing factors are analyzed by scatter plot of these factors, and the main factors, which have a great effect on the haze, are selected as the haze forecast basis. Furthermore, PM2.5 prediction model is built by multiple linear regression method. The prediction results combined with meteorological factors are in the form of the criterion of haze judgment. Finally, this paper provided an actual example for haze prediction, which utilizes multiple linear regression to forecast the weather condition of Beijing after one day, three days and seven days.

Keywords Multiple linear regression, Main influencing factors analysis, Online update, PM2.5, Haze forecast

1 引言

近年来,多次大范围持续雾霾天气先后侵袭我国中东部地区,给人们生产生活造成了严重影响,雾霾现象已经成为我国重要的环境公害[1]。准确地预测雾霾程度,提前做好防护措施,对降低雾霾对人们生产生活造成的危害有着极为重要的意义。

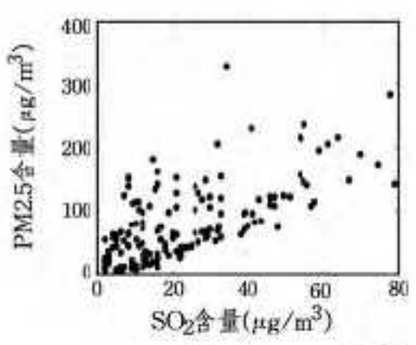
针对雾霾现象,学者提供了一些预报方法。文献[2]分析了雾霾形成过程中大气成分如 SO_2 和 NO_2 浓度与低层大气中雾霾特征的关系以及人气低空湍流特征对雾霾低能见度天气形成的机理,并给出了预测实例。文献[3]提出了一种基于统计结果的雾霾预测方法,方法根据当地风速和湿度进行预测,方法简单但通用性尚需检验。文献[4]提出了一种 BP 神经网络的雾霾预测方法,但其只考虑了气象因素,没有考虑大气成分,且 BP 神经网络易局部收敛。文献[5]通过建立 3 次指数平滑模型,分析 2002—2012 年我国 SO_2 和烟尘的排放量以及每年在环境污染治理方面的投资总额等指标,得出未来 3 年内我国雾霾天气仍会频发的结论,并分析了原因。多元线性回归预测具有模型简单、预测结果准确、模型解释能力强的特点,在模型预测中得到了广泛的应用[6-9]。文献[10]针对具体的雾霾天数情况,运用主成分分析法,建立对雾霾天数的线性回归模型,通过预测数据和实际数据的情况可以判断未来的年度雾霾天数。

本文提出了一种在线更新的多元线性回归 PM2.5 含量预测方法,并结合气象要素作为雾霾的判断标准。对未来一天、三天及一周的 PM2.5 含量进行较为精确的预测。每天根据当天的检测结果,不断更新模型,既保证了预测的精度,又无需大量的预测数据,还可以及时反映新情况的变化。

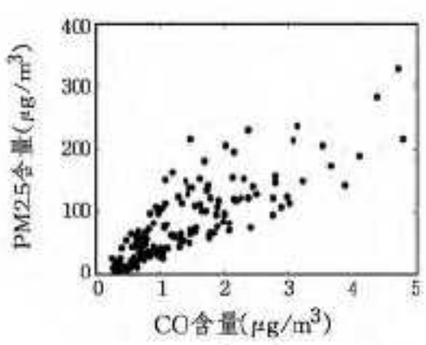
2 主要影响因素分析

主要影响因素分析可以通过画散点图来实现。图 1 给出了 PM2.5 含量与各影响因素的散点图。通过散点图可以看出湿度、风级、二氧化硫(SO_2)、一氧化碳(CO)、二氧化氮(NO_2)、臭氧(O_3)以及前一天的 PM2.5 含量与 PM2.5 含量的相关度较大;而温度的观测点很分散,表明对 PM2.5 值相关度不大,不作为影响因素。此外,SO_2、CO、NO_2、湿度与 PM2.5 含量正相关;O_3、风级与 PM2.5 含量负相关,这与常识相符合。SO_2、CO、NO_2 是污染物排放的含量,而湿度高时有利于污染物的积累;当风级较高时,这些污染物被扩散,从而降低了 PM2.5 含量。

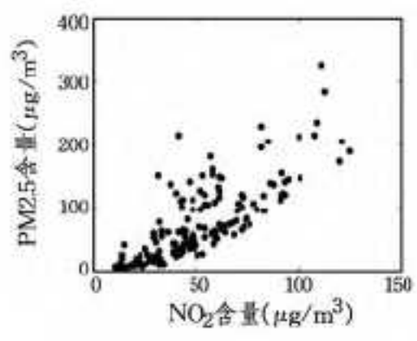
本文受国家自然科学基金(71171164,71401134),陕西省自然科学基础研究计划项目(2015JM1003)资助。
付倩娆(1990—),女,硕士生,主要研究方向为数理统计、可靠性分析,E-mail:1551742802@qq.com。



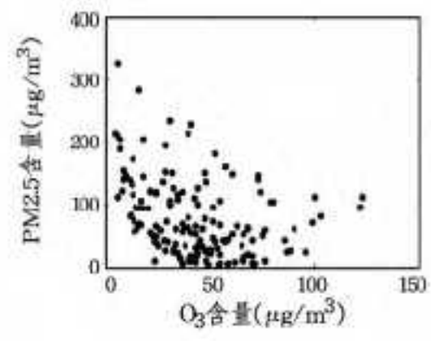
(a)PM2.5 含量与 SO_2 含量



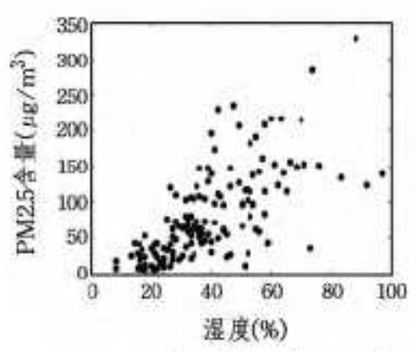
(b)PM2.5 含量与 CO 含量



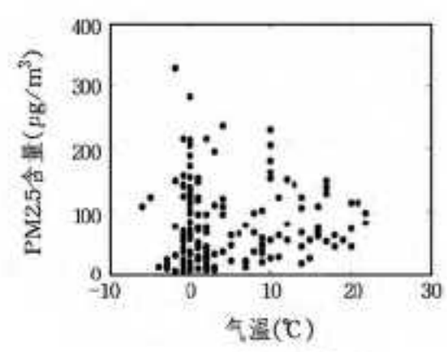
(c)PM2.5 含量与 NO_2 含量



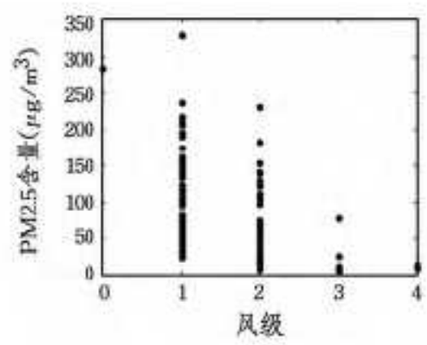
(d)PM2.5 含量与 O_3 含量



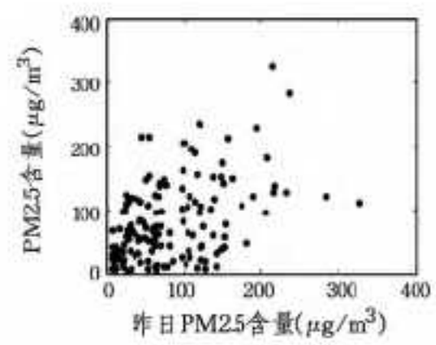
(e)PM2.5 含量与相对湿度含量



(f)PM2.5 含量与气温



(g)PM2.5 含量与风级



(h)PM2.5 含量与昨日 PM2.5 含量

图 1　PM2.5 含量和各影响因子的散点图

通过散点图分析，可以得出结论：湿度、风级、SO_2、CO、NO_2、O_3，以及前一天的 PM2.5 浓度可作为多元线性回归分析预测的输入变量。而温度的观测点很分散，表明对 PM2.5 值相关度不大，不作为影响因素。

3　多元线性回归

3.1　多元线性回归简介

设 $(y_i, x_{1i}, x_{2i}), i=1,2,\cdots,n$，是取自总体的一组随机样本。在该组样本下，回归模型可表示为：

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \tag{1}$$

记 $y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$，$X=\begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix}$，$\beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$，$\mu=\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$

则在该组样本下，回归模型的矩阵表示为：

$$y=X\beta+\mu \tag{2}$$

记 $\hat{\beta}=\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_n \end{bmatrix}$。则样本回归模型的矩阵表示为：

$$y=X\hat{\beta} \tag{3}$$

3.2　参数估计和预测

系数向量 β 的 OLS 估计为：

$$\hat{\beta}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \tag{4}$$

其中，X^{T} 为 X 的转置矩阵。

给定样本外一点 $x_f=(1,x_{1f},x_{2f},\cdots,x_{nf})^{\mathrm{T}}$，则因变量 y_f 的点预测为：

$$\hat{y}_f=x_f\hat{\beta} \tag{5}$$

3.3　实际预测结果

利用上节所述的回归分析方法，研制了雾霾预测软件——雾霾预测小助手。雾霾预测小助手是在 Windows 操作系统 Matlab 仿真平台上，通过图形用户界面（Graphical User Interface，GUI）搭建的测试平台。如图 2 所示，用户通过该软件可以直观地观测未来 7 天内预测的 PM2.5 浓度的变化以及雾霾预报情况。

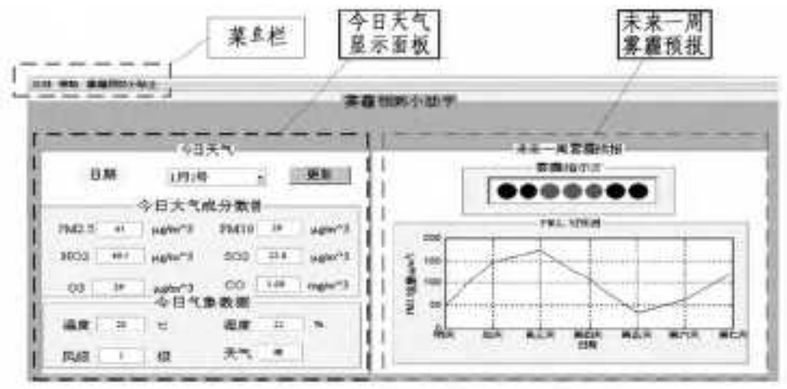


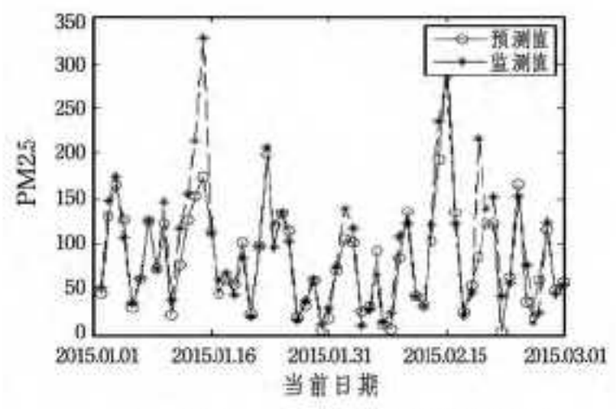
图 2　雾霾预测小助手操作界面

传统多元回归分析预测方法依赖大样本数据，采用固定的回归预测模型；而本文中每天预测都以该天前一月的数据作为样本，进行多元线性回归分析，不断更新模型，既保证了预测的精度，又无需大量的预测数据，还可以及时反映新情况的变化。

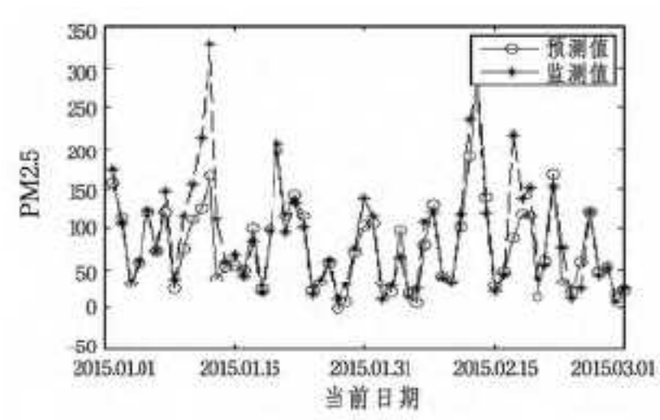
2010 年中国气象局发布的《气象行业标准霾的观测和预报等级》规定，“当大气成分站 PM2.5 日均浓度大于 0.075 mg/m³ 时，可作为判始霾的重要标准”，同时大量统计结果显示霾往往出现在风速小于 4m/s（风速三级以下）和相对湿度大于 40% 以及小于 90% 的情况[2]。本文雾霾判断标准为 PM2.5 日均浓度大于 0.075mg/m³，风速小于三级以及湿度大于 40% 以及小于 90%。

本文采用了北京市 2014 年 12 月—2015 年 4 月的大气成分数据和气象数据，对未来一周的 PM2.5 值及雾霾情况进行预测，其未来第一天到第七天的预测值与真实监测值的结果分别如图 3 所示：图 3 给出了 1 月 1 日至 3 月 31 日对未来一周的 PM2.5 含量预测结果。从图 3 可以看出，预测结果与真实监测值总体相符，表现出模型良好的预测能力，且稳定性较好，即长期预测仍然有较好的预测结果。

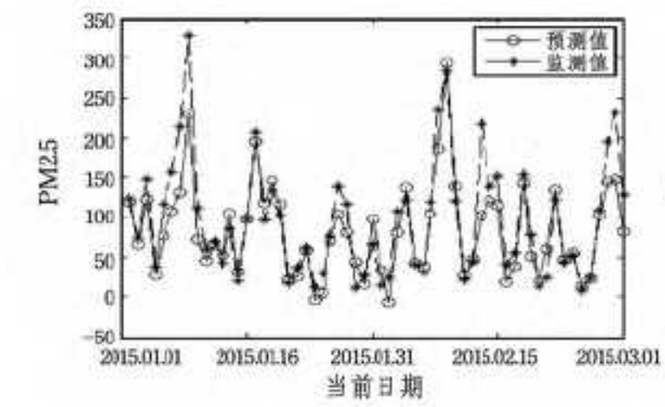
模型存在个别 PM2.5 的预测值与监测值存在突变的现象，如 2015 年 1 月 14 日，此日是重度污染，且当天的 SO_2、CO、NO_2 相对来说特别大，相对湿度也很大，而风级又偏小，与此同时还是一个霾转小雪的天气，从而使得预测结果出现了很大的偏差，而且当天在预测过程中都出现这个突变，分析认为是由于样本集有限且样本集未出现类似现象导致的，这是模型的固有缺陷。



(a)第一天的 PM2.5 预测与监测结果



(b)第三天的 PM2.5 预测与监测结果



(c)第七天的 PM2.5 预测与监测结果

图 3　未来 7 天 PM2.5 预测结果

表 1 给出了其未来七天雾霾预报结果的统计值。从表 1 可以看出,在预测过程中,随着预报周期的变长,空报率和漏报率都有所增加,命中率有所降低,但是命中率都维持在 80%以上,预测精度较高,预报结果比较满意,从而验证了所采用预测模型及预测方法的有效性。

表 1　其未来 7 天雾霾预报结果的统计值

指标 \ 天数	第一天	第三天	第七天
空报率	0.1000	0.1000	0.1167
漏报率	0.0500	0.0500	0.0667
命中率	0.8500	0.8500	0.8167

结束语　所提模型通过多元线性回归,采用了在线更新的预测方式,在保证预测精度同时,无需大量预测数据。与传统的大数据离线预测模式不同,每天根据当天检测结果,不断更新模型,既保证了预测的精度,又无需大量的预测数据,并可以及时反映新情况的变化。根据本文的原理设计了雾霾预测小助手之类的软件,为后续手机 APP 设计提供了参考。

参 考 文 献

[1] 杨准. 雾霾现象成因初步探讨[J]. 科技创新导报,2014(34):21

[2] 王继志,杨元琴,周春红,等. 雾霾低能见度天气分析与预测方法研究[C]//2007 年中国气象学会年会论文集. 2007:145-149

[3] 张琳,胡雪红. 德州市雾霾客观预报方法[J]. 农技服务,2010,27(11):1493-1493

[4] 艾洪福,石莹,等. 基于 BP 人工神经网络的雾霾天气预测研究[J]. 计算机仿真,2015,32(1):402-405

[5] 侯琼煌,杨航. 基于三次指数平滑模型的雾霾天气分析与预测[J]. 环境保护科学,2014(6):73-77

[6] 王勇,黄国兴,彭道刚. 带反馈的多元线性回归法在电力负荷预测中的应用[J]. 计算机应用与软件,2008,25(1):82-84

[7] 李军成,陈国华,石小芳. 基于灰色多元线性回归的粮食产量预测[J]. 安徽农业科学,2010,38(16):8281-8282

[8] 周晨,冯宇东,肖匡心,等. 基于多元线性回归模型的东北地区需水量分析[J]. 数学的实践与认识,2014,44(1):118-123

[9] 周永生,肖玉欢,黄润生. 基于多元线性回归的广西粮食产量预测[J]. 南方农业学报,2011,42(9):1165-1167

[10] 李莉,孙永霞. 基于均值化主成分分析的雾霾环境分析与研究[J]. 计算机应用研究,2015(5):1373-1375

(上接第 504 页)

例如先天性盲人疾病就是一种基因病,可以通过基因分析对其进行有效筛选。在未来,基因测序服务能够为个性化健康管理提供科学依据。

结束语　针对人类基因这样的大数据分析处理问题,文中介绍了一种新的云计算平台 Spark。Spark 较传统的 Hadoop 平台有了很大程度的改进,极大地缩短了数据处理的运行时间,同时集成了一些资源管理平台,使得对数据的分析处理更加方便、简单,基于 Spark 的人类基因系统的研究具有广阔的应用前景。

参 考 文 献

[1] 赵广荣,杨冬,白姝,等. 现代生命科学与生物技术[M]. 天津:天津大学出版社,2008

[2] 严霄凤,张德馨. 大数据研究[J]. 计算机技术与发展,2013,23(4):168-172

[3] 孙磊,胡学龙,张晓斌,等. 生物医学大数据处理的云计算解决方案[J]. 电子测量与仪器学报,2014,28(11):1190-1197

[4] 胡秀. 基于 Web 的数据挖掘技术的研究[J]. 软件导刊,2015,14(1):149-150

[5] 米允龙,米春桥,刘文奇. 海量数据挖掘过程相关技术研究进展[J]. 计算机科学与探索,2015,9(6):641-659

[6] Zhang Shu-fen, Yan Hong-can, Chen Xue-bin. Research on Key Technologies of Cloud Computing[C]//2012 International Conference on Medical Physics and Biomedical Engineering. 2012:1791-1797

[7] 卢小宾,王涛. Google 三大云计算技术对海量数据分析流程的技术改进优化研究[J]. 图书情报工作,2015,59(3):6-11

[8] 崔杰,李陶深,兰红星. 基于 Hadoop 的海量数据存储平台设计与开发[J]. 计算机研究与发展,2012,49(21):12-18

[9] 林清滢. 基于 Hadoop 的云计算模型[J]. 现代计算机,2010(7):114-116

[10] 王家林. 大数据 Spark 企业级实战[M]. 北京:电子工业出版社,2015

[11] Wu Leng-dong, Yuan Li-yan, You Jia-huai. Survey of Large-Scale Data Management Systems for Big Data Applications[J]. Journal of Computer Science and Technology, 2015, 30(1):163-183

[12] 陈虹君. 基于 Hadoop 平台的 Spark 框架研究[J]. 电脑知识与技术,2014,10(35):8407-8408

[13] 胡俊,胡贤德,程家兴. 基于 Spark 的大数据混合计算模型[J]. 计算机系统应用,2015,24(4):214-218

[14] 所剑,王大广,刘泽锋. 早期胃癌诊断和治疗[J]. 中国实用外科杂志,2011,31(8):717-719