# 集对分析在城市空气污染预报中的应用研究*

诸晓明 王国强
(浙江省绍兴市气象台, 绍兴 312000)

摘 要

在城市空气污染预报模型中,精心挑选的因子具有较好的预报性能,但是因子的优良性能并非始终不变,而有时个别因子的不良表现往往可能导致预报的失败。根据集对分析(SPA)把不确定性和确定性作为一个动态的同异反系统处理的思想,动态地分析和处理每次预报中因子作用的变化,即每次预报前,先对因子进行态势判别和同异反分析,然后使可能干扰预报的弱势因子的作用受到有效抑制,使有助于预报的强势因子的作用得到充分发挥,从而实现了因子作用大小在各次预报中的动态变化,取得了较为满意的效果。在预报模型中增加不确定性处理有助于提高预报准确率。

关键词: 集对分析; 不确定性; 联系度; 动态多元回归模型; 空气污染预报

## 引 言

城市空气质量与一定范围内污染源的分布和排放有关,与大气运动对空气中污染物的稀释、扩散、清除和聚集的强度有关。前者可用当地环境监测站的实测空气质量记录来反映,并认为污染源在短期内有相对稳定性,而空气污染预报主要从天气过程与污染物的关系出发进行研究[1-2]。大气运动具有随机性,天气预报具有不确定性,城市空气污染预报同样具有不确定性。本文应用不确定性系统理论和方法——集对分析[3](set pair analysis: SPA)来处理城市空气污染预报中的不确定性问题,以提高预报准确率。

## 1 集对分析的基本思路

天气系统具有确定性和不确定性的双重特性[4],相应的预报模型也应有既确定又不确定的品质,即同时具有处理确定性问题和不确定性问题的能力。在多元回归预报模型中,进入模型的因子虽然经过天气学和数理统计方法的精选,具有所谓的"优良性能",但是这种"优良性能"实际上是对于整个样本或次数众多的预报过程而言,是一个整体概念,对于样本中的某一个例或者具体应用中的某次预报则并非完全如此。有时其中的一个或几个因子"性能不佳"就导致了一次预报的失败。也就是说,在使用多元回归模型所进行的多次天气预报中,模型中各因子的预报性能是在不断变化的。一些因子在这次预报中表现出较强的预报性能,而在另一次预报中则表现出较弱的预报性能,甚至还有可能起干扰预报的负作用。这给我们提出了如下的问题:假设有一个多元回归模型,因变量是一维随机变量 $Y$,自变量为 $m$ 维变量 $X$。在某次预报中有 $p(p<m)$ 个自变量分量性能不佳,那么,能否使这 $p$ 个自变量分量在这次预报中少发挥作用或不发挥作用,而由其余的 $(m-p)$ 个性能优良的自变量分量来决定模型的预报结论?换句话说,要提高多元回归模型在多次天气预报中的预报准确率,必须使模型中各自变量分量的作用大小动态地变化,每当自变量分量的预报性能下降时,它的作用就要受到某种抑制,而让其它预报性能较好的自变量分量的作用得到充分发挥。

## 2 空气污染预报中自变量性能的优劣识别

### 2.1 邻近估计和变异系数

人们可能认为,从差别甚微的 $n$ 个初始场出发通过预报模式的积分,得到的 $n$ 个预报结果应该"差别甚微",但是大量的预报实践表明这 $n$ 个预报

* 2005-03-11 收到,2005-10-20 收到再改稿。

结果可能发散到较大区域。这是由空气污染预报的不确定性所造成的。用不同的预报模式分别对 $n$ 个初始场作预报,如果这种发散越小,就认为预报模式的质量越好[5]。

同样在用非参数回归制作天气预报时,近邻估计[6]要在样本中为估计点 $X$ 找到最为相近的 $k$ 个个例,记为$(X_i, Y_i)$(其中 $i=1,2,\cdots,k$)。此时 $k$ 个近邻的 $X_i$ 是 $k$ 个初始场,$k$ 个近邻的 $Y_i$ 是 $k$ 个预测值。如果 $k$ 个近邻的 $Y_i$ 分布越集中,则表示用自变量 $X$ 去预测 $Y$ 的效果越好;反之如果 $k$ 个 $Y_i$ 的分布越分散,则表示用 $X$ 去预测 $Y$ 的效果越差。$k$ 个近邻的 $Y_i$ 分布是集中或分散的程度称离散度,可以用标准差或方差来定量描述。考虑不同单位和不同数量级别的两组数字的离散度比较,则可采用标准差除以平均值所得到的变异系数[7](coefficient of variability)来表示。当预报量为一维时,如果用$\sqrt{s}$表示标准差,$\overline{Y}$表示均值,则变异系数 $c_v$ 表示为:

$$c_v = \frac{\sqrt{s}}{\overline{Y}} \tag{1}$$

有时为了计算方便把变异系数的平方当变异系数使用,此时变异系数 $c_v$ 可表示为:

$$c_v = \frac{s}{\overline{Y}^2} \tag{2}$$

**2.2　同异反分析**

为了表述简便起见,暂时把多元回归模型中的某一因子记为 $Z_i(i=1,2,\cdots,n)$,记 $Z_i$ 中的最大值为 $Z_{max}$,把第 $i$ 个例的 $Z_i$ 与 $Z_{max}$ 组成集对。对集对进行对比分析可知,当用 $Z_i$ 去预测 $R_i$ 时,可以给出的预测是一个概率分布,它的均值可作为实际使用中的预测值。在某问题下对某集对作分析,它们共有 $n=q+f+p$ 个特性,其中有 $q$ 个特性为两个集合所共有,有 $p$ 个特性为两个集合对立,在其余的 $f$ 个特性上则表现为既不对立又不同一。如果预测的不确定性较小,则同一度 $a=\frac{q}{n}=\frac{Z_i}{Z_{max}}$,对立度 $c=\frac{p}{n}=\frac{(Z_{max}-Z_i)}{Z_{max}}$,差异度 $b=\frac{f}{n}=0$。如果预测的不确定性较大,则同一度 $a=\frac{q}{n}=0$,对立度 $c=\frac{p}{n}=0$,差异度 $b=\frac{f}{n}=\frac{Z_{max}}{Z_{max}}=1$。具体方法详见参考文献[8]。

**2.3　自变量预报性能优劣概念及强势与弱势判别**

根据 SPA 理论,当用同一度 $a$、差异度 $b$ 和对立度 $c$ 去剖析多元回归模型中自变量 $X$ 与因变量 $Y$ 的关系时,事实上还描述了自变量的一种态势,因此为方便计,我们把某次天气预报中具有较好预报性能的自变量分量称为是处于强势的自变量分量,把在某次预报中具有较差预报性能的自变量分量称为是处于弱势的自变量分量。一个自变量分量是处于强势还是弱势,由相应的变异系数决定。变异系数取最大值时,认为该因子处于弱势。

设有 $n$ 次观测 $X_i, Y_i(i=1,2,\cdots,n)$,在自变量定义域中有任意估计点 $X=x, x$ 和 $X_i$ 在 $p$ 维自变量空间的位置写成列向量

$$\boldsymbol{x} = (x_{01}, x_{02}, \cdots\cdots, x_{0p})' \tag{3}$$

和

$$\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots\cdots, X_{ip})' \tag{4}$$

则它们之间的统计距离定义为

$$\rho_i(x) = \left| \frac{(x_{01}-X_{i1})^2}{s_{11}} + \frac{(x_{02}-X_{i2})^2}{s_{22}} + \cdots\cdots + \frac{(x_{0p}-X_{ip})^2}{s_{pp}} \right|^{1/2} \tag{5}$$

其中 $s_{ii}(i=1,2,\cdots\cdots,p)$ 为自变量均方差。如果自变量为一维,即当 $p=1$ 时,那么

$$\rho_i(x) = \frac{|x_{01}-X_{i1}|}{\sqrt{s_{11}}} \tag{6}$$

从式(5)和式(6)可知,统计距离以样本标准差为基本单位。把式(6)与式(1)归纳在一起:

$$\begin{cases} \rho_i(x) = \dfrac{|x_{01}-X_{i1}|}{\sqrt{s_{11}}} \\ c_v = \dfrac{\sqrt{s}}{\overline{Y}} \end{cases} \tag{7}$$

利用式(7)可以对回归模型中每一因子的各个例分别做近邻分析,找出 $k$ 个近邻,对 $k$ 个近邻因变量计算变异系数。在每一个例中,由 $m$ 个因子计算出 $m$ 个变异系数,变异系数取最大值时,可认为该因子可能处于弱势。

以绍兴市气象台城市空气污染预报中可吸入颗粒物($PM_{10}$)的污染指数预报为例来说明态势的分析与判断过程。在该预报过程中并不直接预报 $PM_{10}$ 的污染指数,而先预报它的变量,即先预报 $\Delta Y = Y_t - Y_{t-1}$,再计算出 $Y_t$。其中 $t$ 为要预报的日期,$t-1$ 为 $t$ 日的前一天日期。资料选用 2003 年和 2004 年 5 月 21 日—9 月 10 日,模型中涉及样本的容量为 222(缺 4 天资料),预报因子为 6 个,其中 $X_1$ 为 MM5 模式输出的 12 ~ 24 h 雨量;$X_2$ 为 24 ~

36 h 雨量; $X_3$ 为 T213 的 850 hPa 24 h 温度梯度; $X_4$ 为 700 hPa 24 h 垂直速度; $X_5$ 为 850 hPa 24 h 温度露点差; $X_6$ 为24 h海平面气压梯度。整个分析与判断过程分两步进行:第一步,对由数值预报产品格点资料进行天气学和统计学方面的分析加工,组合成为关键区或锋区因子;第二步,根据各个例中每个因子的 $c_v$ 值来判断因子的强势或弱势状态。表 1 列出了各因子各个例的变异系数 $c_v$,计算时取近邻数为 $k = 7$[9]。对表中变异系数的进一步统计分析可知,当某因子的 $c_v$ 值为 6 个因子中最大的 $c_v$ 值,且($c_v \geqslant 51.8$)时,该因子的预报能力较差,表明其处于弱势。如在个例 1 中因子 $X_6$ 的变异系数 $c_v = 289.0$ 为 6 个因子中的最大值,表明该因子 $X_6$ 处于弱势;在个例 2 中没有因子处于弱势;而在个例 3 和个例 4 中,可以看到都是因子 $X_1$ 处于弱势,其余个例可类推。

表 1 各预报因子的变异系数

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| 个例 1 | 63.2 | 155.8 | 30.5 | 11.6 | 16.5 | 289.0 |
| 个例 2 | 38.0 | 22.6 | 15.6 | 0.8 | 5.7 | 37.0 |
| 个例 3 | 71.1 | 15.1 | 56.3 | 6.3 | 11.0 | 2.2 |
| 个例 4 | 54.4 | 22.2 | 22.0 | 37.8 | 6.4 | 18.1 |
| 个例 5 | 53.8 | 30.4 | 13.1 | 37.8 | 14.9 | 599.0 |
| … | … | … | … | … | … | … |

## 3 自变量分量处于弱势时的分解

在预报模型的自变量中,自变量分量之间相互联系、相互制约,有机地组成一个整体。在预报时如果发现一个分量处于弱势,说明它在模型中的重要性已下降,甚至可能干扰模型作出正确预报结论,因而希望让这些分量在这次预报中减少作用甚至失去作用。要达到此目的,显然不能简单地剔除这个自变量分量。本章从 SPA 的原理出发,用联系度公式导出解决这一问题的具体方法。

在联系度表达式中,差异度 $b = \frac{f}{n}$,它表示在 $n$ 个特征中有 $f$ 个特征表现为既不同一又不对立,即有 $f$ 个特征对预报量的预测持“含糊”态度,该因子与其勉强参与表态,还不如放弃“投票权”,而把预报结论的决定权让给其他强势因子。为此这里令

$$i = \frac{\sum q}{\sum q + \sum p} + \frac{\sum p}{\sum q + \sum p} j \qquad (8)$$

或表示为

$$i = \frac{\sum a}{\sum a + \sum c} + \frac{\sum c}{\sum a + \sum c} j \qquad (9)$$

如果回归模型共有 $m$ 个因子,则式中 Σ表示对 $m$ 个因子求和,在上式中求和与求平均等价。

式(8)和式(9)的含义是:当因子处于弱势时,它的差异度的 $f$ 个特征按一定比分配给它的同一度和对立度,这个比就是所有因子的平均同一度与平均对立度之比。按 SPA 的规定,取 $j = -1$,可得

$$\mu = \frac{q - p}{n} + \frac{f \sum q - \sum p}{n \sum q + \sum p} \qquad (10)$$

或

$$\mu = (a - c) + \frac{b(\sum a - \sum c)}{\sum a + \sum c} \qquad (11)$$

式(10)和式(11)是适用于多元回归预报模型的联系度表达式。用此式对表 2 的数据进行计算,可得到各因子的联系度值,表 3 仅列出 $X_1$ 和 $X_2$ 的联系度,其余可类推。

表 2 各个例中预报因子 $X_1$ 的同异反分析

| | 同一度 $a$ | 差异度 $b$ | 对立度 $c$ |
|---|---|---|---|
| 个例 1 | 0.25 | 0.000 | 0.75 |
| 个例 2 | 0.16 | 0.000 | 0.84 |
| 个例 3 | 0.00 | 1.000 | 0.00 |
| 个例 4 | 0.00 | 1.000 | 0.00 |
| 个例 5 | 0.14 | 0.000 | 0.86 |
| … | … | … | … |

表 3 预报因子的联系度

| | $\mu_1$ | $\mu_2$ | |
|---|---|---|---|
| 个例 1 | -0.501 | -0.722 | …… |
| 个例 2 | -0.686 | -0.859 | …… |
| 个例 3 | -0.583 | -0.571 | …… |
| 个例 4 | -0.689 | -0.343 | …… |
| 个例 5 | -0.724 | -0.715 | …… |
| … | … | … | … |

## 4 动态多元回归模型应用效果比较

通过上面一系列的分析处理,现在可以用多元线性回归模型和最小二乘法,对表 3 的联系度资料作为自变量,建立新的预报模型。本文称新的回归模型为基于 SPA 的动态多元回归模型,简称为动态回归模型,相应的工作称为动态多元回归分析。

前面的例子中有 6 个预报因子,分别为 $X_1$ ~ $X_6$。通过一系列的处理,可得到它们相应的映射,映射主要应用式(10)以及因子的态势判断式(7)来

进行的。式(10)是针对多元回归模型的特点从集对分析的联系度表达式推导而来。如果注意一下前面的推导过程,不难发现当因子处于强势时,因子的映射只不过是线性变换;而当因子处于劣势时,该因子在模型中的作用已消失,它的作用已由其它因子取代,自然这是非线性变换。如果所有因子都是线性变换,映射并不能使模型的质量有所变化。映射前后因子与预报量的相关系数见表 4,可见大部分因子的相关系数有了提高。复相关系数和残差平方和,新模型为 0.609 和 50260.1,传统多元回归模型为 0.514 和 58765.1,新模型的复相关系数有了提高,预报误差则比原模型减少了 14.47%,可见新模型预报能力有了明显提高。

表 4　预报因子映射前后的相关系数比较

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| 传统多元回归模型 | 0.445 | 0.417 | 0.425 | 0.426 | 0.464 | 0.442 |
| 动态多元回归模型 | 0.443 | 0.457 | 0.460 | 0.471 | 0.505 | 0.480 |
| 效果评定 | 不提高 | 提高 | 提高 | 提高 | 提高 | 提高 |

SPA 有两个基本观点:不确定性和确定性可以放在同一个系统中进行分析和处理;不确定性与确定性在一定条件下可以相互转换。在把 SPA 应用于多元预报模型的过程中,本文主要做了两件工作:一是设计了自变量分量态势的判断方法,用以辨认确定性部分和不确定性部分;二是推导了适用多元回归分析的联系度表达式,用以使不确定性转化为确定性。

表 5 是以第 66 号个例所作的分析。根据变异系数判断,第 1 因子处于弱势,$X_1$ 对预报量的估计为 12.15,预报量实况为 43,$X_1$ 的估计比实况明显偏小。由于 $X_1$ 被判别为弱势,用式(11)作非线性映射得到新因子,新因子是 $X_2$, $X_3$, $X_4$, $X_5$ 和 $X_6$ 的函数,用新因子计算得到它对预报量的估计为31.07,新因子的预报值比原因子预报值有了明显提高。其他因子由于并不处于弱势,它的一元回归值没有变化。原预报模型的预报值为 25.0,动态多元回归模型的预报值为 32.6,它们的预报误差分别为 18.0 和 10.4,后者比前者减少了误差42.22%。可见通过因子的势态判断、因子的同异反分析和线性非线性变换等一系列处理,预报误差有了明显减少,同时也可看到动态多元回归模型对因子处理机制是一般预报方法难以实现的。

表 5　SPA 对因子的订正

| 因子类型 | 因子分析 | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | 模型预测值 |
|---|---|---|---|---|---|---|---|---|
| 原因子(映射前) | 变异系数 | 88.7 | 55.8 | 16.5 | 19.8 | 40.3 | 21.7 | 25.0 |
| | 一元回归值 | 12.15 | 21.20 | 44.21 | 34.51 | 23.04 | 30.35 | |
| | 势态评定 | 弱势 | 强势 | 强势 | 强势 | 强势 | 强势 | |
| 新因子(映射后) | 因子值 | -0.658 | -0.803 | -0.517 | -0.584 | -0.714 | -0.671 | 32.6 |
| | 一元回归值 | 31.07 | 21.20 | 44.21 | 34.51 | 23.04 | 30.35 | |

## 5　结　语

多元回归模型是在城市空气污染预报中应用广泛的一种预测模型,回归分析对因子的筛选有许多行之有效的方法[10],而如何合理地使用因子则未引起人们的足够重视。实际上,合理地选择回归因子和合理地使用因子同样重要。从这个意义上说,本文给出了一种科学地使用因子的新思路。而在各种各样的预测问题中,城市空气污染预报是一种典型的复杂的预测问题,它既含有确定性,又含有不确定性。根据 SPA 把不确定性和确定性作为一个动态的同异反确定不确定系统处理的思想,动态地分析和处理每次预报中因子作用的变化,效果较为满意,说明为预报模型增加不确定性处理的能力有助于提高城市空气污染预报准确率[11]。

## 参 考 文 献

[1] 孙明华,徐大海,朱蓉,等. 城市空气臭氧污染业务预报方案研究. 气象,2002,28(4):3-8.

[2] 徐大海,朱蓉. 大气平流扩散的箱格预报模型与污染潜势指数预报. 应用气象学报,2000,11(1):1-12.

[3] 赵克勤. 集对分析及其初步应用. 杭州:浙江科技出版社,2000.

[4] 史国宁. 概率天气预报的兴起及其社会意义. 气象,1996,22(5):3-8.
[5] 李小泉. 美国国家气象中心中期预报时段内的集合预报. 气象科技,1994,22(2):7-11.
[6] 陈希孺,柴根象. 非参数统计教程. 上海:华东师范大学出版社,1993.
[7] 中国科学院数学研究所. 常用数理统计方法. 北京:科学出版社,1973.
[8] 黄德才,赵克勤. 用联系数描述和处理网络计划中的不确定性. 系统工程学报,1999,(3):112-117.
[9] 王国强,蒋延龙,陈红梅. 近邻估计——线性回归预报模型及其台风暴雨预报试验. 气象科技,1999,27(4):25-29.
[10] 陈希孺. 近代回归分析. 合肥:安徽教育出版社,1987.
[11] 赵克勤. 集对分析对不确定性的描述和处理. 信息与控制,1995,24(3):165-168.

# The Application of Set Pair Analysis on City Air Pollution Index Forecasting

Zhu Xiaoming Wang Guoqiang
(*Shaoxing Meteorological Observatory, Zhejiang Province, Shaoxing* 312000)

**Abstract**

The predictors of city air pollution forecast models which are specially selected have prefer forecast ability in general. But sometimes when the situation changes, some predictors are negative and may result in failure of the forecast model. According to the principle of Set Pair Analysis (SPA), regarding uncertainty and certainty as a dynamic systematic procession, the evolution of predictors' action in every forecasting is analyzed and processed dynamically. That is to say, the judgment of potential states and Identical-Discrepancy-Contrary Analysis are made about predictors before being used to calculate weather forecast, then the effect of predictors on weak potential state which may interfere forecasting is suppressed effectively, while those predictors on strong potential state which may be contributed to forecasting is given full play. As a result the dynamic evolution in the structure of predictors is made in the forecast model, and the rationality of forecasting mechanisms and the ability of models are intensified. So adding the processing uncertainty to the forecasting model is help to improve the forecasting accuracy.

**Key words:** set pair analysis; uncertainty; connection degree; dynamic multiple regression model; air pollution forecasting