文章编号:1000-582X(2010)09-092-06

偏最小二乘回归在地表沉陷预测中的应用

蒋建平[1],陈功奇[1],章杨松[2]
(1.上海海事大学 海洋环境与工程学院,上海 201306;
2.南京理工大学 理学院土木工程系,江苏 南京 210094)

摘　要:考虑地下开采引起的地表沉陷的众多影响因素,基于偏最小二乘二次多项式回归这一非线性方法,对地表沉陷的最大值进行了预测。以地表最大沉陷值为因变量,以采高、采深、煤层倾角、硬度系数等为自变量,得出了地表最大沉陷值的预测模型。结果发现,Press 残差值随潜变量个数的增加而降低,由两者关系图可确定潜变量的个数为 4 对;采高的标准回归系数最大,说明 4 个影响因素中采高对地表沉陷值的影响最大;预测模型的决定系数为 0.915 7,预测值的误差率为 ±10.41%,表明用偏最小二乘二元多项式回归方法预测地表沉陷是可行的。

关键词:地表沉陷;偏最小二乘回归;预测;非线性
中图分类号:TU433　　**文献标志码**:A

Application of partial least-squares regression in the forecast of ground subsidence

JIANG Jian-ping[1], *CHEN Gong-qi*[1], *ZHANG Yang-song*[2]
(1. College of Ocean Environment and Engineering, Shanghai Maritime University, Shanghai 201306, P. R. China; 2. Department of Civil Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, P. R. China)

Abstract: Taking into account many influence factors of ground subsidence induced by underground exploitation, based on partial least-squares multinomial regression, a forecast analysis on the maximum of ground subsidence is carried out. Taking height, depth, obliquity of coal clay and rigidity coefficient as independent variables, and maximum of ground subsidence as dependent variable, the forecast model of maximum of ground subsidence is obtained. It is found that, Press residual value decreases with the increase of number of latent variables, and the number of latent variables is four by Press residual value versus number of latent variables. The normal regression coefficient of height is the largest in the four influence factors, and this indicates that the influence of height is the largest on maximum of ground subsidence. The determination coefficient of forecast model obtained in this paper is 0.915 7, the error of forecast model is ±10.41%. The following conclusion can be drawn that the model based on partial least-squares multinomial regression is a better and feasible non-linear method.

Key words: ground subsidence; partial least-squares regression; forecast; non-linear

收稿日期:2010-04-17
基金项目:国家自然科学基金项目(40872172);上海市教委科研创新项目(09YZ250);上海海事大学科研基金项目(2009160);港口、海岸及近海工程校重点学科项目(A2010030);上海市第四期本科教育高地建设项目(B210008G)
作者简介:蒋建平(1966-),男,上海海事大学副教授,主要从事岩土工程、地下工程、港航工程的研究,(E-mail)jjpwx@163.com。

地表在地面荷载和地下掏空作用下都会产生沉降[1-4]。如建筑地基、公路地基、铁路地基在荷载作用下会产生沉降，各种地下工程特别是采矿工程也会引起地表的沉降[5-7]。

地下开采将引起上覆岩层直至地表产生沉陷和变形。开采沉陷是岩石力学领域的一个分支，开采后引起岩层与地表移动一直是矿山生产经常遇到的问题。由于开采沉陷造成地面各种设施的损害，给矿山生产和安全带来了重大的经济损失，故开采引起的地面沉陷是采矿工程中的一个重要研究课题[8-10]。为保证地表建筑物的安全，不仅需要了解地表动态沉陷规律，还必须对沉陷动态观测资料进行分析，并作趋势预报[11]。

开采沉陷预测方法一般可归为 2 大类：一类是以地表移动作为研究对象，不考虑岩体特性的唯像学理论模型；另一类是以力学原理为基础的正演法和反分析法。第 1 类方法，选用的参数，物理意义不明确，很难反映岩层内部的移动规律，不适用于复杂地质条件和复杂采空区的地表沉陷预计。第 2 类方法，能对岩层移动过程做出解释，计算中所需参数有各自的物理意义，概念比较清楚。

近几十年来，开采地面沉陷预测研究已取得巨大进展[12]，先后提出了预测开采地面沉陷的经验方法、剖面函数法、数值模拟法、薄板弯曲理论、物理模拟法、概率积分法、随机介质理论法、三维层状介质理论法、人工神经网络法等。

但由于矿山开采引起的地表沉陷受各种地质和采矿等因素的综合影响，岩土体结构及其力学行为、实际开采条件非常复杂，地表沉陷表现出复杂性和非线性性。目前还没有成熟的计算岩体力学性态的模型和方法。因此，开采沉陷预测一直是岩土体力学和采矿工程研究中的难点和热点。

各种预测方法各有其优势、缺陷和适用范围。这里基于偏最小二乘二次多项式回归法对地表沉陷值进行预测。

1　偏最小二乘二次多项式回归方法

目前国内外广泛采用最小二乘法[13-14]（least-squares regression，LSR）进行回归参数的无偏估计，以实现对各种数据的模型拟合与预测。然而，最小二乘法回归是建立在自变量因子之间不存在密切相关关系的假定基础上，而实际工程情况往往与该假定不符。各种自变量因子之间总是存在着一定的相关关系，即多重相关性，它会导致回归分析的正则方程组出现病态，从而使最小二乘法的参数估计不稳定，模型拟合精度难以保证，在此基础上进行预测，将可能产生严重的偏差甚至错误。

长期以来，模型式方法和认识性方法之间的界限分得十分清楚。而偏最小二乘法[15-16]（partial least-squares regression，PLSR）则把它们有机的结合起来了，在一个算法下，可以同时实现回归建模（多元线性回归）、数据结构简化（主成分分析）以及两组变量之间的相关性分析（典型相关分析），这是多元统计数据分析中的一个飞跃。偏最小二乘法是由 S. Wold 和 C. Albano 于 1983 年提出。该方法是一种新型的多元统计分析方法，它能有效解决常规最小二乘法回归难以克服的自变量因子间多重相关性影响的问题，即与常规多元线性回归分析相比，PLSR 并不直接考虑因变量集合与自变量集合的回归建模，而是在变量系统中提取若干对系统具有最佳解释能力的新综合变量（即潜变量或成分），也就是在解释变量空间和反应变量空间中分别寻找某些线性组合（潜变量），并使得两个变量空间的协方差最大。然后利用这些潜变量进行回归建模，从而避免了模型因子之间的多重相关性干扰。具体的思路如下。

对于 q 个因变量 $\{y_1, y_2 \cdots, y_q\}$ 和 p 个自变量 $\{x_1, x_2 \cdots, x_p\}$，利用对其观测得到的 n 个样本点，可分别构成自变量和因变量矩阵系统 $\boldsymbol{X}=[x_1, x_2 \cdots, x_p]_{n\times p}$、$\boldsymbol{Y}=[y_1, y_2 \cdots, y_q]_{n\times q}$，利用 PLSR 建模，可分别在 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 中提取出潜变量 t_1 和 u_1（所提取的潜变量 t_1 是 $x_1, x_2 \cdots, x_p$ 的线性组合，u_1 是 $y_1, y_2 \cdots, y_q$ 的线性组合）。潜变量提取时，应满足两个要求：

1）t_1 和 u_1 应尽可能大地携带它们各自原变量系统中的数据信息；

2）t_1 和 u_1 的相关程度能够达到最大。

这两个要求表明，t_1 和 u_1 应尽可能好地代表 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，同时自变量的潜变量 t_1 对因变量的潜变量 u_1 有最强的解释能力。在第一对潜变量 t_1 和 u_1 被提取后，PLSR 分别实施 $\boldsymbol{X}$ 对 t_1 的回归以及 $\boldsymbol{Y}$ 对 t_1 的回归，如果回归方程已达到满意的精度，则算法终止；否则，将利用 $\boldsymbol{X}$ 被 t_1 解释后的残余信息以及 $\boldsymbol{Y}$ 被 t_1 解释后的残余信息进行第二轮的成分提取；如此往复，直到达到一个较满意的精度为止。若最终对 $\boldsymbol{X}$ 共提取了 m 个潜变量 $t_1, t_2, \cdots, t_m$，PLSR 将通过实施 y_k 对 $t_1, t_2, \cdots, t_m$ 的回归，然后再表达成 Y_k 关于原变量 $x_1, x_2 \cdots, x_p$ 的回归方程，即

$$y_k^* = a_{k1}x_1^* + a_{k2}x_2^* + \cdots + a_{kp}x_p^* + Y_{Ak}。\quad (1)$$

式(1)即为所建的 PLSR 拟合模型，其中 Y_{Ak} 是残差矩阵 $\boldsymbol{Y}_k$ 的第 k 列。许多情况下，PLSR 并不需要提取全

部的潜变量 $t_1,t_2,\cdots,t_A$(A 为自变量矩阵 $\boldsymbol{X}$ 的秩)进行建模,而是采取截尾的方式,只通过提取前 h 个潜变量,$h<A$,即能得到一个精度足够高、拟合与预测性能均较好的 PLSR 模型。对于建模所需提取的潜变量个数 h,可以通过交叉有效性检验来确定。

总之,偏最小二乘回归方法综合了多元线性回归法(MLR)和主成分回归法(PCR)的优势,同时从自变量矩阵和因变量矩阵中提取偏最小二乘成分,可以有效地降维,开辟了有效的回归分析途径,利用潜变量提取的思路,采用了信息综合和筛选技术,有效地克服了在应用最小二乘回归时遇见的自变量间的多重相关性,明显地改善了数据结果的可靠性和准确度,因而得到了日益广泛的应用。

然而应用偏最小二乘回归分析建立的模型是一种多元线性回归模型,在实际应用中,由于因变量和自变量之间的关系往往会呈现非线性关系,用线性模型 PLSR 处理此类数据时,势必会造成较大的偏差。对于这类问题的处理,通常采用 2 种方法:一种是对数据进行预处理,尽可能地去除噪声部分,如多倍分散校正(MSC)、正交信号处理(OSC)等。但是,这样的话存在着有用信息同时被去除和模型"过拟合"等问题。另一种就是建立非线性模型,如人工神经网络(ANN),但是 ANN 需要较多的训练集,另外还存在着容易"过拟合"的现象。而偏最小二乘二次多项式回归是一种非线性的多元回归,能有效克服上述缺陷,其回归拟合的精度比一般偏最小二乘线性回归方法要高。

2 基于偏最小二乘二次多项式回归的地表沉陷预测

2.1 理论思想

地表沉陷是一个复杂的系统,其影响因素众多。主要的影响方面有:1)地层情况,主要指岩土层的空间分布、各岩土层的物理力学参数;2)开采深度;3)开采工艺,包括采高、开采规模、开采方法等。经综合分析,从这 3 个影响方面析出的主要影响有:1)地层情况中有煤层倾角,一般煤层倾角越大,地表沉陷值越大;2)地层情况中还有围岩的硬度系数,一般硬度系数越大,地表沉陷值越小;3)开采深度,一般开采深度越大,地表沉陷值越小;4)开采工艺中的采高,一般采高越大,地表沉陷值越大。

地表沉陷值是上述各影响因素综合作用的结果,且各影响因素的值是已知的。如何由这些已知的影响因素值来综合求得地表沉陷值,这就是笔者要研究的。这里采用上述的偏最小二乘二次多项式回归方法进行分析。

2.2 预测

实例分析的数据来源于文献[17],如表 1 所示。设置自变量 x_1、x_2、x_3、x_4 分别代表采高、采深、煤层倾角和硬度系数,设置因变量 y 代表地表最大沉陷值。计算过程中发现,Press 残差值随潜变量个数的增加而降低(图 1)。因此,潜变量的个数确定为 4 对。计算结束后,得出的模型的 T_i、U_i 如表 2 所示;模型误差平方和、Press 残差和决定系数如表 3 所示;标准回归系数与模型效应负荷量如表 4 所示。

表 1 沉陷及其影响因素数据表

采高 x_1/m	采深 x_2/m	煤层倾角 x_3/(°)	硬度系数 x_4	最大下沉值(实测)y/m	最大下沉值(预测)/m	误差/%
2.20	126	25	5.5	1.395 7	1.476 8	5.807 747
3.00	225	4	5.5	2.094 8	2.098 5	0.175 673
2.50	120	41	5.5	1.320 7	1.342 7	1.664 434
3.78	113	38	5.5	2.085 0	1.909 4	−8.420 790
0.96	157	41	5.5	0.507 2	0.603 6	19.011 610
0.90	160	12	3.0	0.704 3	0.892 9	26.771 460
1.44	168	11	3.0	1.130 8	1.037 3	−8.270 360
2.15	197	22	3.0	1.594 8	1.208 6	−24.217 700
1.26	173	38	3.0	0.794 3	0.883 8	11.264 200
1.20	42	14	5.5	0.815 0	0.771 8	−5.302 110
2.00	400	0	5.5	0.740 0	0.814 3	10.039 980
2.80	155	22	3.0	0.934 6	1.160 9	24.215 360
1.65	147	3	3.0	1.153 4	1.024 1	−11.206 500
2.20	125	10	5.5	1.516 6	1.604 3	5.785 110

续表

采高 x_1/m	采深 x_2/m	煤层倾角 x_3/(°)	硬度系数 x_4	最大下沉值(实测)y/m	最大下沉值(预测)/m	误差/%
2.00	333	8	5.0	1.386 4	1.323 7	−4.519 80
1.67	272	12	5.5	1.143 4	1.225 1	7.148 21
2.80	310	12	5.5	1.917 2	1.856 4	−3.171 48

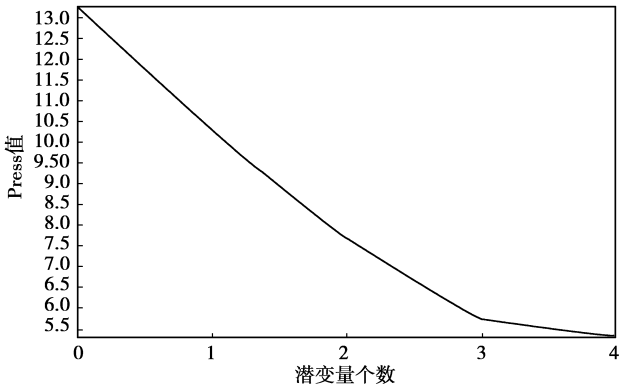

图 1　Press 残差值与潜变量个数关系图

表 2　模型的 T_i、U_i

t_1	t_2	t_3	t_4	u_1	u_2	u_3	u_4
0.561 3	−0.368 6	0.516 3	−0.579 6	0.303 6	0.003 4	−0.140 8	−0.255 8
1.252 4	−2.031 6	1.422 2	1.213 4	1.751 2	−1.066 3	0.309 1	−0.007 7
0.922 0	1.857 5	1.352 2	0.847 1	0.148 3	0.355 9	0.336 4	0.035 1
4.400 4	1.749 4	−0.423 7	−1.458 9	1.730 9	0.675 7	−0.023 8	0.070 6
−1.839 6	2.454 3	1.873 2	0.210 8	−1.536 1	0.530 0	0.384 7	−0.032 6
−0.835 7	−0.062 0	−0.976 6	−1.365 7	−1.128 0	0.671 0	−0.694 1	−0.476 5
−0.857 9	−0.425 1	−0.467 4	−0.585 0	−0.244 9	−0.224 3	0.065 9	0.170 0
−0.690 5	−0.839 1	0.024 6	−0.391 1	0.715 9	−1.093 5	0.780 8	0.775 3
−1.636 9	0.421 5	0.930 8	0.017 1	−0.941 7	0.046 4	0.110 7	−0.096 7
−1.157 3	0.166 3	−0.646 2	−2.295 4	−0.898 8	0.265 9	−0.203 9	−0.059 9
−0.363 4	1.430 3	−2.244 7	3.078 5	−1.054 1	0.855 4	−0.322 3	0.177 8
−0.259 7	−0.514 2	−0.607 7	−1.059 8	−0.651 2	0.509 1	−0.700 8	−0.565 4
−0.981 7	−0.176 7	−0.173 8	0.029 7	−0.198 1	−0.338 8	0.272 9	0.311 7
0.553 5	−1.269 0	0.421 5	−0.840 4	0.554 0	−0.251 2	−0.221 7	−0.315 6
0.208 5	−0.214 5	−0.699 6	1.455 7	0.284 4	−0.170 3	0.090 4	0.246 3
0.004 8	−0.377 9	−0.774 6	−0.118 1	−0.218 8	0.221 4	−0.362 3	−0.189 7
0.719 8	−1.800 6	0.473 2	1.841 6	1.383 5	−0.989 8	0.318 7	0.213 3

表 3 模型误差平方和、Press 残差和决定系数

潜变量个数	数据标准化后模型误差平方和	数据标准化后模型 Press 残差	决定系数 R^2
1	6.286 7	13.252 5	0.607 1
2	2.338 0	10.215 7	0.853 9
3	1.886 0	7.670 2	0.882 1
4	1.349 5	5.717 7	0.915 7

表 4 标准回归系数与模型效应负荷量

组分	x_1	x_2	x_3	x_4	x_1^2	x_2^2	x_3^2
标准回归系数	0.709 9	0.185 1	−0.110 8	0.162 4	−0.073 1	−0.297 2	−0.090 8
组分	x_4^2	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4
标准回归系数	−0.192 1	0.050 4	−0.060 7	0.311 5	0.139 1	−0.103 8	−0.031 1
组分	x_1	x_2	x_3	x_4	x_1^2	x_2^2	x_3^2
模型效应负荷量	0.122 0	0.627 2	−0.209 5	0.216 2	−0.270 4	0.386 2	0.243 5
组分	x_4^2	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4
模型效应负荷量	−0.259 5	0.047 1	−0.260 2	0.002 4	−0.534 8	0.520 7	−0.199 6

标准回归系数可无量纲地比较各个自变量 x_i（采高、采深、煤层倾角、硬度系数）对因变量 y（地表最大沉陷值）的影响。从表 4 可发现，采高对地表沉陷值的影响最大。

计算得出的预测方程式为：

$$y = -1.388\,340\,1 - 0.220\,826x_1 + 0.006\,534x_2 + 0.002\,935x_3 + 0.816\,367x_4 - 0.046\,327x_1^2 - 0.000\,012x_2^2 - 0.000\,243x_3^2 - 0.105\,798x_4^2 + 0.000\,444x_1x_2 - 0.002\,373x_1x_3 + 0.175\,714x_1x_4 + 0.000\,061x_2x_3 - 0.000\,63x_2x_4 - 0.000\,99x_3x_4。 \quad (2)$$

2.3 有效性判别

按多元回归方法，当 $R>0.7$、$R^2>0.45$ 时，表明因素 x 对 $\boldsymbol{Y}$ 变化的影响在 50% 以上，称为强相关，即认为 x 对 $\boldsymbol{Y}$ 的影响很大；当 $R<0.3$、$R^2<0.09$ 时，表明 x 对 $\boldsymbol{Y}$ 变化的影响不到 10%，称为弱相关，即认为 x 对 $\boldsymbol{Y}$ 影响不大，可忽略；当 $0.3<R<0.7$，称为相关。从表 3 可发现，预测模型的决定系数 R^2 高达 0.915 7，说明式(2)的拟合精度高，是可靠的。表 1 中还列出了据式(2)得出的地表最大沉陷值的预测值及其与实测值的误差，可发现，误差率为±10.41%，这一精度对岩土工程问题是足够的。

得出的方程式(2)，可直接采用作为地表沉陷的预测。不像人工神经网络等方法，这些方法虽然预测精度较高，但它无最终的预测公式，它需要使用者去直接使用人工神经网络等方法，而人工神经网络等方法又是相当复杂的，一般使用者难以掌握和使用。

3 结 语

由以上的分析可看出，偏最小二乘二元多项式回归方法是一种较好的非线性方法，与别的方法相比有一定的优越性。由于地表沉陷的影响因素众多，非常复杂，使得在沉陷预测模型方面的研究还有待加强，因此，希望基于偏最小二乘法的地表沉陷预测模型更加完善。

参考文献：

[1] GAYARRE F, ALVAREZ-FERNANDEZ M I, GONZALEZ-NICIEZA C, et al. Forensic analysis of buildings affected by mining subsidence [J]. Engineering Failure Analysis, 2010,17(1):270-285.

[2] MENG Q J, FENG Q Y, WU Q Q, et al. Distribution characteristics of nitrogen and phosphorus in mining induced subsidence wetland in Panbei coal mine, China [J]. Procedia Earth and Planetary Science, 2009,1(1):1237-1241.

[3] 朱泽兵，张永兴，刘新荣，等. 特大断面车站隧道爆破开挖对地表建筑物的影响[J]. 重庆大学学报，2010，33(2):110-116.

ZHU ZE-BING, ZHANG YONG-XING, LIU XIN-RONG, et al. Influence of blasting vibration on adjacent buildings of station tunne [J]. Journal of Chongqing University, 2010,33(2):110-116.

[4] YUAN G L, LI S M, XU G A, et al. The anti-deformation performance of composite foundation of transmission tower in mining subsidence area [J]. Procedia Earth and Planetary Science, 2009, 1(1): 571-576.

[5] REN W Z, GUO C M, PENG Z Q, et al. Model experimental research on deformation and subsidence characteristics of ground and wall rock due to mining under thick overlying terrane [J]. Internation Journal of Rock Mechanics & Mining Sciences, 2010,47(4): 614-624.

[6] LI W X, WEN L, LIU X M. Ground movements caused by deep underground mining in Guan-Zhuang iron mine, Luzhong, China [J]. International Journal of Applied Earth Observation and Geoinformation, 2010,12(3):175-182.

[7] ALESHINA I N, SNYTKO V A, SZCZYPEK S, et al. Mining-induced ground subsidences as the relief-forming factor on the territory of the Silesian Upland (Southern Poland) [J]. Geography and Natural Resources, 2008,29(3):288-291.

[8] 任松，姜德义，杨春和. 复杂开采沉陷分层传递预测模型[J]. 重庆大学学报，2009,32(7):823-828.
REN SONG, JIANG DE-YI, YANG CHUN-HE. Stratification transfer model for predicting complex mining subsidence [J]. Journal of Chongqing University, 2009,32(7):823-828.

[9] 曹树刚，刘玉成，刘延保，等. 基于观测资料的沉陷盆地主断面曲线拟合[J]. 重庆大学学报，2009,32(7): 804-808.
CAO SHU-GANG, LIU YU-CHENG, LIU YAN-BAO, et al. Curve fitting of main section for subsidence basin by oberving data [J]. Journal of Chongqing University, 2009,32(7):804-808.

[10] WU K, LI L, WANG X L, et al. Research of ground cracks caused by fully-mechanized sublevel caving mining based on field survey [J]. Procedia Earth and Planetary Science, 2009,1(1):1095-1100.

[11] GUEGUEN Y, DEFFONTAINES B, FRUNEAU B, et al. Monitoring residual mining subsidence of Nord/Pas-de-Calais coal basin from differential and persistent scatterer interferometry: Northern France [J]. Journal of Applied Geophysics, 2009,69(1):24-34.

[12] LI W X, LIU L, DAI L F. Fuzzy probability measures (FPM) based non-symmetric membership function: engineering examples of ground subsidence due toundergroud mining [J]. Engineering Applications of Artifcial Intelligence, 2010,23(3):420-431.

[13] 柴洪洲，崔岳，明锋. 最小二乘配置方法确定中国大陆主要块体运动模型[J]. 测绘学报，2009,38(1):61-65.
CHAI HONG-ZHOU, CUI YUE, MING FENG. The determination of Chinese mainland crustal movement model using least-squares collocation [J]. Acta Geodaetica et Cartographica Sinica, 2009,38(1):61-65.

[14] 陈伟根，周恒逸，黄会贤. 变压器油中溶解气体光声光谱检测最小二乘回归定量分析[J]. 重庆大学学报，2010,33(2):22-27.
CHEN WEI-GEN, ZHOU HENG-YI, HUANG HUI-XIAN. Quantitative analysis of photoacoustic spectroscopy detection for dissolved gas in transformer oil based on least-squares regress [J]. Journal of Chongqing University, 2010,33(2):22-27.

[15] 朱洵，荣起国. 基于偏最小二乘回归的基因网络数学模型[J]. 系统仿真学报，2009,21(4):1148-1154.
ZHU XUN, RONG QI-GUO. Modeling of gene networks using partial least square method[J]. Journal of System Simulation, 2009,21(4):1148-1154.

[16] 唐启义，冯明光. DPS 数据处理系统：实验设计、统计分析及数据挖掘[M]. 北京：科学出版社，2007.

[17] 毕忠伟，王春来，丁德馨. MATLAB 神经网络工具箱在矿山开采沉陷中的应用[J]. 采矿技术，2002,2(2): 50-51.
BI ZHONG-WEI, WANG CHUN-LAI, DING DE-XIN. The application of MATLAB neural network in ground subsidence of mining[J]. Mining Technology, 2002,2(2):50-51.

（编辑　赵　静）