

DOI :10.3876/j.issn.1000-1980.2010.02.006

偏最小二乘法与人工神经网络耦合的小流域产沙模型

李世欣^{1,2} 温 建³ 邵孝侯¹ 王晓亚⁴ 王玉英⁴

(1.河海大学水利水电学院,江苏 南京 210098;2.河南农业大学机电工程学院,河南 郑州 450002;
3.河南农业大学信息与管理科学学院,河南 郑州 450002;4.河南省南召县水土保持科学研究站,河南 南召 474650)

摘要:针对小流域侵蚀产沙的复杂性,将偏最小二乘回归与人工神经网络耦合,建立了小流域降雨侵蚀产沙检验模型,并应用于小流域降雨侵蚀产沙预报.采用偏最小二乘法对多维自变量中的信息进行组合和提取,从而得到对因变量解释能力最强并可很好概括自变量信息的主成分,有效克服了变量之间的多重相关问题,实现了对高维数据的降维处理.把提取的主成分作为神经网络的输入,提高了网络的学习效率和稳健性.应用结果表明,偏最小二乘神经网络耦合模型的拟合和检验精度均优于偏最小二乘回归模型和人工神经网络模型精度.

关键词:小流域;产沙;偏最小二乘法;人工神经网络

中图分类号: P333.4 **文献标识码:** A **文章编号:** 1000-1980(2010)02-0149-05

水土流失是当今世界人们普遍关注的重大生态问题之一,已成为我国首要的和最大的环境问题^[1].目前,我国土壤侵蚀面积(包括水蚀、风蚀和冻融蚀面积)达 484.74 万 km²,占全国国土面积的 51.1%^[2].土壤侵蚀的加剧是造成土地荒漠化、洪涝灾害、水土资源危机的主要原因之一.多年来,国内外学者对土壤侵蚀的机理和规律进行了系统研究,并从不同角度、采用不同理论、利用不同影响因子建立了适应于不同地域特征的土壤侵蚀模型,有效指导了土壤侵蚀的预防和治理工作.尤其是近 20 多年来,随着应用数学和信息技术的不断发展,一些新的技术理论被应用到土壤侵蚀规律及应用模型的研究中,如分形理论^[3]、人工神经网络^[4]、RS 和 GIS 技术^[5]等,拓宽了土壤侵蚀的研究思路,提高了土壤侵蚀模型的预测精度.

偏最小二乘回归(partial least-squares regression,PLSR)是一种新型多元统计数据分析法.该方法集多元线性回归分析、典型相关分析和主成分分析的基本功能于一体,通过在变量系统中提取若干个对系统具有最佳解释功能的综合信息变量,实现了对高维数据空间的降维处理.该方法有效解决了自变量间存在严重多重相关性以及样本点个数较少等建模问题.人工神经网络(artificial neural network,ANN)具有模拟人类大脑思维功能的能力,理论上可以实现任意函数的逼近,达到人们希望的精度要求.本文将偏最小二乘回归(PLSR)与神经网络(ANN)结合,建立了偏最小二乘神经网络耦合模型(PLS-ANN),并应用于小流域侵蚀产沙预报.

1 偏最小二乘神经网络模型的构建方法

1.1 偏最小二乘回归模型(PLSR 模型)^[6-8]

设有因变量集合 $Y = \{y_1, y_2, \dots, y_q\}_{n \times q}$ 和自变量集合 $X = \{x_1, x_2, \dots, x_p\}_{n \times p}$,偏最小二乘回归分别在 X, Y 中提取主成分 t_1 和 u_1 ,并要求 (a) t_1 和 u_1 尽可能多地携带它们各自数据表中的变异信息,尽可能好地代表 X 和 Y ; (b) t_1 和 u_1 的相关程度能够达到最大, t_1 对 u_1 有最强的解释能力.

在第 1 成分 t_1 和 u_1 被提取后,偏最小二乘回归分别实施 X 对 t_1 的回归和 Y 对 t_1 的回归.如果回归方程已经达到满意精度,则算法终止;否则,将 X 被 t_1 解释后的残余信息和 Y 被 u_1 解释后的残余信息进行第 2 轮的成分提取,如此往复,直到达到一个较为满意的精度为止.若提取了 h 个主成分 $t_1, t_2, \dots, t_h$,则这 h 个

主成分即为从原数据信息中提取的综合变量.这些变量对系统的解释性最强,并克服了自变量之间的多重相关性,剔除了噪音信息的干扰.

方程精度可采取交叉有效性原则进行检验.在偏最小二乘回归建模中,究竟应选取多少个成分为宜,一般采用交叉有效性原则进行判别^[9-10].除去某个样本*i*的所有集合并使用*h*个成分拟合一个回归方程,把排除的样本点*i*代入前面的回归方程,得到 y_i 在样本点*i*上的拟合值 $\hat{y}_{h|k-i}$.对每一个样本点重复上述计算,定义 y_i 的预测误差平方和为 P_{hj} ,则有 $P_{hj} = \sum_{i=1}^n (y_{ij} - \hat{y}_{h|k-i})^2$.记 Y 的预测误差平方和为 P_h ,则 $P_h = \sum_{j=1}^q P_{hj}$.变量 P_h 取最小值时表明模型的拟合效果最好,这时提取的成分个数*h*即为最佳成分数.

当 $q = 1$ 时,则上述问题由多因变量 PLSR 模型转变为单因变量的 PLSR 模型,它是多因变量 PLSR 模型的一种特例,小流域侵蚀产沙研究的即为单因变量问题.

1.2 偏最小二乘神经网络模型(PLS-ANN)^[8-11]

人工神经网络系统具有大规模的并行处理和分布式信息存储能力,能自动调整内部神经元之间的连接权重,以匹配输入输出响应关系,具有良好的自组织性和学习功能.本文选取典型的3层前馈型 BP(back propagation)网络模型^[4,12].BP 网络模型通常由输入层、隐含层和输出层构成.PLS-ANN 模型构建方法为:用偏最小二乘(PLS)方法提取的*h*个主成分 $t_1, t_2, \dots, t_h$ 作为 BP 模型的输入,输出层为1个节点*y*,隐含层传输函数选择 sigmoid 函数,隐含层节点数的确定采用试错法,经多次调试选择最佳节点个数,模型结构如图1所示.

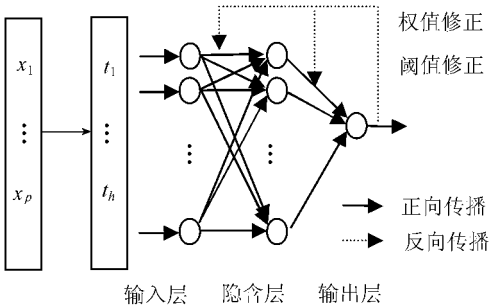


图1 PLS-ANN 模型结构
Fig. 1 Structure of PLS-ANN model

2 小流域侵蚀产沙的 PLS-ANN 模型

2.1 研究区概况

和平沟小流域位于河南省南召县西部,北纬 112°15′~112°18′,东经 33°17′~33°20′之间,面积 9.62 km²,地质岩性为斑状花岗岩,土壤类型主要为沙壤和黄棕壤.多年平均降雨量 1020 mm,年均气温 14.8℃.流域上宽下窄,为一扇形闭合集水单元,林草覆盖率 88%,海拔高程 236~919 m.

南召县水土保持科学实验站在流域内设 3 个基本雨量站,均设雨量计测定降雨量,在小流域下游设总控制断面,测定径流和泥沙.

2.2 数据及方案

影响流域侵蚀产沙的因素主要有地质地貌、土壤类型、降雨径流、植被状况等自然因素和人为因素.对一个流域而言,一定时期内地质地貌、土壤类型等下垫面因素对小流域产流、产沙的影响是相对稳定的,降雨成为影响流域侵蚀产沙的直接动力因子,而只有产生径流的降雨才能产生侵蚀,因此降雨因子和径流因子是侵蚀产沙的重要影响因子^[13-14].林草植被对侵蚀产沙有一定的减缓作用,是影响土壤侵蚀的关键因子^[15].输沙率是水文泥沙监测的重要指标,与输沙量关系密切.事实上,小流域内降雨、土壤、地形、植被、人类活动相互作用结果最终都体现在流域出口处,在洪水和泥沙等特征中表现出来^[16].所以本文采用 1984~2006 年南召县水保站的实测资料,选取流域年降水量 x_1 、年产流雨量 x_2 、年降水日数 x_3 、年径流深 x_4 、年最大 1 日径流量 x_5 、年最大流量 x_6 、林草覆盖率 x_7 、年最大 1 日平均输沙率 x_8 等 8 个因子作为自变量,年输沙量*y*作为因变量.为了评价 PLS-ANN 模型的效果,分别采用 PLSR 模型、BP 神经网络模型和 PLS-ANN 模型进行计算,并对计算结果进行了比较.

2.3 结果与分析

利用前 1984~2002 年的数据建立产沙模型,2003~2006 年的数据用于模型检验.首先,利用偏最小二乘方法对 8 个自变量进行成分提取.根据交叉有效性原则,提取的 3 个成分为最佳成分,见图 2.

3 个主成分 t_1, t_2, t_3 表达式为

$$\begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix} = \begin{bmatrix} 0.4588 & -0.3992 & 0.2090 \\ 0.4465 & -0.1781 & 0.0671 \\ 0.1517 & -0.5666 & 0.5589 \\ 0.4903 & -0.2707 & 0.1490 \\ 0.4545 & -0.1381 & -0.3693 \\ 0.3574 & -0.1462 & -0.5514 \\ -0.0002 & -0.2820 & 0.4031 \\ 0.3621 & 0.6182 & 0.2245 \end{bmatrix}^T E_0$$

其中 E_0 为原始数据 $X([x_1, x_2, \dots, x_8]_{9 \times 8})$ 标准差标准化处理后的矩阵,采用下式进行标准差标准化,即

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j}$$

式中: x_{ij} ——矩阵 X 的第 i 行第 j 列元素的值; $s_j, \bar{x}_j$ ——矩阵 X 第 j 个自变量的标准差和平均值; x_{ij}^* —— x_{ij} 的标准化值.将偏最小二乘法提取的 3 个主成分 t_1, t_2, t_3 作为神经网络的输入, y 作为输出,设定精度为 0.0001,进行学习训练,训练 1368 次就可满足精度要求.若将 8 个自变量作为输入样本,直接输入 BP 神经网络,同样的精度要求则需要训练 6841 次方可收敛.

a. 拟合误差分析.分别用 PLS-ANN 模型、BP 神经网络模型和 PLSR 模型建立产沙模型,对 1984~2002 年的产沙量进行拟合,并将拟合值与观测值进行比较(图 3).

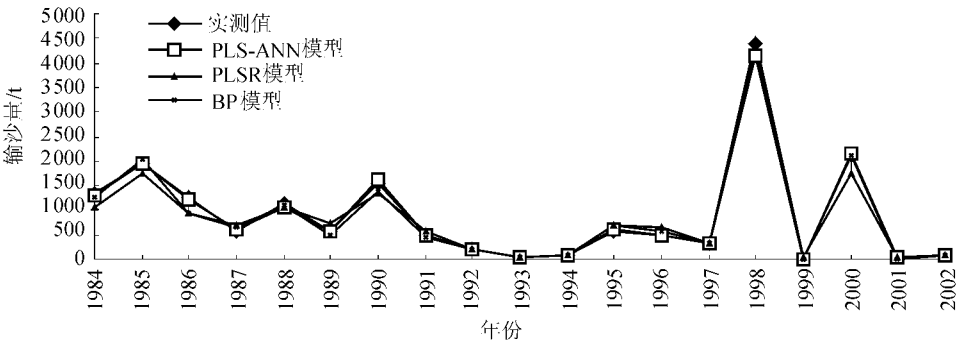


图 3 小流域产沙量拟合精度比较

Fig. 3 Comparison of fitting precision for sediment yield in small watershed

PLSR 模型拟合误差平均为 18.66%,误差在 10% 以下的年份有 6 个,误差大于 50% 的年份有 1 个,BP 神经网络的模型误差稍低,平均值为 13.99%,误差为 10% 以下的年份有 7 个,误差大于 50% 的年份没有,PLS-ANN 模型误差最小,平均值达到 6.57%,仅有 3 个年份的模拟误差大于 10%,最高为 19.92%,其余年份的误差均在 10% 以下.

b. 检验误差分析.应用所建立的 PLS-ANN 模型对 2003~2006 的数据进行检验,并与 PLSR 模型和 BP 神经网络模型进行比较(表 1).由表 1 可看出,BP 网络模型有 1 个年份的误差高达 27.70%,其中有 2 个年份的误差均在 10% 以下,平均误差为 14.11%;PLSR 模型有 2 个年份的误差在 20% 以上,2 个年份的误差在 10% 以下,平均误差为 20.33%,最大误差为 25.27%;PLS-ANN 模型误差均在 20% 以下,最大误差为 16.08%,有 3 个年份的误差均在 10% 以下,平均误差仅为 7.5%.可见,无论是拟合还是检验的精度,偏最小二乘神经网络模型都比单一方法高.

表 1 检验结果对比

年份	实测 输沙量/t	计算输沙量/t			相对误差/%		
		BP 模型	PLSR 模型	PLS-ANN 模型	BP 模型	PLSR 模型	PLS-ANN 模型
2003	123	110.8511	154.0815	129.7032	9.3667	25.2695	5.4498
2004	1260	1247.2780	1450.4590	1057.3450	1.2032	15.1158	16.0838
2005	934	773.5859	721.2124	882.5719	18.1758	22.7824	5.5062
2006	344	251.6670	281.5100	333.2584	27.7061	18.1657	3.1226

3 结 语

a. 流域侵蚀产沙是一个复杂的自然过程,受自然因素、人为因素等诸多因素影响,本文选取 8 个自变量,采取 PLS 方法提取的主成分,不仅携带更多的自变量信息,而且对因变量影响程度最大。本文所建立的 PLS-ANN 模型,既解决了自变量之间相关性和样本数据较少的问题,又降低了神经网络的输入维数,加快了网络收敛速度,增强了网络稳定性,降低了模型的模拟误差和检验误差。

b. PLS-ANN 模型拟合误差平均为 6.57%,检验误差平均为 7.5%,均比 PLSR 模型和 BP 网络模型要低,表明对复杂的非线性小流域产沙系统,PLS-ANN 模型能更好地反映其本质特征。

c. PLS-ANN 模型选取的自变量均为容易获取或测定数据,模型计算方便,实用性较强,对研究豫西南山区水土流失规律和支持当地生态环境建设具有一定实用价值。

参考文献:

- [1] 郑粉莉,王占礼,杨勤科.我国土壤侵蚀科学回顾和展望[J].自然杂志,2008,30(1):12-16.(ZHENG Fen-li,WANG Zhan-li,YANG Qin-ke.The Retrospection and prospect on soil erosion research in China[J]. Chinese Journal of Nature,2008,30(1):12-16.(in Chinese))
- [2] 李智广,曹伟,刘秉正,等.我国水土流失状况与发展趋势研究[J].中国水土保持科学,2008,6(1):57-62.(LI Zhi-guang,CAO Wei,LIU Bing-zheng,et al.Current status and developing trend of soil erosion in China[J]. Science of Soil and Water Conservation,2008,6(1):57-62.(in Chinese))
- [3] 崔灵周,李占斌,郭彦彪,等.基于分形信息维数的流域地貌形态与侵蚀产沙关系[J].土壤学报,2007,44(2):197-203.(CUI Ling-zhou,LI Zhan-bin,GUO Yan-biao,et al. Fractal-information-dimension-based relationship between sediment yield and topographic feature of watershed[J]. Acta Pedologica Sinica,2007,44(2):197-203.(in Chinese))
- [4] 赵西宁,王万忠,吴普特,等.坡面入渗的人工神经网络模型研究[J].农业工程学报,2004,20(3):48-50.(ZHAO Xi-ning,WANG Wan-zhong,WU Pu-te,et al. Artificial neural network model for soil infiltration in slope farmland[J]. Transactions of the Chinese Society of Agricultural Engineering,2004,20(3):48-50.(in Chinese))
- [5] 马修军,谢昆青.GIS 环境下流域降雨侵蚀动态模拟研究:以 PCRaster 系统和 LISEM 模型为例[J].环境科学进展,1998,7(5):137-144.(MA Xiu-jun,XIE Qun-qing. Dynamic simulation of rainfall erosivity by GIS:based on the system of PCRaster and the model of LISEM[J]. Advances in Environmental Science,1998,7(5):137-144.(in Chinese))
- [6] 付强.数据处理方法及其农业应用[M].北京:科学出版社,2006:164-204.
- [7] WOLD S,TRYGG J,BERGLUND A,et al. Some recent developments in PLS modeling[J]. Chemometrics and Intelligent Laboratory Systems,2001,58:131-150.
- [8] 陈南祥,黄强,曹连海.基于偏最小二乘回归与神经网络耦合的岩溶泉预报模型[J].水利学报,2004(9):68-72.(CHEN Nan-xiang,HUANG Qiang,CAO Lian-hai. Model for prediction of karst spring flow based on the coupling of neural network model with partial least square method[J]. Journal of Hydraulic Engineering,2004(9):68-72.(in Chinese))
- [9] EFRON B,GONG G. A leisurely book at the bootstrap,the jackknife and cross validation[J]. The American Statistician,1983,37(1):36-48.
- [10] 曾九孙,刘祥官,罗世华,等.主成分回归和偏最小二乘法在高炉冶炼中的应用[J].浙江大学学报:理学版,2009,36(1):33-36.(ZENG Jiu-sun,LIU Xiang-guan,LUO Shi-hua,et al. Application of principal component regression and partial least square in blast furnace iron-making[J]. Journal of Zhejiang University:Science Edition,2009,36(1):33-36.(in Chinese))
- [11] 戈汉权,施泽进,任在清.基于偏最小二乘与神经网络耦合的储层参数预测[J].成都理工大学学报:自然科学版,2007,34(6):618-620.(GE Han-quan,SHI Ze-jin,REN Zai-qing. Prediction of the reservoir parameters based on the coupling of neural network model with partial least square method[J]. Journal of Chengdu University of Technology:Sci & Technol Ed,2007,34(6):618-620.(in Chinese))
- [12] 朱大奇,史慧.人工神经网络原理及应用[M].北京:科学出版社,2006:36-63.
- [13] 许全喜.人工神经网络模型在流域水沙预报中的应用[J].人民长江,2000,31(5):30-32.(XU Quan-xi. Application of artificial neural network in predicting soil erosion[J]. Yangtze River,2000,31(5):30-32.(in Chinese))
- [14] 蔡强国,刘纪根,刘前进.岔巴沟流域次暴雨产沙统计模型[J].地理研究,2004,23(4):433-439.(CAI Qiang-guo,LIU Ji-gen,LIU Qian-jin. Research of sediment yield statistical model for single rainstorm in Chabagou drainage basin[J]. Geographical Research,2004,23(4):433-439.(in Chinese))
- [15] 张志强,王盛萍,孙阁,等.流域径流泥沙对多尺度植被变化响应研究进展[J].生态学报,2006,26(7):2356-2364.(ZHANG

Zhi-qiang ,WANG Sheng-ping ,SUN Ge et al.Runoff and sediment yield response to vegetation change at multiple scales :a review[J].
Acta Ecologica Sinica 2006 26(7) 2356-2364.(in Chinese))

[16] 侯建才 ,李占斌 ,李勉 ,等 .流域次降雨侵蚀产沙的 BP 神经网络模拟[J].水土保持通报 ,2007 ,27(3) :79-83.(HOU Jian-cai ,
LI Zhan-bin ,LI Mian et al.Back propagation neural network simulation on sediment yield of watershed under single rainfall[J].Bulletin
of Soil and Water Conservation 2007 27(3) :79-83.(in Chinese))

Sediment yield model for small watersheds based on coupling of
partial least square regression and artificial neural network

LI Shi-xin^{1 ,2} , WEN Jian³ , SHAO Xiao-hou¹ , WANG Xiao-ya⁴ , WANG Yu-ying⁴

- (1. College of Water Conservancy and Hydropower Engineering , Hohai University , Nanjing 210098 , China ;
2. College of Mechanical & Electrical Engineering , Henan Agricultural University , Zhengzhou 450002 , China ;
3. College of Information and Management Science , Henan Agricultural University , Zhengzhou 450002 , China ;
4. Nanzhao Experimental Station of Soil and Water Conservation , Nanzhao 474650 , China)

Abstract : With regard to the complexity of sediment yield in small watersheds , a coupling partial least square regression and artificial neural network model was proposed by combining the partial least square regression with the artificial neural network . The proposed model was applied to the prediction of sediment yield owing to precipitation in small watersheds . The information of multi-dimensional independent variables was combined and extracted by use of the partial least square regression so as to find the most important components with strong interpretation capacity for dependent variables and satisfactory depiction for independent variables . Accordingly , the problem of multi-correlations among variables could be solved , and the amount of input dimensions could be reduced . The extracted components were regarded as the input of the artificial neural network . The learning efficiency and robustness of the network were improved . The results show that the proposed model is of higher fitting and prediction precisions than the models based on the partial least square regression or the artificial neural network .

Key words : small watershed ; sediment yield ; partial least square regression ; artificial neural network