中国农业大学学报 2010,15(6):102-107
Journal of China Agricultural University

基于支持向量机和神经网络的土壤水力学参数预测效果比较

聂春燕[1] 胡克林[2] 邵元海[1] 陈 薇[1*]
(1.中国农业大学 理学院,北京 100193; 2.中国农业大学 资源与环境学院,北京 100193)

摘 要 在美国土壤水分物理性质数据库(UNSODA 2.0)的基础上,考虑土壤质地不分类和分类2种情况,分别构建了基于支持向量回归机(SVR)的土壤传递函数模型,比较了在土壤质地不分类和分类情况下预测土壤水力学参数(水分特征曲线和饱和导水率)的效果,并与建立在相同数据库上的基于神经网络的Rosetta模型的预测效果进行了比较。结果表明:土壤质地不分类的情况下,输入参数越多,基于SVR模型的预测效果越好;土壤质地分类情况下,基于SVM分类建模的预测结果普遍好于不分类情况。无论土壤质地是否分类,样本和输入参数相同的条件下,基于SVR的模型预测的效果都优于Rosetta模型。

关键词 土壤质地分类; 传递函数; 支持向量回归机; 神经网络
中图分类号 S 117 **文章编号** 1007-4333(2010)06-0102-06 **文献标志码** A

Comparison of predicting results of soil hydraulic parameters by SVR and rosetta models

NIE Chun-yan[1], HU Ke-lin[2], SHAO Yuan-hai[1], CHEN Wei[1*]
(1. College of Science, China Agricultural University, Beijing 100193, China;
2. College of Resources and Environmental Sciences, China Agricultural University, Beijing 100193, China)

Abstract Based on the database of America soil water physical characteristics (UNSODA 2.0), the PTFs models for both classified and unclassified soil textures were constructed by using Support Vector Regression (SVR) approach, and the predictive results of soil texture under both classified and unclassified situations for soil hydraulic parameters (V-G model parameter and soil saturated hydraulic conductivity), as well as the results from the Rosetta model based on neural network and constructed by the same database, were compared. The results indicated that the more the number of parameters input, the better the results obtained by using SVR model in the situation of soil texture unclassified; however, in the situation of the classified, the results obtained from separately constructed SVM model are generally better than that of the unclassified. No matter whether the soil texture classified or not, on the occasion of the same sample and parameters input, the results obtained from the model based on SVR have an advantage over Rosetta model.

Key words soil texture classification; pedotransfer functions; support vector regression; neural network

土壤水力学参数(水分特征曲线和导水率)是模拟水分和溶质在土壤中运移的关键参数,对于水分、盐分、养分和污染物等物质平衡的计算非常重要。对于大多数农田土壤水分和溶质运动的研究,能够通过实际测定获得这些参数,但对于区域尺度,这些参数的测定费时耗力,且具有很大的空间变异性,难以采用实测方法获取足够的数据。研究表明,进行大量高精度的实测是不必要的,一些粗略的估算公式完全可以满足区域模拟计算的需要[1]。因此,如何通过一些已知的容易测定的土壤基本性质获取土

收稿日期: 2010-04-19
基金项目: 国家科技支撑计划项目(2008BADA7B05); 公益性行业科技项目(200803036); 教育部新世纪优秀人才支持计划项目(NCET-07-0809)
第一作者: 聂春燕,硕士研究生,E-mail:niechunyanzhy@163.com
通讯作者: 陈薇,教授,主要从事统计分析、可拓学及其应用研究,E-mail:chenwei@cau.edu.cn

壤水力学参数的研究具有现实意义。土壤传递函数(pedotransfer functions，PTFs)[2-6]是目前应用最为广泛的获取土壤水力学参数的一种间接方法[7]，可以根据土壤基本性质如粒径分布、容重、土壤孔隙度、有机质含量和土壤结构等估计土壤水力学参数。

目前已经开发了很多 PTFs 模型。按照其理论基础，可分为统计模型[8]、物理经验模型[9]、分形方法[10]等。常用的模型主要有回归统计模型和基于神经网络的模型[11]，其中最具代表性的是基于回归统计的 Vereecken 模型[12]、HYPRES 模型[13]及基于神经网络的 Rosetta 模型[14]。回归分析是最早也是应用最多的一种方法[15-17]，人工神经网络与回归分析方法相比，主要优点在于不需要先验假设，只通过迭代校验过程便可得到输入层(土壤基本理化性质)和输出层(土壤水力学参数)之间的最优关系。一些研究结果表明，人工神经网络的预测误差要明显小于普通的线性回归技术[18-19]。

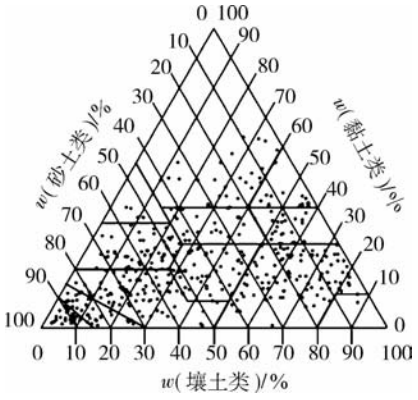
近年来支持向量回归机[20](support vector regression，SVR)也用来建立 PTFs 模型，能较好地解决小样本、非线性、高维数和局部极小点等实际问题。Lamorski 应用 SVR 建立了波兰的 PTFs 模型对土壤水力学参数进行了预测，发现 3 参数的基于 SVR 的 PTFs 模型与 7 参数的基于人工神经网络的 PTFs 模型相比，预测精度相当或更好[21]。杨绍锷等利用黄淮海平原曲周县的 143 点试验资料建立了基于 SVR 的 PTFs 模型，结果表明预测值和实测值不存在显著性差异，用该方法预测土壤水力学参数是可行的，但由于受到样本数少的限制，该结果需要进一步证实[22]。

已有方法和模型都是将所有质地的土壤作为一个整体进行估算的，很少将土壤按照其质地性质的不同进行分类。不同质地的土壤，无论基本理化性质还是水力学性质相差都很大，必然导致结果存在较大误差。为了更精确的估算土壤水力学参数，本研究拟在美国土壤水分物理性质 UNSODA 数据库的基础上，按照土壤质地性质的不同对土壤进行分类，利用 SVR 分别构建土壤传递函数模型，并将预测结果与基于神经网络 Rosetta 模型的结果进行比较，探讨此方法的可行性。

1　材料与方法

1.1　数据来源及分类

美国盐土实验室建立的 UNSODA 数据库，收集了世界各地(12 个国家 140 多个地区)的土壤基本物理性质(土壤粒径分布、土壤容重、有机物含量等)和水力学性质(土壤水分特征曲线、饱和与非饱和导水率)[22]，土壤质地分布情况见图 1。



粒径：砂粒，>50 μm；粉粒，2～50 μm；黏粒，<2 μm。

图 1　基于美国制土壤质地三角形 UNSODA 数据分布

Fig. 1　Distribution of soil data across the USDA-SCS soil textural triangle

根据最常用的 Van Genuchten (V-G)模型中参数的定义域($0\leqslant\alpha\leqslant0.3$，$1\leqslant n\leqslant5$，$0\leqslant\theta_r\leqslant0.2$)，过滤掉其中不理想的数据，最后得到 316 组有效数据用于土壤传递函数建模和验证。由于饱和导水率 K_s 在 UNSODA 数据库中有缺失，只得到 315 组有效数据可以用于 K_s 的建模和验证。

本研究主要根据砂粒(粒径＞50 μm)、粉粒(2≤粒径≤50 μm)和黏粒(粒径＜2 μm)3 粒级的组成进行分类，分别为砂土类、壤土类、黏土类及中间类共 4 类(图 2)。每个组的样本数分别为 207、2、62 和 45；饱和导水率 K_s 的数据分组情况为 173、11、73 和 58。

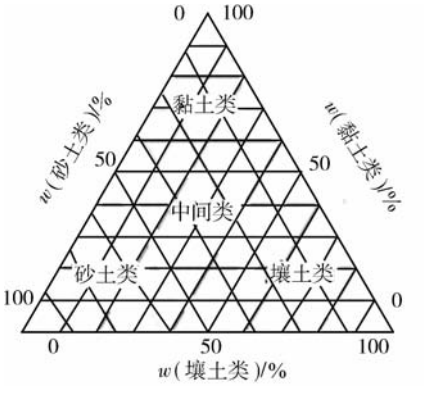


图 2　自定义土壤分类

Fig. 2　Self-defined soil textural classification

选取土壤质地和容重作为聚类分析变量，采用欧式距离衡量样本间亲疏程度。将样本分为 N 类，并按照 60%和 40%的比例随机选取训练和检验样本，这样样本可以按照属性随机抽取。试验中的数据均归一化(0-1)。

1.2 Rosetta 模型

Rosetta 是由美国盐土实验室 Marcel G. Schaap 构建的一个应用模型。Rosseta 模型基于 BP 神经网络理论建立[23-24]，用于估算饱和含水量、残余含水量、van Genuchten 参数；该模型的建立用了 2 085 个数据，其中用于估算饱和导水率的模型使用了 1 306 个数据。该模型提供了分层的 PTFs，这种分等级的处理允许使用者灵活地提供数据选择合适的模型。

1.3 支持向量回归机

设 $\Omega = \{(x_1, y_1), (x_2, y_2), \cdots, (x_1, y_1)\}$ 为样本训练数据，其中：x_i 为输入样本，$x_i \in \mathbf{R}^n$；y_i 为输出样本，$y_i \in \mathbf{R}^n$；$i = 1, 2, \cdots, l$，l 为样本数。SVR[26]的基本思想是定义线性超平面 $y = f(x)$，并把寻找最优线性超平面的算法归结为求解一个凸规划问题，进而运用 Mercer 核 $k(x, x')$ 展开定理，通过非线性映射 $x \to \varphi(x)$，把样本空间映射到一个高维乃至无穷维的特征空间(Hilbert 空间)，在特征空间中应用线性学习机的方法解决样本空间中的高度非线性分类和回归等问题。

根据 Lagrange 函数和 KKT 条件得：

$$f(x) = (\omega \cdot \varphi(x)) + b = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)(\varphi(x) - \varphi(x_i)) + b = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x \cdot x_i) + b \qquad (1)$$

式中：$\omega \in \mathbf{R}^n$；b 为常量，$b \in R$；α_i, α_i^* $(i = 1, 2, \cdots, l)$ 为 lagrange 乘子。式(1)即为 SVR 方法最终确定的非线性回归函数。支持向量机在计算 $f(x)$时，无需计算权向量 ω 和非线性映射 $\varphi(x)$的具体数值，而只需计算 Lagrange 乘子 α_i 和 α_i^* 以及核函数 $k(x, x_i)$即可。与传统回归方法相比，SVR 的主要优点是：应用了核函数的展开原理，可以在高维特征空间中应用线性学习机的方法，所以与线性模型相比几乎不增加计算的复杂性；应用了 ε-不敏感损失函数，使得 SVR 在一个小的 ε-带内不计算损失，从而抗干扰能力强，提高了泛化能力。通过比较分析，本研究选择常用的 Mercer 核——RBF 核函数：

$$k(x, x_i) = \exp\left(-\frac{\| x - x_i \|^2}{\gamma^2}\right) \qquad (2)$$

式中 γ 为核参数。本研究中参数的选取按照十折交叉验证准则[26]。

1.4 模型检验及数据处理

通过分析土壤水力学参数的预测值与实测值的差异程度，采用决定系数 R^2 作为判断指标评价土壤转移函数的预测效果，R^2 的取值越接近 1，表示预测效果越好。

土壤样本的分类和聚类分析采用了 SPSS 统计软件[27]。SVR 模型的构建采用 LIBSVM 软件。

2 结果与分析

2.1 土壤质地不分类 2 种模型的预测效果

从 UNSODA 数据库中挑选出含有土壤质地、容重、饱和含水量 θ_s、残余含水量 θ_r、饱和导水率 K_s 以及根据实测水分特征曲线拟合得到的 V-G 模型参数 α 和 n 作为模型比较的样本，得到检验样本数 166，UNSODA 数据库中饱和导水率 K_s 有缺失，因此检验 K_s 的样本数只有 103 个。分别采用这 2 个样本(样本数为 166 和 103)判断不同输入情况下 Rosetta 模型和 SVR 模型的预测效果。

SSC-BD 表示输入参数为砂粒、粉粒、黏粒和容重；SSC-BD-T33-T1500 表示输入参数为粒径组成、容重以及压力分别为 33 和 1 500 kPa 下的含水量。由表 1 可以看出，2 类模型在输入参数较多(6 个)情况下的预测效果均好于输入参数较少(4 个)的情况；输入参数较少(4 个)的情况下，SVR 模型预测得到的结果明显优于 Rosetta 模型，但是 2 种模型对参数 θ_r 和 α 的预测效果均很差。在输入参数较多(6 个)的情况下，除参数 n 和 K_s 外，SVR 模型预测得到的土壤水力学参数结果要明显优于 Rosetta 模型，而且 SVR 模型预测得到的参数 θ_r 和 α 均远好于 Rosetta 模型。

表 1　不同输入参数个数 Rosetta 和 SVR 模型预测参数的决定系数 R^2

Table 1　R-square obtained by the SVR model with the Rosetta model under different inputs

输入参数①	输出参数	样本个数	R^2	
			Rosetta	SVR
SSC-BD	θ_r	166	0.015 6	0.091 2
	θ_s	166	0.641 1	0.677 1
	$-\lg\alpha$	166	0.148 3	0.232 1
	$\lg n$	166	0.483 9	0.493 3
	$\lg K_s$	114	0.320 3	0.361 5
SSC-BD-T33-T1500	θ_r	166	0.018 1	0.202 4
	θ_s	166	0.652 3	0.644 7
	$-\lg\alpha$	166	0.147 6	0.515 5
	$\lg n$	166	0.851 7	0.750 1
	$\lg K_s$	114	0.449 1	0.383 9

注：①SSC 表示粒径组成，即砂粒、粉粒和黏粒；BD 为容重，g/cm^3；T33 和 T1500 分别表示压力为 33 和 1 500 kPa 下的土壤含水量，cm^3/cm^3。θ_r 和 θ_s 分别为残余和饱和含水量，cm^3/cm^3；α 和 n 为 V-G 模型参数；K_s 为饱和导水率，cm/d。下同。

2.2　土壤质地分类 2 种模型的预测效果

将 UNSODA 中的有效数据，进行自定义土壤分类(图 2)。由于黏土类样本较少(<11 个)，故不参与建模及分析。选取土壤质地及容重作为分类指标，采用欧式距离衡量样本间的亲疏程度，将样本进行聚类分析，并按照 60%和 40%的比例随机选取训练集和检验集样本(表 2)。

不同土壤质地 SVR 模型与 Rosetta 模型预测效果的比较见表 3。可以看出，对于砂土类在检验样本和输入参数都相同的情况下，除了 $\lg n$ 预测值与实测值的 R^2 小于分类的 Rosetta 模型外，不同土壤质地的 SVR 模型对这些参数的预测值和实测值的 R^2 均好于 Rosetta 模型。另外，不同土壤质地的 SVR 模型对于砂土类这 5 个参数的预测精度基本上都随着输入参数的增加而增加。对于壤土类在同样检验样本和输入参数的情况下，不同质地的 SVR 模型对这 5 个参数的预测值和实测值的 R^2 均好于 Rosetta 模型，也就是说分类建模后对各参数的预测精度更高。自建模型对于壤土类各个参数的预测精度都随着输入参数的增加而增加。对于中间类在检验样本和输入参数都相同的情况下，自建模型对参数 θ_r、θ_s、$\lg\alpha$、$\lg n$ 的预测值和实测值的 R^2，基本上好于 Rosetta 模型和不同质地的 BP-ANN 模型，也就是说分类建模后的预测精度更高。而对于饱和导水率 K_s 的预测，由于样本的数量有限，训练样本仅有 30 多个，故预测效果非常差，这可能与该类所包含的土壤质地性质比较复杂有关。

表 2　经 SPSS 聚类和抽样后的样本数据

Table 2　Distribution of the samples after clustering and case selecting with SPSS

土壤质地	V-G 模型样本			K_s 样本		
	60%	40%	聚类	60%	40%	聚类
砂土类	8	2	1	10	15	1
	1	1	2	21	13	2
	3	2	3	4	3	3
	20	14	4	13	7	4
	66	49	5	5	4	5
	1	1	6	15	9	6
	7	2	7	17	6	7
	12	8	8	8	6	8
	1	1	9	8	3	9
	4	3	10	3	3	10
壤土类	31	19	1	11	9	1
	2	2	2	19	12	2
	2	2	3	6	4	3
	1	1	4	6	3	4
	1	1	5	1	1	5
中间类	2	1	1	6	4	1
	12	5	2	5	4	2
	4	3	3	12	7	3
	8	5	4	2	2	4
	3	2	5	11	5	5

表 3 不同土壤质地 SVR 模型与 Rosetta 模型预测参数的决定系数 R^2

Table 3 R-square obtained by the SVR model for sandy soil with the Rosetta model in different soil texture

土壤质地	模 型	输 入	θ_r	θ_s	lgα	lg$n$	lgK_s
砂土类	SVR	SSC	0.205 4	0.208 7	0.113 5	0.317 6	0.307 5
		SSC-BD	0.248 1	0.528 9	0.329 1	0.380 9	0.347 5
		SSC-BD-T33	0.345 6	0.580 4	0.507 8	0.438 4	0.450 9
		SSC-BD-T33-T1500	0.850 6	0.636 5	0.476 6	0.680 9	0.451 7
	Rosetta	SSC	0.048 4	0.171 9	0.106 6	0.330 8	0.227 5
		SSC-BD	0.057 7	0.542 8	0.074 0	0.338 4	0.311 1
		SSC-BD-T33	0.132 4	0.574 7	0.075 3	0.394 5	0.366 6
		SSC-BD-T33-T1500	0.762 8	0.639 7	0.000 2	0.778 1	0.385 5
壤土类	SVR	SSC	0.543 7	0.660 8	0.139 6	0.595 5	0.757 7
		SSC-BD	0.689 3	0.695 3	0.650 9	0.828 8	0.793 7
		SSC-BD-T33	0.705 0	0.714 6	0.478 3	0.916 0	0.624 7
		SSC-BD-T33-T1500	0.705 0	0.745 8	0.392 2	0.949 9	0.784 3
	Rosetta	SSC	0.046 1	0.364 9	0.059 6	0.344 8	0.196 3
		SSC-BD	0.067 5	0.332 8	0.025 8	0.266 1	0.352 1
		SSC-BD-T33	0.098 5	0.372 2	0.799 9	0.144 9	0.412 2
		SSC-BD-T33-T1500	0.050 7	0.379 9	0.204 3	0.811 8	0.507 3
中间类	SVR	SSC	0.106 5	0.336 6	0.283 2	0.536 7	0.291 4
		SSC-BD	0.126 5	0.801 7	0.638 4	0.638 4	0.293 5
		SSC-BD-T33	0.282 5	0.967 8	0.768 4	0.769 4	0.293 6
		SSC-BD-T33-T1500	0.184 9	0.932 1	0.358 4	0.354 9	0.303 8
	Rosetta	SSC	0.023 9	0.336 5	0.008 8	0.303 3	0.026 5
		SSC-BD	0.092 0	0.867 3	0.392 1	0.035 7	0.445 0
		SSC-BD-T33	0.018 3	0.816 1	0.217 3	0.000 3	0.533 2
		SSC-BD-T33-T1500	0.136 3	0.736 6	0.002 8	0.165 9	0.574 8

3 结 论

本研究在美国土壤水分物理性质数据库的基础上，构建了基于 SVR 的土壤传递函数模型，并且与建立在相同数据库上的基于神经网络的 Rosetta 模型的预测效果进行了比较，主要结论如下：

1）在土壤质地不分类的情况下，输入参数越多，两类土壤传递函数模型的预测效果越好。在相同样本和输入参数的情况下，基于 SVR 的传递函数模型预测得到的土壤水力学参数（V-G 模型参数和饱和导水率）的效果优于 Rosetta 模型。

2）在土壤质地分类的情况下，除了中间类，大多数情况下输入参数越多，模型预测性能越好。这与中间类所包含的土壤质地性质可能比较复杂有关。在绝大部分情况下，在相同样本和输入参数的情况下，基于 SVR 的传递函数模型预测得到的土壤水力学参数的效果要优于 Rosette 模型。

3）无论是基于 SVR 的土壤传递函数，还是 Rosetta 模型，除了中间类，土壤质地分类情况下的模型预测效果普遍好于质地不分类情况下的模型预测效果，这说明了土壤质地分类情况下所建立的模型更有针对性，需要在实践中进一步检验。

参 考 文 献

[1] Saxton K E, Rawls W J, Romberger J S. Estimating generalized soil-water characteristics from texture [J]. Soil Sci Soc Am J, 1986, 50: 1031-1036

[2] Wösten J H M, Pachepsky Y A, Rawls W J. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics [J]. Journal of Hydrology, 2001, 251: 123-150

[3] Wosten J H M, Van Genuchten M Th . Using testure and other soil properties to predict the unsaturated soil hydraulic functions [J]. Soil Sci Soc Am J, 1988, 52: 1762-1770

[4] 胡振琪，张学礼. 基于 ANN 的复垦土壤水分特征曲线的预测研究 [J]. 农业工程学报，2008，24(10)：15-19

[5] Rawls W J, Gish T J, Brakensiek D L. Estimating soil water retention from soil physical properties and characteristics [J]. Advanced Soil Science, 1991, 16: 213-234

[6] Van Genuchten M Th. A closed form equation for predicting the hydraulic conductivity of unsaturated soils [J]. Soil Sci Soc Am J, 1980, 44: 892-898

[7] Børgesen C D, Schaap M G. Point and parameter pedotransfer functions for water retention predictions for Danish soils [J]. Geoderma, 2005, 127: 154-167

[8] Arya L M, Paris J F. A physicoempirical model to predict the soil moisture. characteristic from particle-size distribution and hulk density data [J]. Soil Sci Soc Am J, 1981, 45: 1023-1030

[9] 刘建立，徐绍辉. 据颗粒大小分布估计土壤水分特征曲线：分形模型的应用 [J]. 土壤学报，2003，40(1)：46-52

[10] Schaap M G, Leij F J, Van Genuchten M Th. Neural network analysis for hierarchical prediction of soil water retention and saturated hydraulic conductivity [J]. Soil Sci Soc Am J, 1998, 62: 847-855

[11] 郭焱，李保国. 预测土壤水分运动参数的 PTFs 法[C]//石元春，刘昌明，龚元石. 节水农业应用基础研究进展. 北京：中国农业出版社，1995：56-63

[12] 黄元仿，李韵珠. 土壤水力性质的估算：土壤传递函数 [J]. 土壤学报，2002，39(4)：517-523

[13] 高如泰，陈焕伟，李保国，等. 基于 BP 神经网络的土壤水力学参数预测[J]. 土壤通报，2005，36(5)：641-646

[14] Gupta S C, Larson W E. Estimating soil water retention characteristics from particle size distribution, organic matter percent, and bulk density [J]. Water Resources Research, 1979, 15: 1633-1635

[15] Rawls W J, Brakensiek D L, Saxton K E. Estimation of soil water properties [J]. Transactions of the ASAE, 1982, 25: 1316-1320

[16] Vereecken H, Maes J , Feyen J , et al. Estimating the soil moisture retention characteristic from 124 texture, bulk density, and carbon content [J]. Soil Science, 1989, 148: 389-403

[17] Schaap M G, Bouten W. Modeling water retention curves of sandy soils using neural networks[J]. Water Resources Research, 1996, 32: 3033-3040

[18] Minmsny B, McBratney A B, Bristow K L. Comparison of diferent approaches to the development of pedotransfer functions for water-retention curves [J]. Geoderma, 1999, 93: 225-253

[19] 邓乃扬，田英杰. 数据挖掘中的新方法：支持向量机 [M]. 北京：北京科学出版社，2004

[20] Cristianini N, Shawe-Taylor J. 支持向量机导论[M]. 李国正，王猛，曾华军，译. 北京：电子工业出版社，2004

[21] Lamorski K, Pachepsky Y, Sławiński C, et al. Using support vector machines to develop pedotransfer functions for water retention of soils in poland [J]. Soil Sci Soc Am J, 2008, 72: 1243-1247

[22] 杨绍锷，黄元仿. 基于支持向量机的土壤水力学参数预测 [J] . 农业工程学报，2007，23 (7)：42-47

[23] 袁曾任. 人工神经元网络及其应用[M]. 北京：清华大学出版社，1999：66-131

[24] Pachepsky Y A, Timlin D, Varallyay G. Artificial neural networks to estimate soil water retention from easily measurable data [J]. Soil Sci Soc Am J, 1996, 60: 727-733

[25] Vapnik V. The Nature of Statistical Learning Theory [M]. New York: Spring, 1995

[26] 卢纹岱. SPSS for Windows 统计分析[M]. 北京：电子工业出版社，2006