doi: 10.3969/j.issn.1006-1576.2010.12.001

基于偏最小二乘回归分析的试验装备修理成本预测

张翀[1]，郑绍钰[2]，王璐璐[3]
（1. 装备指挥技术学院 研究生管理大队，北京 101416；2. 装备指挥技术学院 装备采办系，北京 101416；
3. 中国民航大学 电子信息工程学院，天津 300300）

摘要：为了科学预测试验装备修理成本，提高维修经费决策质量，引入偏最小二乘回归分析（Partial Least Squares Regression, PLSR）对试验装备修理成本进行预测。针对试验装备修理成本小样本、贫数据、特征量相关性强的不利条件，构建预测模型；基于以往数次大修相关数据，预测试验专用装备使用期的某次大修成本。同时，为保持模型的稳健性，提高模型解释能力和预测精确度，尝试利用变量投影重要性分析对模型进行优化，取得了较好的效果。实例证明，该方法不仅能在多变量间存在严重多重相关性情况下建立模型，而且能够有效筛选与因变量关系不大的自变量，简化输入样本集。

关键词：偏最小二乘回归分析；试验装备；修理成本；预测；变量投影重要性分析
中图分类号：O241.5; E257　　**文献标识码**：A

Forecast of Tentative Equipment Repair Cost Based on Partial Least Squares Regression

Zhang Chong[1], Zheng Shaoyu[2], Wang Lulu[3]
(1. Administrant Brigade of Postgraduate, Institute of Command & Technology of Equipment, Beijing 101416, China;
2. Dept. of Equipment Acquisition, Institute of Command & Technology of Equipment, Beijing 101416, China;
3. College of Electronic Information Engineering, Civil Aviation University of China, Tianjin 300300, China)

Abstract: In order to forecast tentative equipment's repair cost scientifically, and improve the decision-making quality of maintenance outlay, partial least squares regression (PLSR) is introduced to forecast tentative equipment's repair cost. Aiming at the limitation of tentative equipment repair cost's small sample, inadequate statistics, close relative eigenvector, forecasting model is constructed; based on several heavy repair data before, the heavy repair cost of special tentative equipment in use is forecasted. Meanwhile, it has the good effect to attempt optimizing the model by using variable importance in projection (VIP) in order to keep the model's stability and improve its explaining ability and forecasting accuracy. It is proved by examples that this method can not only construct models in the case that high multi-correlation exist between variables, but also filter effectively independent variable which is of little relation to dependent variable, and simplify sample set.

Keywords: partial least squares regression; tentative equipment; repair costs; forecast; variable importance in projection

0　引言

在我军试验装备维修保障中，维修费用直是制约性关键矛盾。作为维修费用主体部分的修理成本，由于军方对其了解不够，难以科学确定数额，导致在与地方承修单位签订合同时处于被动局面，直接影响了维修费用的使用效益。如何科学预测试验装备修理成本，提高维修经费决策质量，是当前必须高度关注的问题。试验装备修理成本预测是一项复杂的系统工程。试验装备的单件和小批量特点，为预测数据收集造成了较大困难；试验装备涵盖范围广、差异大，分类界定尚不明确，特征量繁多、且相关性强，为建立预测模型又带来了不利影响。因此，笔者在密切关注国内外小样本、贫数据、相关性强条件下预测建模的发展动态基础上，引入偏最小二乘回归分析法（Partial Least Squares Regression, PLSR）构建预测模型，以预测试验专用装备使用期的某次大修成本。

1　偏最小二乘回归分析法

1.1　原理

PLSR 建模包含基本回归和交叉有效性分析 2 个部分：

设有单因变量 y 和 k 个自变量 $\{x_1, x_2, \cdots, x_k\}$，观测 n 个样本点，构成数据表 $X = \{x_1, x_2, \cdots, x_k\}_{n\times k}$ 和

收稿日期：2010-06-28；修回日期：2010-08-13
基金项目：2009 年国家社会科学基金重大项目“国防建设科学发展重大问题研究”（09&ZD066）
作者简介：张翀（1984-），男，四川人，硕士，助理工程师，在读博士，从事装备采办理论及应用。

$Y=\{y\}_{n\times 1}$。PLSR 分别在 X 和 Y 中提取主成分 t_1 和 u_1，并满足 2 个条件：一是两组成分分别最大限度地承载变量的变异信息；二是对应成分的潜因变量与潜自变量之间协方差最大化。当第一成分 t_1 和 u_1 被提取后，PLSR 分别实施 X 对 t_1、Y 对 u_1 的回归。若回归已达到满意精度，则算法终止；否则，将利用 X 被 t_1 及 Y 被 t_1 解释后的残余信息进行第二轮成分提取，如此往复。当对 X 提取 $h(h=1,2,\cdots,m)$ 个成分时达到满意精度，则实施 y 对潜变量 $t_1,t_2,\cdots,t_m$ 的回归，最后可还原成 y 关于原变量 $x_1,x_2,\cdots,x_k$ 的回归方程。

可见，PLSR 建模的关键是成分确定方法。一般地，PLSR 不必选用全部成分，而是采用“截尾”方式，以交叉有效性作为判定成分个数的标准。交叉有效性的验证如下：

设 y_i 为原始数据，$t_1,t_2,\cdots,t_m$ 为基本回归中提取的成分；$\hat{y}_{hi}$ 是使用全部样本点并取 m 个成分回归后，第 i 个样本的拟合值；$\hat{y}_{h(-i)}$ 为分析时删去第 i 个样本点并取 m 个成分回归后，模型计算 y_i 的拟合值，分别计算前、后者的预测误差平方和 $S_{ss,h}$、$S_{SPRESS,h}$。一般地，总有 $S_{SPRESS,h}>S_{ss,h}$，而 $S_{ss,h}<S_{ss,h-1}$（其中 $S_{ss,h-1}$ 是用全部样本点回归后的 $(m-1)$ 个成分方程的拟合误差）。当 $S_{SPRESS,h}/S_{ss,h-1}\le 0.95^2$ 时，即：$Q_h^2=\left(1-S_{SPRESS,h}/S_{ss,h-1}\right)\ge(1-0.95^2)=0.0975$，说明加入的新成分 t_k 的贡献显著[2]。

1.2 辅助分析

PLSR 建模集合了主成分分析、典型相关性分析和回归分析的特点，可利用 PLSR 的辅助分析技术，考证模型建立的优劣度和可靠性。

1.2.1 变量投影重要性分析

PLSR通过引入变量投影重要性指标（Variable Importance in Projection, VIP）测度自变量对因变量的解释能力。由于 x_j 对 y 的解释是通过 t_h 传递的，若 t_h 对 y 的解释能力很强，而 x_j 构造 t_h 时具有重要作用，则 x_j 对 y 的解释能力被视为很大，可定义为：

$$VIP_j=\sqrt{\frac{k}{Rd(y;t_1,\cdots,t_m)}\sum_{h=1}^{m}Rd(y;t_h)w_{hj}^2} \tag{1}$$

式中[3]：w_{hj} 用于衡量 x_j 对构造 t_h 的边际贡献；$Rd(y;t_h)$ 代表 t_h 对 y 的解释能力，$Rd(y;t_1,\cdots,t_m)$ 代表 $t_1,t_2,\cdots,t_m$ 对 y 的累计解释能力。

当 k 个自变量 x_j 在解释 y 时作用相同，则所有 VIP_j 均为 1；否则，对于 VIP_j 越大的 x_j，解释 y 的作用更为重要[3]。一般认为，VIP_j 大于 1 的自变量重要，在 0.5 到 1 之间比较重要，小于 0.5 则不重要。

1.2.2 特异点分析

类似主成分分析，定义样本点 i 对成分 $t_1,t_2,\cdots,t_m$ 的累计贡献率 T_i^2，用于发现样本中的特异点，即：

$$T_i^2=\frac{1}{(n-1)}\sum_{h=1}^{m}\frac{t_{hi}^2}{s_h^2} \tag{2}$$

式中，s_h^2 为成分 t_h 的方差。

T_i^2 值一般不易过大。Simca-p+12.0 软件采用了 Tracy 证明的统计量 $n^2(n-m)T_i^2/m(n^2-1)\sim F(m,n-m)$，当 $T_i^2\ge m(n^2-1)F_{0.05}(m,n-m)/n^2(n-m)$ 时，可认为在 95% 检验水平上，样本点 i 对成分 $t_1,t_2,\cdots,t_m$ 的贡献过大，此时称样本点 i 为特异点。特别地，当提取 2 个成分时，此判别条件为：

$$\left(\frac{t_{1i}^2}{s_1^2}+\frac{t_{2i}^2}{s_2^2}\right)\ge\frac{2(n-1)(n^2-1)}{n^2(n-2)}F_{0.05}(2,n-2) \tag{3}$$

该式实质可看作一个 T^2 椭圆[4]。若所有样本点都在椭圆内，认为样本分布均匀；否则，落在椭圆外的点是特异点。

2 基于 PLSR 的预测模型构造策略

基于 PLSR 建模原理，其预测模型的构造策略如下：

1) 收集特征量数据，建立样本集，标准化处理后确定自变量与因变量集合；

2) 利用 PLSR 回归分析，经交叉有效性验证后，逐步提取适当个数的主成分；

3) PLSR 回归拟合，得到关于潜变量的拟合方程，代换为标准化变量方程后，经标准化的逆过程，得到因变量 y 关于自变量 x 的函数式；

4) 再次收集自变量数据，建立预测样本集，带

入函数式可得到预测值。

3　基于 PLSR 预测模型的实证分析

为验证 PLSR 预测模型的正确性和可靠性，以某类型试验装备为例进行模型实证分析，收集同类十型现役试验装备 1989～2004 年发生的 25 次大修数据如表 1（鉴于保密性，对数据进行了一定处理）。

表 1　十型试验装备大修成本相关历史数据

	Y/万	*X*1/万	*X*2/万	*X*3/万	*X*4/h	*X*5/h	*X*6	*X*7	*X*8/‰	*X*9/km	*X*10[5]	*X*11
A-1	259.23	143 5	160.8	74.8	24	5.5	5	5.2	2.5	150	126.4	10
B-1	252.05	137 0	142.5	66.05	24	5.5	5.3	5.3	2.4	200	126.4	7
C-1	280.6	163 0	159.8	75.2	26	4	6.8	5	2.1	200	126.4	5
C-2	297.95	163 0	166.4	80	32	3.5	7	5.5	1.6	200	105.6	5
A-2	265.45	143 5	158	82	34.5	4	6.7	6.5	1.7	150	135.1	7
A-3	280.66	143 5	174.5	87.8	35	3.5	7.7	8	1.6	200	103.9	7
C-3	312.55	163 0	177.6	78.3	36	2.5	8.5	7.5	1.5	250	95.8	3
D-1	320.97	155 0	194.4	96.5	35.5	3.8	6.5	7.5	2	150	96.7	6
B-2	270.34	137 0	159.5	75.5	36	4.5	6.6	6.8	1.9	250	105.1	7
D-2	315.67	155 0	185	92	39	3.7	6.8	7.5	1.8	200	104.8	5
E-1	225.6	120 0	130.8	69.5	20	6	4.6	5.5	2.8	100	126.4	18
F	232.45	125 0	138	72	24	5	4.8	6	2.3	100	109.1	15
G	280.34	138 0	166.5	85.2	24.5	5.3	5	6.2	2.3	100	118.2	18
H	290.23	145 0	167.2	85	25.5	5.4	5.2	6.3	2.1	150	103.9	20
E-2	243.26	120 0	148.3	77.8	29	3.5	5.5	6.8	2	120	95.8	12
F-2	265.98	138 0	170	83	30	4.5	5.6	7.5	2	130	96.7	13
F-3	278.67	138 0	176.4	85.15	34	4	6	7.8	1.8	120	97.7	12
G-2	285.12	145 0	177.5	88.5	33	4	6	8.2	1.8	180	104.8	12
I-1	315.89	165 0	185.5	88	20	4.5	4.5	5	2.8	210	126.4	15
J-1	368.12	190 0	218	105	22.5	4.1	5	6.2	2.7	200	111	15
I-2	320.23	165 0	195.5	95	25	4	5.1	6.5	2.4	210	135.1	12
J-2	380.7	190 0	228.3	107.5	26	3	5.9	7.2	1.9	230	101.3	10
I-3	325.4	165 0	196	100	27.5	3.5	5.7	7.3	1.8	250	96.7	10
J-3	365.92	190 0	215.5	102	29	2	6.8	8	1.7	250	104.8	8
I-4	332.25	165 0	202.8	103.5	29.5	3	6.3	7.8	1.8	250	104.8	10

3.1　PLSR 预测模型

基于 PLSR 的预测原理，将原始数据输入数据分析软件 SIMCA-P+12.0，软件根据交叉有效性指标自动选择最佳成分提取数。拟合结果见图 1，提取 1～3 个成分时的解释能力如图 2。

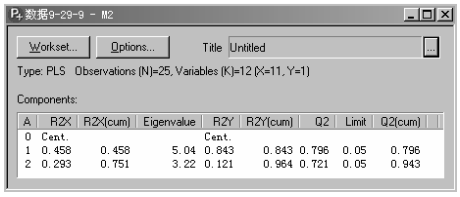

图 1　拟合结果

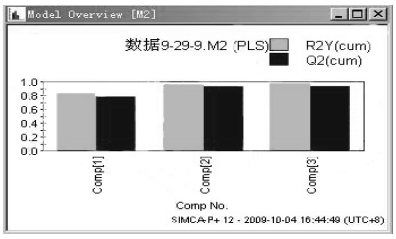

图 2　提取 1-3 个成分时的解释能力

提取第一成分 t_1 时，模型对 X、Y 的解释能力分别为 0.458 和 0.843，交叉有效性 Q^2=0.796＞0.097 5，继续提取成分；提取第二成分 t_2 时，模型对 X、Y 的解释能力分别上升 0.293 和 0.121，累计解释能力分别达到 0.751 和 0.964，交叉有效性 Q^2=0.147＞0.097 5，继续提取成分；提取第三成分 t_3 时，模型对 X 和 Y 的解释能力已上升不大，此时交叉有效性 Q^2=-0.314＜0.097 5，故停止提取成分。

综合分析结果，提取 2 个成分时，模型能够满足交叉有效性要求，实施 y 对潜变量 t_1、t_2 回归，得到标准化变量方程及标准化变量系数。最后，通过数据标准化的逆过程，可还原成 y 关于原始变量 $x_1, x_2, \cdots, x_k$ 的回归方程如下：

$$y = 49.326\ 4 + 0.059\ 324\ x_1 + 0.465\ 798\ x_2 + 0.918\ 102\ x_3 - 0.519\ 308\ x_4 - 5.147\ 774\ x_5 - 0.833\ 02\ x_6 + 1.110\ 209\ x_7 + 4.261\ 435\ x_8 + 0.127\ 383\ x_9 - 0.053\ 904\ x_{10} + 0.159\ 941\ x_{11} \quad (4)$$

式 (4) 体现出样本数据表中的所有自变量，PLSR 模型具有良好的变量解释能力。

3.2　模型分析

利用 PLSR 辅助分析技术，考证模型建立优劣度，为预测提供可靠信息。

3.2.1　变量投影重要性分析

按照式 (1) 计算各输入自变量对因变量的影响关系大小，即 VIP 值，结果如图 3。

图 3 中，x_1、x_2、x_3的 VIP 值都达到 1.5（显著大于 1），证明它们对修理成本的重要度影响明显；x_5、x_9的 VIP 值大于 1，可见对修理成本的也具有重要影响；x_4、x_6、x_7、x_8的 VIP 值在 0.5 到 1 之间，说明对修理成本影响比较重要。

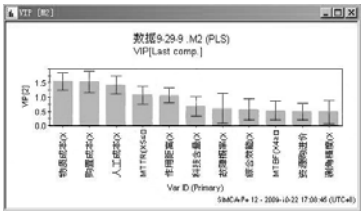

图 3 变量投影重要性分析

3.2.2 特异点分析

根据特异点分析原理，画出关于主成分t_1、t_2的散点图，即T^2椭圆图，如图 4。

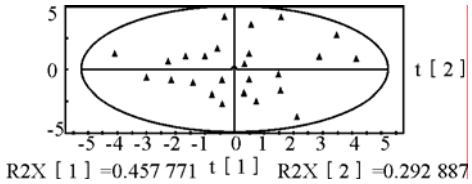

图 4 T^2椭圆图

所有样本点都分散在以t_1、t_2为主轴构成的椭圆内，证明所选样本符合建模要求，样本质量能够得到保证。

3.2.3 模型拟合质量分析

将大修成本历史数据 y，与预测模型数据 Y 相比较，大部分数据的绝对误差和相对误差都在 3% 以下（仅 1/3 数据突破 3%，但都在 5% 以下）。绘制模型预测值 Y 与实际值 y 的比较图如图 5。

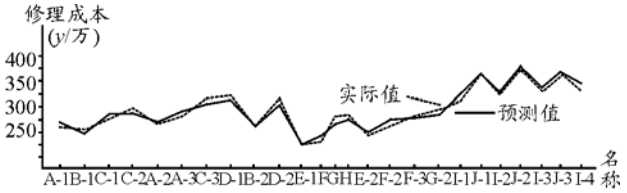

图 5 模型拟合值与实际值比较图

实线（预测值）与虚线（实际值）重合度较高，PLSR 预测模型拟合质量较好。

3.3 预测结果对比分析

收集我军此类现役 5 型试验装备最近一次大修成本相关数据如表 2，分别利用支持向量机、灰色理论、*PLSR* 对当前数据进行预测，结果如表 3。

表 2 5 型试验装备最近一次大修成本相关数据

	*X*1/万	*X*2/万	*X*3/万	*X*4/h	*X*5/h	*X*6	*X*7	*X*8/‰	*X*9/km	*X*10[5]	*X*11
B-3	137 0	159.8	78.2	39.5	4	7	8	1.8	250	108.3	3
D-3	155 0	175.7	86.8	41.5	2.5	7.5	8.5	1.3	220	94.7	3
E-4	120 0	143	75	33.5	2	6.1	9	1.6	150	108.3	10
H-3	145 0	175.8	90	34	2.5	6.6	9	1.5	200	94.7	10
J-4	190 0	230.7	115.6	35.5	1.5	7.5	9	1.6	280	94.7	5

表 3 3 种方法预测结果

名称	实际值	PLSR 预测		灰色理论预测		支持向量机预测	
		预测值	相对误差	预测值	相对误差	预测值	相对误差
B-3	280.76	273.01	2.76%	270.58	3.63%	270.82	3.54%
D-3	305.82	300.52	1.73%	298.16	2.50%	295.2	3.47%
E-4	249.55	254.89	-2.14%	240.9	3.47%	236.96	5.05%
H-3	295.45	302.19	-2.28%	285.64	3.32%	291.15	1.46%
J-4	387.68	391.40	-0.96%	394.78	1.83%	395.88	2.16%

由于原始数据出自近 20 年来同类十型试验装备大修成本实际发生值，在时间上不连续，因此灰色理论预测结果并不满意。支持向量机预测的精度受核函数本身影响较大，模型解释性差，误差变动幅度明显，预测仍不成熟。可见，采用 PLSR 预测修理成本获得了较高的预测精度，其预测效果更具优势。

4 基于 PLSR 的预测模型优化

4.1 优化思路

在 PLSR 预测建模中，提取 2 个成分时能解释较高比例(0.964) Y 的变化信息，但只能解释较低比例(0.751) X 的变化信息。只有包含更多主成分，预测效能才会不断改善；但模型本身限于交叉有效性检验($Q^2 \geq 0.097\ 5$)，提取过多主成分并不可行。究其原因，主要是原始输入自变量 X 中含有与因变量 Y 无关的变化信息[6]。

根据 PLSR 辅助分析原理，变量投影重要性可以测度每一自变量对因变量的贡献关系，定义的 VIP 值能够比较出这种关系大小。故利用变量投影重要性(VIP≥0.5 为标准)筛选自变量，以此作为模型的新输入变量，最大限度地排除变量中的不相关冗余因素。将层层筛选出的自变量再进行 PLSR 建模，就能有效提高预测精度。

4.2　优化验证

仍以表 1 数据建立 PLSR 预测模型，以表 2 数据检验优化后模型预测精度，并比较分析预测结果。

4.2.1　优化过程

基于当前预测结论，图 3 中 x_{10}、x_{11} 的 *VIP* 值明显低于 0.5，说明对修理成本 Y 的贡献关系不大。故剔除 x_{10}、x_{11}，将剩余原自变量组成新数据表，进行 PLSR 拟合。变量投影重要性分析结果表明除 x_4（$\mathrm{VIP}=0.440\,898$）不符合要求外，其余都在 0.5 以上。同理，剔除该变量，第三次拟合结果和变量投影重要性分析结果如图 6 和图 7。

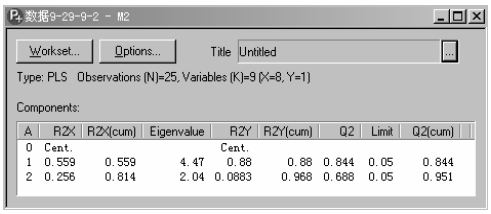

图 6　第三次拟合结果

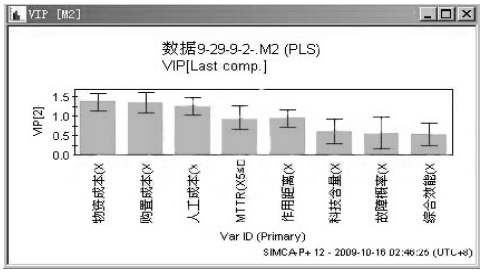

图 7　第三次变量投影重要性分析

此时，所有输入变量的 VIP 值都大于 0.5，说明选取的变量建模有效。虽然模型对 X 的累计解释能力下降了 0.003，但对 Y 解释能力上升了 0.004。

综上，模型经两次变量筛选优化后，对 X、Y 的累计解释能力分别上升 0.063 和 0.004，预测函数式为：

$$y=21.843\,3+0.062\,686x_1+0.486\,871x_2+0.950\,280x_3-4.429\,697x_5-1.815\,086x_6+0.200\,081x_7+7.440\,993x_8+0.122\,15x_9 \quad (5)$$

4.2.2　模型拟合质量分析

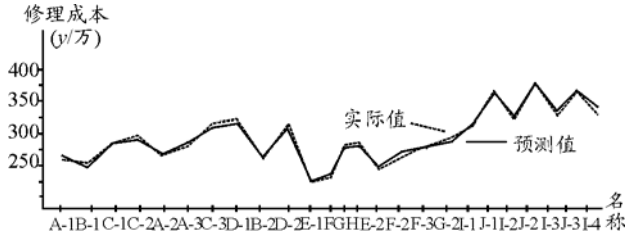

图 8　优化后模型的拟合值与实际值比较图

对比历史与预测数据后发现，绝大部分的相对误差在 3% 以下（极少数样本突破 3%，但都在 5% 以下），模型预测精度优化明显。绘制模型拟合值 Y 与实际值 y 的比较图如图 8。

图中虚实连线的重合程度比优化前更加提高。

4.2.3　模型预测结果分析

以表 2 为检验数据，优化前后模型预测结果如表 4。

表 4　优化前后模型预测结果对比

名称	实际值	优化前		优化后	
		预测值	相对误差	预测值	相对误差
B-3	280.76	273.01	2.76%	274.95	2.07%
D-3	305.82	300.52	1.73%	300.60	1.71%
E-4	249.55	254.89	-2.14%	250.06	-0.20%
H-3	295.45	302.19	-2.28%	298.19	-0.93%
J-4	387.68	391.40	-0.96%	390.77	-0.80%

除 B 型装备第三次大修成本预测值的相对误差略大外，其余误差均小于优化前。试验装备修理成本预测由于自身特点，不应强调某型装备某次大修成本误差，而应从总体上综合考虑预测误差水平，因此模型优化后的预测效果更加优异。

5　结论

该方法不仅能在多变量间存在严重多重相关性的情况下建立模型，而且能够有效筛选与因变量关系不大的自变量，简化输入样本集。此外，利用变量投影重要性分析优化预测模型也能较大程度地提高预测精度。因此，基于 PLSR 的试验准备修理成本预测具有其他方法不可比拟的优势，能为我军成本预测领域相关工作提供良好的方法支持。

参考文献：

[1] 王惠文. 偏最小二乘回归方法及其运用[M]. 北京: 国防工业出版社, 1999: 14-19.

[2] 李寿安, 张恒喜, 童中翔, 等. 偏最小二乘回归在军用飞机价格预测中的应用[J]. 航空学报, 2006(5): 601-602.

[3] 朱洵, 荣起国. 基于偏最小二乘回归的基因网络数学建模[J]. 系统仿真学报, 2009(2): 1149-1150.

[4] 王惠文, 吴载斌, 孟洁. 偏最小二乘回归的线性与非线性方法[M]. 北京: 国防工业出版社, 2006: 124-125.

[5] 中国统计年鉴: 各类价格指数[M]. 北京: 中华人民共和国国家统计局, 2008.

[6] 李波, 顾冲时, 李智录, 等. OSC-PLS 法在大坝安全监控模型中的应用[J]. 水利水电科技进展, 2008(8): 5-6.

[7] 周大伟, 何宝民, 冯楠. 基于预知维修技术的装备维修管理[J]. 四川兵工学报, 2009(3): 105-106.