Partial least squares regression for predicting economic loss of vegetables caused by acid rain

WANG Ju 1,†, MENG He 1, DONG De-ming 1, LI Wei 2, FANG Chun-sheng 1,‡
1 College of Environment and Resources, Jilin University, Changchun 130012, P.R. China
2 College of Environment, Beijing Normal University, Beijing 100875, P.R. China

Received 10 October 2008; received in revised form 12 February 2009

Abstract: To predict the economic loss of crops caused by acid rain, we used partial least squares (PLS) regression to build a model of single dependent variable – the economic loss calculated with the decrease in yield related to the pH value and levels of Ca²⁺, NH₄⁺, Na⁺, K⁺, Mg²⁺, SO₄²⁻, NO₃⁻, and Cl⁻ in acid rain. We selected vegetables which were sensitive to acid rain as the sample crops, and collected 12 groups of data, of which 8 groups were used for modeling and 4 groups for testing. Using the cross validation method to evaluate the performance of this prediction model indicates that the optimum number of principal components was 3, determined by the minimum of prediction residual error sum of squares, and the prediction error of the regression equation ranges from −2.25% to 4.32%. The model predicted that the economic loss of vegetables from acid rain is negatively correlated to pH and the concentrations of NH₄⁺, SO₄²⁻, NO₃⁻, and Cl⁻ in the rain, and positively correlated to the concentrations of Ca²⁺, Na⁺, K⁺ and Mg²⁺. The precision of the model may be improved if the non-linearity of original data is addressed.

Keywords: acid rain; partial least-squares regression; economic loss; dose-response model

1 Introduction

Acid rain refers to the precipitation of pH value below 5.6, in thunder storm, fog, hail and other forms. Acid rain is also called acid deposition, and includes wet deposition and dry deposition, caused by emissions of sulphur dioxide and nitrogen oxides to the atmosphere.

Acid deposition seriously damages environment, and has become one of the ten major problems threatening the global environment. It first became an environmental problem in European industrialized countries and the United States, but now has extended to developing countries. Acid rain emerged as an important environmental problem in P. R. China in the late 1970s [1]. During 1980s, acid rain there occurred mainly in several provinces of south China including Sichuan, Guizhou, Guangdong, Guangxi and all the eastern regions of the country [2-3]. In 2000, the annual average volume-weighted pH was 3.53 and the acid rain frequency (i.e. pH<5.6) was 90% [4]. Xie [5] reported that almost 40% of China is subject to average annual rainwater conditions with a pH value less than 5.6. China has become the third largest acid rain-prone region in the world after northern Europe and North America [6].

The chemical components of acid deposition are sulphur oxides (about 60%), nitrogen oxides (between 21% and 23%), chloride (between 10% and 12%), carbon dioxide, ammonia and organic acids. Human activities have significant impacts on the formation of acid deposition. Chemical analysis of acid deposition...
Economic loss from acid rain

2 Methods

2.1 Partial least-squares regression

The partial least-squares (PLS) regression method was first proposed by Word et al. [13]. Later, PLS was developed and applied in statistical fields to solve problems difficult to handle by multiple linear regressions. This method can be applied to analyzing not only single-variable regression, but also multi-variable regression [14-18].

PLS can efficiently screen out integrated and variable information that is the most explainable in the system, and avoid the interference caused by overlapping correlations and useless information. Thereby, it overcomes the negative effects of overlapping correlations of variations in system modeling. PLS provides a regression modeling method between multiple dependent and independent variables, especially when they are closely correlated. The use of PLS modeling usually results in more reliable conclusions than other modeling approaches. PLS effectively solves the problems of overlapping correlations among the variables and incorporates a variety of multivariate statistical analyzing methods, and hence is especially suitable when the sample size is less than the number of variables [14].

In this work, we applied PLS regression to a single dependent-variable case to build a model for estimating the economic loss of crops from acid deposition. The investigated influencing factors were pH, Ca$^{2+}$, NH$_4$$^+$, Na$^+$, K$^+$, Mg$^{2+}$, SO$_4^{2-}$, NO$_3^-$, and Cl$^-$, denoted by $x_j$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, $x_7$, $x_8$, $x_9$, respectively. Economic loss from acid deposition was taken as the objective indicator denoted by $y$.

2.2 Cross validation

We applied the cross-validation method to evaluating the reliability of the established model, using Visual Basic 6.0 in programming to predict estimated results and test the validity of the established fitting equation.

The cross-validation method is a procedure for accurately assessing the forecast performance of a model and determining the optimum number of principal components. The main index in the assessment is prediction residual error sum of squares (PRESS) that is defined as follows:

$$\text{PRESS}_h = \sum_{i=1}^{n} \left( y_i - \hat{y}_{(i-h)} \right)^2,$$

where $y_i$ is the $i$-th original value, $\hat{y}_{(i-h)}$ is the $i$-th fitting value of modeling set up with sample $n-i$. PRESS$_h$ indicates the quality of phase $h$ prediction equation. When the regression equation is not stable and at considerable variance, the regression equation is sensitive to the change of samples. This disturbance increases the value of PRESS$_h$, which is as an index for measuring the reliability of the model. A model has the best prediction ability when the value of PRESS$_h$ reaches minimum, wherein the value of $h$ is the optimum number of principal components. Written as $h^*$, this optimum number is given by

$$\text{PRESS} \left( h^* \right) = \min_{1 \leq h \leq r} \text{PRESS}_h, \quad r = \text{rank}(X^T X),$$

where $r$ is the regression coefficient; and $X$ is the
matrix of independent variable.

2.3 Data collection

We used vegetables as the sample crops sensitive to acid deposition, with 12 sets of data from Refs. [19-22]. The data were divided into two groups: eight sets for analysis modeling, and the rest four sets for model testing.

3 Results and discussion

3.1 Multiple correlations

The correlation coefficients between independent variables and the dependent variable are shown in Table 1. Strong multiple correlations between the variables were observed, with a particular strong negative correlation (correlation coefficient is $-0.936$) between pH and the economic loss from acid deposition. The lower the pH of acid deposition was, the greater the economic loss from acid deposition; this agrees well with the actual situation.

3.2 Optimum number of principal components

The results of PRESS and cumulative PRESS are listed in Table 2. PRESS$_h$ reached its minimum when the number $h$ of principal components was 3, indicating 3 principal components were adequate to generalize the information of the original samples.

3.3 Modeling regression equation

Given the optimum principal component number of 3, the coefficients of each of the component variables are shown in Table 3.

According to partial least-squares regression algorithm, the prediction model is derived as follows.

The origin variable regression equation was

$$y = 7.3050 - 0.7910x_1 + 0.0240x_2 - 0.7747x_3 + 0.0194x_4 + 0.2770x_5 + 0.1645x_6 - 0.0054x_7 - 0.0501x_8 - 0.1742x_9. \quad (3)$$

The standardization variable regression equation was

$$\hat{y}^* = -0.6344x_1^* + 0.0090x_2^* - 0.4540x_3^* + 0.0720x_4^* + 0.1330x_5^* + 0.2731x_6^* - 0.0783x_7^* - 0.1280x_8^* - 0.0703x_9^*, \quad (4)$$

where the asterisked symbols represent the standardized variables of corresponding original ones in Eq. (3).

<table>
<thead>
<tr>
<th>$R(.)$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
<th>$x_8$</th>
<th>$x_9$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>1.000</td>
<td>-0.213</td>
<td>0.330</td>
<td>-0.316</td>
<td>0.036</td>
<td>-0.385</td>
<td>0.026</td>
<td>0.018</td>
<td>0.266</td>
<td>-0.936</td>
</tr>
<tr>
<td>$x_2$</td>
<td>1.000</td>
<td>0.783</td>
<td>0.991</td>
<td>0.759</td>
<td>0.846</td>
<td>0.933</td>
<td>0.719</td>
<td>0.803</td>
<td>0.149</td>
<td></td>
</tr>
<tr>
<td>$x_3$</td>
<td>1.000</td>
<td>0.702</td>
<td>0.598</td>
<td>0.390</td>
<td>0.788</td>
<td>0.641</td>
<td>0.754</td>
<td>-0.457</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_4$</td>
<td>1.000</td>
<td>0.752</td>
<td>0.887</td>
<td>0.910</td>
<td>0.672</td>
<td>0.752</td>
<td>0.255</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_5$</td>
<td>1.000</td>
<td>0.651</td>
<td>0.726</td>
<td>0.686</td>
<td>0.758</td>
<td>0.051</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_6$</td>
<td>1.000</td>
<td>0.846</td>
<td>0.571</td>
<td>0.730</td>
<td>0.441</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_7$</td>
<td>1.000</td>
<td>0.654</td>
<td>0.907</td>
<td>-0.023</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_8$</td>
<td>1.000</td>
<td>0.816</td>
<td>0.005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_9$</td>
<td>1.000</td>
<td>-0.185</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y$</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: $x_1$ is pH; $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, $x_7$, $x_8$ and $x_9$ are respectively the concentrations of $Ca^{2+}$, $NH_4^+$, $Na^+$, $K^+$, $Mg^{2+}$, $SO_4^{2-}$, $NO_3^-$, and $Cl^-$; $y$ is economic loss from acid deposition.
### Table 2  Prediction residual error sum of squares (PRESS) and cumulative PRESS of different components \( t_1 \) to \( t_7 \) extracted form dependent variable

<table>
<thead>
<tr>
<th>Component</th>
<th>PRESS</th>
<th>Cumulative PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_1 )</td>
<td>5.599</td>
<td>5.599</td>
</tr>
<tr>
<td>( t_2 )</td>
<td>1.566</td>
<td>7.165</td>
</tr>
<tr>
<td>( t_3 )</td>
<td>1.477</td>
<td>8.642</td>
</tr>
<tr>
<td>( t_4 )</td>
<td>2.848</td>
<td>11.490</td>
</tr>
<tr>
<td>( t_5 )</td>
<td>4.475</td>
<td>15.965</td>
</tr>
<tr>
<td>( t_6 )</td>
<td>7.149</td>
<td>23.114</td>
</tr>
<tr>
<td>( t_7 )</td>
<td>21.976</td>
<td>45.090</td>
</tr>
</tbody>
</table>

### 3.4 Model evaluation

#### 3.4.1 Precision of regression model

Table 4 shows the precision analysis of regression model.

Precision analysis shows that components \( t_1 \), \( t_2 \) and \( t_3 \) explain 91.99% of the variability in the original dependent variable system, and they have a good representation for the original variable system. Component \( t_1 \) explains almost all of the variability of the dependent variable system; whereas \( t_2 \) accounts for a major part of the variability of the original independent variable system.

#### 3.4.2 Correlation analysis

The correlativity was expressed by the dependence of a component on \( u_1 \), the ratio of the standardized vector \( F_0 \) of economic loss \( y \) to the regression coefficient \( r_1 \) of the component; \( u_1 = F_0/r_1 \). The scatter charts (Fig. 1) of components \( t_1 \), \( t_2 \) and \( t_3 \) versus \( u_1 \) show an obvious linear relationship between \( t_1 \) and \( u_1 \) (Fig. 1a), which supports the result of the sum of squares of correlation coefficient between the dependent variable \( y \) with extracted components, i.e. \( \text{RdY}(t_1) = 91.01\% \). The linear correlation between \( t_2 \) and \( u_1 \) is weaker (Fig. 1b); and that between \( t_3 \) and \( u_1 \) is the weakest (Fig. 1c).

#### 3.4.3 Variable importance in projection

The variable importance in projection indicates the importance of an independent variable \( x_i \) to the dependent variable \( y \) [14].

The results of variable importance in projection suggest that the dependent variable \( y \) can be clearly explained by the independent variable \( x_1 \) (Fig. 2).

### 3.5 Model fitting and testing

According to the model of partial least-squares regression, the group of eight sets of sample data were used for model fitting analysis. The results show that the model fitting error ranged from \(-8.08\%\) to \(9.68\%\), with a maximum relative error of \(9.68\%\) and a minimum relative error of \(0.07\%\). The regression equation fitted well and the model was feasible.

### Table 3  Component coefficients \( R \) and regression coefficient \( r \)

<table>
<thead>
<tr>
<th>Component</th>
<th>( x_1^* )</th>
<th>( x_2^* )</th>
<th>( x_3^* )</th>
<th>( x_4^* )</th>
<th>( x_5^* )</th>
<th>( x_6^* )</th>
<th>( x_7^* )</th>
<th>( x_8^* )</th>
<th>( x_9^* )</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_1 )</td>
<td>-0.7897</td>
<td>0.1255</td>
<td>-0.3857</td>
<td>-0.2155</td>
<td>0.0429</td>
<td>0.3727</td>
<td>-0.0196</td>
<td>0.0040</td>
<td>-0.1559</td>
<td>0.7686</td>
</tr>
<tr>
<td>( t_2 )</td>
<td>-0.2981</td>
<td>-0.3986</td>
<td>-0.6029</td>
<td>-0.3646</td>
<td>-0.1622</td>
<td>-0.4122</td>
<td>-0.1445</td>
<td>-0.3233</td>
<td>0.0692</td>
<td></td>
</tr>
<tr>
<td>( t_3 )</td>
<td>-0.0299</td>
<td>-0.2636</td>
<td>-0.6099</td>
<td>-0.3013</td>
<td>0.4897</td>
<td>-0.0047</td>
<td>-0.1499</td>
<td>0.5624</td>
<td>0.3165</td>
<td>0.2271</td>
</tr>
</tbody>
</table>

Notes: The asterisked symbols \( x_1^* \), \( x_2^* \), \( x_3^* \), \( x_4^* \), \( x_5^* \), \( x_6^* \), \( x_7^* \), \( x_8^* \) and \( x_9^* \) are the standardized independent variables, namely, standardized \( \text{pH} \), and standardized concentrations of \( \text{Ca}^{2+} \), \( \text{NH}_4^+ \), \( \text{Na}^+ \), \( \text{K}^+ \), \( \text{Mg}^{2+} \), \( \text{SO}_4^{2-} \), \( \text{NO}_3^- \), and \( \text{Cl}^- \), respectively; \( t_1 \), \( t_2 \) and \( t_3 \) are extracted components.
Table 4  Precision analysis of regression model: herein RdX is the sum of squares of correlation coefficient between independent variables with extracted components $t_1$, $t_2$ and $t_3$; RdY is that between the dependent variable $y$ with $t_1$, $t_2$ and $t_3$; PRESS$_h$ is the prediction residual error sum of squares

<table>
<thead>
<tr>
<th>Component</th>
<th>RdX</th>
<th>Cumulative RdX</th>
<th>RdY</th>
<th>Cumulative RdY</th>
<th>PRESS$_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>0.189</td>
<td>0.189</td>
<td>0.910</td>
<td>0.910</td>
<td>5.599</td>
</tr>
<tr>
<td>$t_2$</td>
<td>0.672</td>
<td>0.862</td>
<td>0.026</td>
<td>0.936</td>
<td>1.566</td>
</tr>
<tr>
<td>$t_3$</td>
<td>0.058</td>
<td>0.919</td>
<td>0.026</td>
<td>0.962</td>
<td>1.477</td>
</tr>
</tbody>
</table>

Fig. 1  Scatter charts of principal components (a) $t_1$, (b) $t_2$ and (c) $t_3$ versus $u_1$ which is defined as $u_1=F_0/r_1$, the ratio of the standardized vector $F_0$ of economic loss $y$ to the regression coefficient $r_1$

Fig. 2  Variable importance in projection (VIP) of different independent variables $x_i$ ($i=1, 2,...,9$) to the dependent variable $y$, where $x_1$ is pH, $x_2$, $x_5$, $x_8$, $x_4$, $x_9$, $x_5$, $x_6$ and $x_9$ are respectively the concentrations of Ca$^{2+}$, NH$_4^+$, Na$^+$, K$^+$, Mg$^{2+}$, SO$_4^{2-}$, NO$_3^-$, and Cl$^-$; $y$ is the economic loss from acid deposition.

Testing the quality of the regression model by the remaining 4 sets of sample data indicated that the prediction error of the regression equation ranged from $-2.25\%$ to $4.32\%$, with a maximum relative error of $4.32\%$ and a minimum relative error of $-2.25\%$. The regression model predicted reliable results.

4 Conclusions

We established an approach to predict the economic loss of vegetables from acid deposition with respect to the ion concentrations in the deposition and used PLS regression to evaluate the reliability of the prediction model. The regression model predicted a direct negative correlation between pH and the economic loss of vegetables from acid deposition. The economic loss was also predicted to be positively correlated to the concentrations of Ca$^{2+}$, Na$^+$, K$^+$ and Mg$^{2+}$, and negatively correlated to the concentrations of NH$_4^+$, SO$_4^{2-}$, NO$_3^-$, and Cl$^-$. The model can be a useful tool to help control acid deposition. To extend the application of this model to other regions, more sample data are
needed for modelling.

Although the method of partial least-squares regression overcomes the problem of the multiple correlation between variables, the precision of this regression model is insufficient because of the complicated mechanism of crop loss caused by acid deposition which is affected by many factors, and of the PLS regression algorithm which neglects the non-linear features of original data. Thus, this regression model requires further improvement.

References


Economic loss from acid rain


Edited by LUO Min