ErosPredict
– A Program for Predicting Soil Erosion

Søren Højsgaard, sorenh@agrsci.dk Helle H. Rasmussen and Jørgen Djurhuus,
Biometry Research Unit, Department of Agricultural Systems, Danish Institute of
Agricultural Sciences, Research Center Foulum, DK-8830 Tjele, Denmark.

A probabilistic model for predicting the risk (and amount) of water erosion is estab-
lished on the basis of data collected in Denmark over a 7 year period. The model is imple-
mented in the program ErosPredict. For parts of the data not all variables to be used
in the models were observed. Using Markov Chain Monte Carlo methods it is shown that
by utilizing also the incomplete data, improved parameter estimates (with respect to the
predictive ability of the model) can be obtained.

Keywords: Bayesian network, Incomplete data, Markov Chain Monte Carlo

1. INTRODUCTION

The aim of the project is to establish an expert system for predicting soil erosion caused by
surface drainage. The expert system is able to be used to spot areas where erosion is a potential
problem for the aquatic environment and to recommend measures for reducing erosion. The
loss of soil to streams and lakes is a problem due to enhanced eutrophication, caused mainly by
phosphorus in the eroded soil.

Data were collected over a 7-year period in Denmark. During this period erosion by wa-
ter was measured on a total of 189 slopes at 20 different locations. A slope was defined as a
field with uniform cultivation, e.g. soil tillage and crop. Soil types were mostly loamy sand or
sandy loam, while a smaller part was either sand, sandy clay loam or sandy silt loam. Mea-
asurements of erosion were performed in late autumn and in spring. The total number of obser-
vations was 1041, of which 213 had erosion (about 20%). The observations with erosion were
highly skewed with a 75% percentile of 1.49 m³ha⁻¹. The expert system contains a number
of measured or recorded variables, e.g. cultivation, soil texture, water impermeable layer, soil
tillage direction, length-slope factors, soil surface roughness and a selected number of climate
variables, e.g. accumulated precipitation, days with precipitation above 20 and 30 mm and precipitation and melting of snow on frozen soil. The measured response is the furrow volume per area ($m^3/ha$). The data consists of two types of explanatory variables: Type I are variables which are easy to measure while type II are those difficult to obtain. For the data in the study, information about the second type of variables is not collected for about 60% of the cases.

The statistical model underlying the expert system is established in two steps. 1) First, on the basis of the complete data, the structure of the prediction model is determined. In this step, parameter estimates to be used in connection with prediction are also obtained. 2) Secondly, the incomplete data are used to achieve improved parameter estimates using Markov Chain Monte Carlo methods.

2. Data

The erosion is measured as furrow volume per area ($m^3/ha$). The type I explanatory variables are presented in the upper part of Table 1. The measurements of the precipitation in Table 1 are since last cultivation (max. one year). The type II variables are presented in the lower part of Table 1. These variables are complicated to measure. Due to a change in the experimental plan, these variables were not measured in about 60% of the cases.

In the following we let $D$ denote the entire data set, $DC$ the set of complete cases and $DI$ the incomplete cases, i.e. cases where the type II explanatory variables are missing. The individual measurements are indexed by $i$ in what follows. Let $Cult(i)$, $Asp(i)$ and $Wil(i)$ denote the cultivation, aspect and type of water impermeable layer for the $i$'th observation and let $YR(i)$ the erosion–year and the region of the $i$'th observation. (See Table I for the definition of an erosion–year.)

![Figure 1: Erosion data follow a right skewed distribution with point mass in zero. When erosion is present, the amount of erosion can be described by a log–normal distribution.](image)

3. Identifying the Models from the Complete Data

Figure 1 illustrates that most frequently there is no erosion at all and that when erosion is present, it follows a right skewed distribution. The erosion is therefore modelled by a right
Table 1: The Type I explanatory variables are those easy to obtain and the Type II explanatory variables are those difficult to obtain and which are only recorded in about 40% of the cases.

<table>
<thead>
<tr>
<th>Type I explanatory variables – covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{d_{20mm}}$ Days with precipitation greater than 20 mm</td>
</tr>
<tr>
<td>$x_{snowmelt}$ Accumulated precipitation and melting of snow on frozen soil (mm)</td>
</tr>
<tr>
<td>$x_{ls_{99}}$ LS 99% quantile</td>
</tr>
<tr>
<td>$x_{ln_{cl_{si}}}$ Ln of sum of clay and silt (2-20 μm) at up slope; ln(%)</td>
</tr>
<tr>
<td>$x_{ln_{prec8}}$ Ln of accumulated precipitation for days with precipitation &gt; 8 mm; ln(mm)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type I explanatory variables – factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{cult}$ System of cultivation – 1: grain; 2: Christmas trees; 3: ploughed; 4: stubble harrowed</td>
</tr>
<tr>
<td>$x_{asp}$ Aspect – 1: northwest, north, northeast, east; 2: southeast, south, southwest, west</td>
</tr>
<tr>
<td>$x_{wil}$ Water impermeable layer; 1: no/some; 2: yes</td>
</tr>
<tr>
<td>$x_{region}$ Region where the slope unit belongs to</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type II explanatory variables – covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{mud_{1}}$ Soil surface roughness as ln of MUD in tillage direction; ln(mm)</td>
</tr>
<tr>
<td>$x_{mud_{2}}$ Soil surface roughness as ln of MUD perpendicular to tillage direction; ln(mm)</td>
</tr>
</tbody>
</table>

Soil physical knowledge dictate that the type II variables could be important predictor variables of erosion. The model is therefore initially based only on the complete data DC, i.e. the data where the mud variables are registered. The models were selected on the basis of statistical significance as well as soil physical background knowledge. Incorporation of the incomplete data DI is discussed in Section 5.

The Basic Idea To model the erosion risk, a random variable $Z$ is defined as $Z = 1$ if erosion takes place and $Z = 0$ otherwise where the probability of erosion $Pr(Z = 1) = p(x)$ is a function of $x$. Given that there is erosion, i.e. that $Z = 1$, the amount of erosion is modelled by a log–normal distribution:

$$Pr(Y|Z = 1, x) \sim log \ N(\mu^y(x), \sigma^2_y)$$

where $\mu^y(x)$ is a function of $x$. If there is not erosion, i.e. if $Z = 0$, then the amount of erosion is certainly zero, i.e. $Y = 0$ with probability 1; i.e. $Pr(Y = 0|Z = 0) = 1$ and
\[ Pr(Y = y \neq 0 | Z = 0) = 0. \] A graphical representation of the models found below is given in Figure 2.

![Figure 2: Illustration of the dependencies in the model. In the Figure, \( \text{logit}_p \) is \( \logit p_i \) and \( \mu \) is \( \mu_i \).](image)

Introduction of the variable \( Z \) can be regarded as a trick to circumvent the point mass in zero. After all, interest is in the marginal distribution of \( Y \) given the covariates \( x \). This distribution is

\[
Pr(Y|Z = 0, x) = Pr(Z = 0|x) + Pr(Y|Z = 1, x)Pr(Z = 1|x).
\]

Using the specific construction of the model above we find

\[
Pr(Y = 0|x) = Pr(Z = 0|x) = 1 - p(x)
\]

\[
Pr(Y = y \neq 0|x) = Pr(Y = y \neq 0|Z = 1, x)Pr(Z = 1|x) = Pr(y|Z = 1, x)p(x).
\]

**Modelling the Erosion Risk** The probability of erosion taking place is modelled by a logistic regression (which is noted to depend on the type II variables \( x_{\text{mud}_1} \) and \( x_{\text{mud}_2} \)):

\[
\text{logit}(p_i) = \delta_{\text{Cult}(i)}^x + \zeta_{\text{arp}(i)}^x + \eta_{\text{wil}(i)}^x + \beta_1 x_{\ln\text{cl}_{si},i} + \beta_2 x_{\text{snowmelt},i} + \\
\beta_3 x_{d_{20mm},i} + \beta_4 x_{ls_{99},i} + \beta_5 x_{\text{mud}_1,i} + \beta_6 x_{\text{mud}_2,i}.
\] (2)

**Modelling the amount of Erosion** Given that there is erosion, the amount of erosion is modelled by a log–normal distribution (which is noted not to depend on the \( x_{\text{mud}} \) variables):

\[
\mu_i^y = \delta_{\text{Cult}(i)}^y + \eta_{\text{wil}(i)}^y + \gamma_1 x_{\ln\text{cl}_{si},i} + \gamma_2 x_{\ln\text{prec8},i} + \gamma_3 x_{ls_{\text{mean}},i} + u_{\text{yr}(i)}.
\] (3)
Modelling the Mean Upslope Depression (MUD)  

The $x_{mud\_1,i}$ and $x_{mud\_2,i}$ variables are modelled by

$$x_{mud\_j,i} \sim N(\mu_{m_j}, \sigma_{m_j}^2),$$

where

$$\mu_{m_j} = \alpha_{1}cult(j) + \alpha_{2}cult(i) \times ln_{cl\_si,i} + \alpha_{3} \times d_{20mm,i} \times B(i) + u_{field(i)}, \tag{4}$$

and $u_{field(i)} \sim N(0, \sigma_{field}^2)$. Here $u_{field(i)}$ describes the field-to-field variation and $B(i)$ is 1 if the system has been cultivated in the fall (winter crop, ploughed or stubbled) and 0 otherwise.

4. Combining the Erosion Risk and Erosion Amount

The model in (1) is a model for the amount of erosion given that erosion has occurred. However, erosion is (fortunately) a rare event. The model can therefore not be taken to predict the amount of erosion one would predict on average under a given set of circumstances. To do so, one needs to combine the expected level from (1) with the probability of erosion occurring. Let for a given set of type I variables

$$Pr(Z = 1) = p, \quad E(Y|Z) = \begin{cases} \eta & Z = 1 \\ 0 & Z = 0 \end{cases}, \quad \text{and} \quad \text{Var}(Y|Z) = \begin{cases} \omega^2 & Z = 1 \\ 0 & Z = 0 \end{cases}.$$  

Then, using text book formulas for conditional means and variances it can be shown that

$$E(Y) = p\eta \quad \text{and} \quad \text{Var}(Y) = p[\omega^2 + \eta^2(1 - p)] = p\omega^2 + \frac{1 - p}{p}E(Y)^2.$$  

However, since the distribution is highly skewed, it is difficult to attribute meaning to the variance $\text{Var}(Y)$. For example, $E(Y) \pm 2\sqrt{\text{Var}(Y)}$ can not be regarded as a 95% prediction interval. Yet, the median and a 95% prediction interval is easy to find by simulation (and compare with the expected amount of erosion) as shown in Figure 3.

5. Improving Parameter Estimates using the Incomplete Data

Only the complete data $DC$ (about 40% of the cases) were used in establishing the models in Section 4. However, it is appealing to try to utilize all available information, i.e. the incomplete data $DI$ as well.

One approach in this connection is to regard the structure of the models, i.e. the choice
Figure 3: Predictive distribution for three different scenarios (i.e. three different values of the Type I variables). The 2.5%, 50% and 97.5% quantiles and the mean for the three scenarios are (from left to right): (0.0, 0.0, 9.7; 1.5), (0.0, 1.6, 10.9; 2.9) and (0.0, 5.2, 13.6; 5.5).

of explanatory variables in (2), (3) and (4) as being fixed. The parameters of the models, on the other hand, can be updated using the information in DI as well. This can be achieved by putting the problem into a Bayesian framework, see e.g. Gilks et al. [1996], by means of the BUGS program, Spiegelhalter et al. [1996].

5.1. Application of the Bayesian Approach

The parameters of the models in Section 3 (symbolically written as \( \theta \)) were estimated on the basis of the complete data DC and we let \( \hat{\theta}_C \) denote the estimate of the parameters based on DC. The BUGS program was used in updating the parameters for the erosion models as follows:

General theory gives that each component \( \hat{\theta}_{kc} \) of the estimator \( \hat{\theta}_C \) is approximately normal with mean \( \theta_k \) and variance \( \sigma_k^2 \) which can be estimated from data. This suggests a prior distribution for \( \theta_k \) in a Bayesian approach, namely \( \pi(\theta_k) \sim N(\hat{\theta}_{kc}, R\sigma_k^2) \). If \( R \approx 0 \) then \( \pi(\theta_k) \) will be very concentrated around \( \hat{\theta}_{kc} \). In this case the data will have only relatively little influence on the updated estimate. On the other hand if \( R \) is large, then data will have a large influence on the updated estimates. We have tried different values of \( R \) and have found that updated values do not change much, which is very comforting. Therefore we have, somewhat arbitrarily, chosen \( R = 2 \). After updating the parameters, one ends up not with a single value for \( \theta_k \) but with the posterior distribution. For use in the prediction model we have take the mode of the distribution, i.e. the value of \( \theta \) for which the distribution has its maximum.

5.2. Evaluating the Predictive Ability

To evaluate the predictive abilities of the models, we have done as follows: The incomplete data DI randomly split into two data sets DI1 and DI2. The model parameters \( \theta \) were updated using only DI1 giving an updated estimate \( \hat{\theta}_{11} \). Recall that erosion is only present in about 20%
of the cases. Therefore one would expect that the models to perform quite well in predicting cases when the erosion risk is low. However, in practice interest is in being able to predict cases where the erosion risk is high. Therefore we took a subset $\text{DI11}$ of $\text{DI1}$ consisting of all cases with erosion in $\text{DI1}$ and a random sample of the same size of cases without erosion. Hence $\text{DI11}$ is balanced with respect to occurrence of erosion. Intuitively, it is therefore expected that when updating the parameters on the basis of these data, the cases with erosion would be given a higher weight thus making the model better in predicting cases with erosion. Updating the parameters as described above yielded and estimate $\hat{\theta}_{i11}$. The models were then used to predict the erosion in $D_{m2}$ using $\hat{\theta}_C$, $\hat{\theta}_1$, and $\hat{\theta}_{i11}$. Measures of predictive ability are discussed in Section 5.3.

5.3. Measures of Predictive Ability

Let $\text{DI21}$ and $\text{DI20}$ be those observations in $\text{DI2}$ with erosion respectively without erosion, where there are $N_1$ cases in $\text{DI21}$ and $N_0$ cases in $\text{DI20}$. The scores below were used to evaluate the prediction of $p_i$ and $\mu_i$. Common to all scores is that a low value indicates a good predictive ability. The scores used were Brier score, log–score and mean squared error of prediction:

\[
\begin{align*}
Brier_1 &= \frac{1}{N_1} \sum_{i \in \text{DI21}} (1 - p_i)^2 \\
Brier_0 &= \frac{1}{N_0} \sum_{i \in \text{DI20}} p_i^2 \\
\text{LogScore}_1 &= -\frac{1}{N_1} \sum_{i \in \text{DI21}} \log p_i \\
\text{LogScore}_0 &= -\frac{1}{N_0} \sum_{i \in \text{DI20}} \log (1 - p_i) \\
MSEP_1 &= \frac{1}{N_1} \sum_{i \in \text{DI21}} (y_k - p_i \mu_i^y)^2 \\
MSEP_0 &= \frac{1}{N_0} \sum_{i \in \text{DI20}} (1 - p_i \mu_i^y)^2
\end{align*}
\]

6. Prediction Results

Table 2 below shows the prediction results for the data set $\text{DI2}$ for the three parameter estimates. Since the score values do not have any intrinsic meaning, they have been made calculated relative to the score for models based on $\hat{\theta}_C$. The general conclusion is that when updating the parameters using $\text{DI1}$ the model get slightly better in predicting cases without erosion, while performance in predicting cases with erosion is marginally worse. When updating the parameter estimates from the balanced data $\text{DI11}$ the model become markedly better in capturing cases with erosion, but this is at the expense of being considerably worse in predicting
cases without erosion.

Table 2: Prediction results for three different sets of model parameters calculated relative to $\hat{\theta}_C$.

<table>
<thead>
<tr>
<th></th>
<th>$Brier_1$</th>
<th>$Brier_0$</th>
<th>LogScore$_1$</th>
<th>LogScore$_0$</th>
<th>MSPE$_1$</th>
<th>MSPE$_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_C$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$\theta_{M_1}$</td>
<td>1.09</td>
<td>0.43</td>
<td>1.18</td>
<td>0.54</td>
<td>0.89</td>
<td>0.50</td>
</tr>
<tr>
<td>$\theta_{M_1}$</td>
<td>0.53</td>
<td>2.57</td>
<td>0.49</td>
<td>2.17</td>
<td>0.71</td>
<td>3.00</td>
</tr>
</tbody>
</table>

7. **The Program ErosPredict**

The Program ErosPredict implements the models described in Section 3. At present only the erosion probabilities and the expected amount of erosion is reported. That is, the 2.5%, the 50% and 97.5% quantiles for the predicted values are not calculated in the present implementation. The features of the program are summarized below.

- Input: Values of the predictor variables of Type I
- Output (currently): Erosion risk given as: i) Probability of erosion, ii) Amount of erosion, given that there is erosion and iii) The combined expectation

8. **Discussion**

We have shown how to model occurrence of erosion by establishing a two step model. It has also been shown that the predictive abilities of the model can be improved considerably by utilizing the incomplete data as well. It was found that the ability to predict cases with erosion is markedly improved when updating the parameters on the basis of a data set which is balanced with respect to erosion. This suggests that in a future study, one should put more emphasis on measuring identifying slopes where erosion takes place. The next step is to perform a more practical evaluation of the model in connection with large scale predictions.

**Literature**
