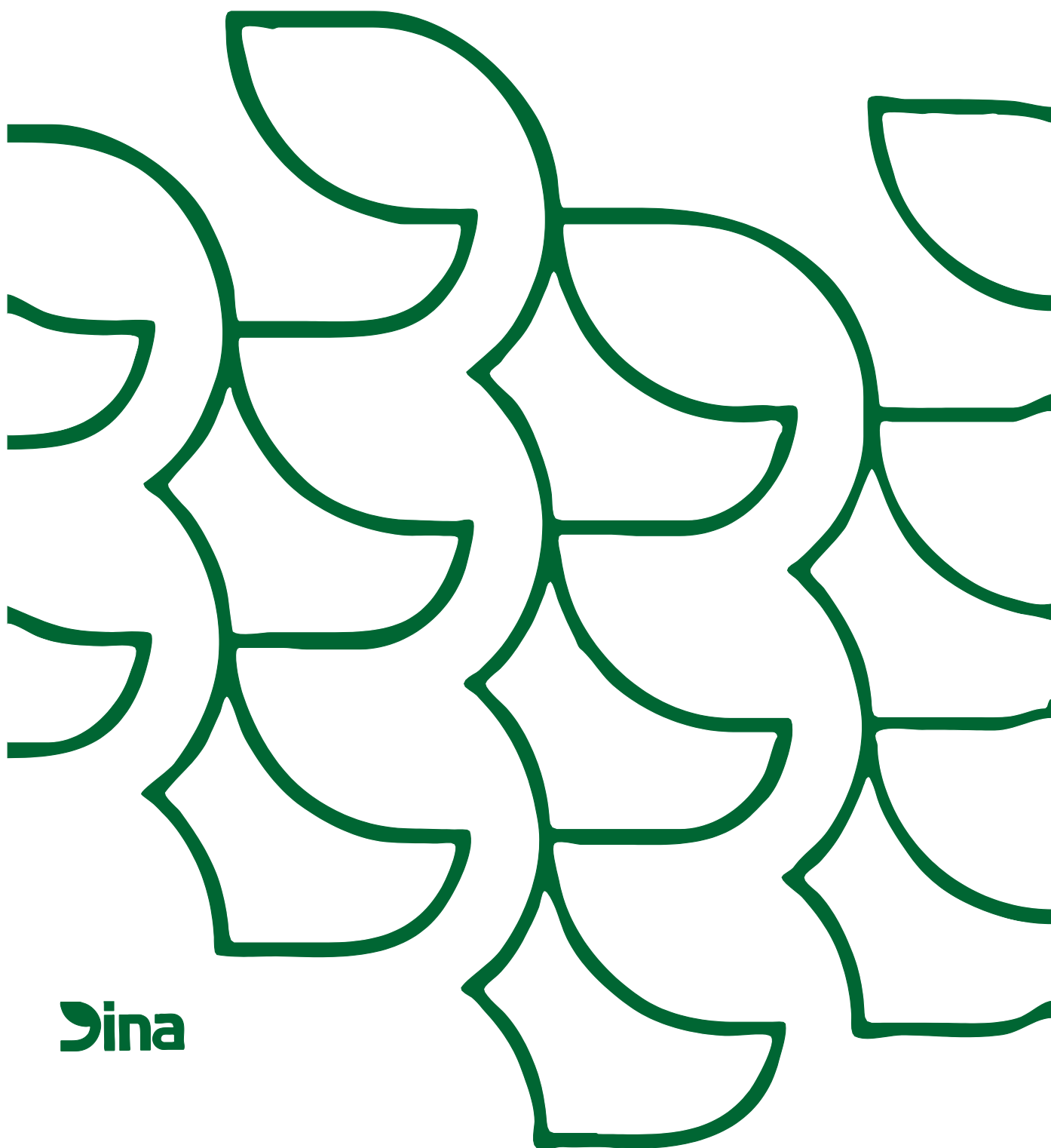


ErosPredict –

A Program for Predicting Soil Erosion

Søren Højsgaard, Helle Højmark Rasmussen & Jørgen Djurhuus

Dina Notat No. 100 · May 2002



Dina



ErosPredict -

A Program for Predicting Soil Erosion

Søren Højsgaard¹), Helle Højmark Rasmussen & Jørgen Djurhuus

Dina Notat No. 100 · May 2002

This note is also available on www at URL:
<http://www.jbs.agrsci.dk/sorenh/Public/notat100.pdf>

Biometry Research Unit
Department of Agricultural Systems
Danish Institute of Agricultural Sciences
Research Center Foulum
DK-8830 Tjele, Denmark
E-Mail:sorenh@agrsci.dk
homepage:<http://www.jbs.agrsci.dk/~sorenh>

ErosPredict – A Program for Predicting Soil Erosion

Søren Højsgaard*, Helle Højmark Rasmussen and Jørgen Djurhuus

Abstract

A probabilistic model for predicting the risk (and amount) of water erosion is established on the basis of data collected in Denmark over a 7 year period. The model is implemented 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 for both 1) the risk of erosion occurring and 2) the amount of erosion.

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.

The data available were collected over a 7-year period in Denmark. During this period erosion by water was measured on a total of 189 slopes at 20 different

*Biometry Research Unit, Department of Agricultural Systems, Danish Institute of Agricultural Sciences, Research Center Foulum, DK-8830 Tjele, Denmark. E-mail: sorenh@agrsci.dk, homepage: <http://www.jbs.agrsci.dk/~sorenh>

locations. A slope was defined as a field with uniform cultivation, eg. 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. Measurements of erosion were performed in late autumn and in spring. The total number of observations was 1041, of which 213 had erosion. The observations with erosion were highly skewed with a 75% percentile of $1.49 \text{ m}^3 \text{ ha}^{-1}$. 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. 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. Secondly, the incomplete data are used to achieve improved parameter estimates using Markov Chain Monte Carlo methods.

2 Data

2.1 Notation

Let D be the entire data set. The data set is divided into two parts: D_C containing all the complete observations and D_M containing the observations where the type II explanatory variables are missing.

2.2 Erosion data

There are 846 measurements of the amount of erosion together with measurements of (some of) the explanatory variables. The erosion is measured as furrow volume per area (m^3/ha). The type I explanatory variables are presented in Table 1. The measurements of the precipitation in Table 1 are since last cultivation (max. one

year). The type II variables are presented in Table 2 below. 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.

Table 1: Type I explanatory variables (those easy to obtain).

Type I – covariates	
$x_{d_{20mm}}$	Days with precipitation greater than 20 mm
$x_{d_{30mm}}$	Days with precipitation greater than 30 mm
$x_{d_{5_8mm}}$	Days with precipitation between 5 and 8 mm
$x_{thaw_{fr}}$	Precipitation and thaw on frozen soil; mm
$x_{ls_{99}}$	LS 99% quantile
$x_{ls_{mean1}}$	Mean of LS-calculations for the region
$x_{ln_{clsi_u}}$	Ln of sum of clay and silt (2-20 μ m) at upslope; ln(%)
$x_{ln_{prec}}$	Ln of precipitation; ln(mm)
Type I – factors	
x_{cult}	System of cultivation – 1: grain; 2: Christmas trees; 3: ploughed; 4: stubble harrowed
x_{asp2}	Aspect – 1: northwest, north, northeast, east; 2: southeast, south, southwest, west
$x_{wst_{12}}$	Water impermeable layer – 1: no/some; 2: yes
x_{year}	Identification of the erosion year: – 1: spring 1994; 2: fall 1994 and spring 1995; ... ; 6: fall 1998 and spring 1999; 7: fall 1999
x_{region}	Region where the slope unit belongs to

Table 2: Type II explanatory variables (those difficult to obtain)

Type II – covariates	
$x_{mud_{1}}$	Roughness of the soil in the direction parallel to the plowing direction, ln of Mean Upslope Depression (MUD) at 0 degree slope; ln(mm)
$x_{mud_{2}}$	Roughness of the soil in the direction perpendicular to the plowing direction, ln of Mean Upslope Depression (MUD) at 0 degree slope; ln(mm)

The individual measurements are indexed by i in what follows. Let $C(i)$ denote the cultivation of the i 'th observation, $A(i)$ the aspect of the i 'th observation, $W(i)$ the type of water impermeable layer for the i 'th observation and $YR(i)$ the

erosion–year and the region of the i 'th observation. (See Table 1 for the definition of an erosion–year.)

About 20% (213/1041) of the observations contained erosion. Figure 1 shows a plot of the logarithm of the amount of erosion against the erosion year. The Figure indicates which data the model for the amount of erosion is based on.

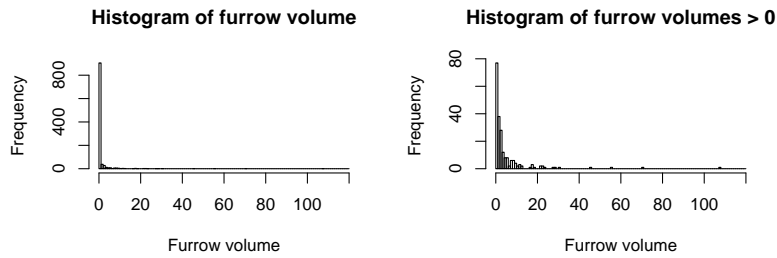


Figure 1: Erosion data follow a right skewed distribution with point mass in zero.

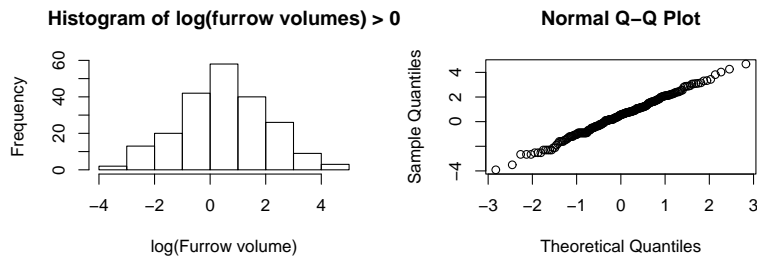


Figure 2: When erosion is present, the amount of erosion can be described by a log–normal distribution.

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 skewed distribution with a point mass in 0 in a two–step way described below. These models, which are described in detail below, are all based on the complete data D_C , 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.

3.1 Modelling the Erosion Risk

To model the erosion risk, a random variable Z is defined as

$$Z = \begin{cases} 1 & : \text{Erosion} \\ 0 & : \text{No erosion} \end{cases}$$

The probability of erosion taking place is modelled by a logistic regression

$$Z_i = 1 | x_i \sim \text{bern}(p_i), \text{ where} \quad (1)$$

$$\begin{aligned} \text{logit}(p_i) &= \delta_{C(i)}^p + \zeta_{A(i)}^p + \eta_{W(i)}^p + \beta_1 x_{\text{ln_clsi_u},i} + \beta_2 x_{\text{thaw_fr},i} + \\ &\beta_3 x_{\text{d_30mm},i} + \beta_4 x_{\text{ls_99},i} + \beta_5 x_{\text{mud_1},i} + \beta_6 x_{\text{mud_2},i}. \end{aligned} \quad (2)$$

The terms $\delta_{C(i)}^p$, $\zeta_{A(i)}^p$ and $\eta_{W(i)}^p$ are the effects of the cultivation, the aspect and the water impermeable layer, respectively. Note that $\text{logit}(p_i)$ depends on the type II variables $x_{\text{mud_1}}$ and $x_{\text{mud_2}}$.

3.2 Modelling the amount of Erosion

Given that there is erosion, the amount of erosion is modelled by a log-normal distribution (See Figure 2) as

$$Y_i | Z_i = 1, x_i \sim \log N(\mu_i^y, \sigma_y^2), \text{ where} \quad (3)$$

$$\begin{aligned} \mu_i^y &= \delta_{C(i)}^y + \eta_{W(i)}^y + \gamma_1 x_{\text{ln_clsi_u},i} + \gamma_2 x_{\text{ln_prec},i} + \\ &\gamma_3 x_{\text{d_5.8mm},i} + \gamma_4 x_{\text{ls_mean1},i} + u_{YR(i)} \end{aligned} \quad (4)$$

Here $u_{YR(i)} \sim N(0, \sigma^2)$ is a random effect of the region in the given erosion year. Note that μ_i^y does not depend on the x_{mud} variables.

3.3 Modelling the Mean Upslope Depression (MUD)

The x_{mud} variables $x_{\text{mud_1},i}$ and $x_{\text{mud_2},i}$ are modelled by

$$x_{\text{mud_j},i} \sim N(\mu_{ij}^m, \sigma_m^2), \text{ where} \quad (5)$$

$$\mu_{ij}^m = \alpha_0 + \alpha_{1D(i)j} + \alpha_{2D(i)} x_{\text{ln_clsi_u},i} B + \alpha_3 x_{\text{d_20mm},i} B + u_{\text{fi_eld}} \quad (6)$$

and $u_{\text{fi_eld}} \sim N(0, \sigma_{\text{fi_eld}}^2)$. Here $u_{\text{fi_eld}}$ describes the field-to-field variation and B is 1 if the system has been cultivated in the fall ($D(i)$: 1,3,4) and 0 otherwise.

3.4 Graphical Representation of the Models

Figure 3 shows the structure of the models established above.

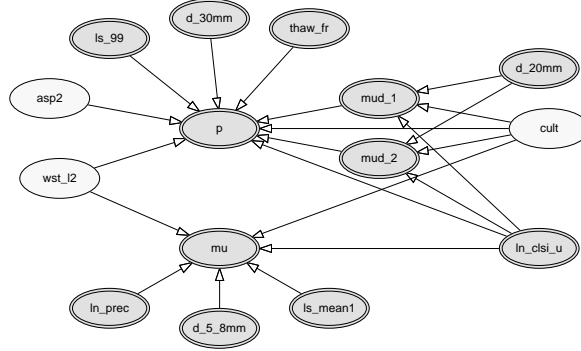


Figure 3: Illustration of the dependencies in the model. In the Figure, p is p_i and μ is μ_i^y .

4 Combining the Erosion Risk and Erosion Amount

The model in (3) 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 combine the expected level from (3) with the probability of erosion occurring, i.e. with (1). 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 formulas for conditional means and variances (which can be found in any standard text book on statistics), it can be shown that

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

However, since the distribution is highly skewed, an interval like

$$E(Y) \pm 2\sqrt{\text{Var}(Y)}$$

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 4. The prediction intervals for the cases shown in Figure 4 are presented in Table 3

Table 3: Median and a 95% prediction interval for the cases shown in Figure 4.

2.5%	50%	97.5%	Mean
0.0	0.0	9.7	1.5
0.0	1.6	10.9	2.9
0.0	5.2	13.6	5.5

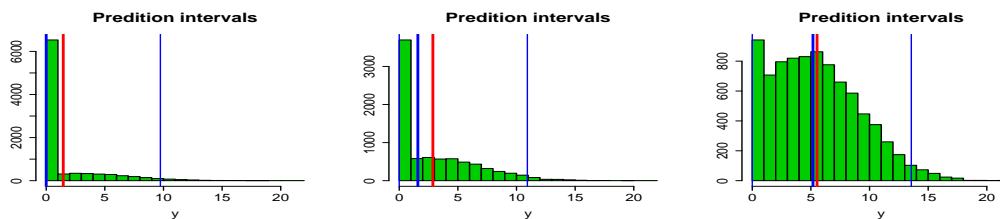


Figure 4: Predictive distribution for three different scenarios (i.e. three different values of the Type I variables). The vertical lines indicate the 2.5%, the 50% and 97.5% quantiles as well as the means given in Table 3.

5 The Program ErosPredict

The Program `ErosPredict` implements the models described in Section 3. Figure 5 shows the graphical interface to the program. Program (and documentation) available as GUI and for batch processing from

<http://www.jbs.agrsci.dk/biometri/software>

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 computations in `ErosPredict` are based on a Bayesian network implemented in HUGIN, (Andersen, Olesen, Jensen and Jensen 1989). The structure

of the Bayesian network is given in Figure 3. In Figure 3 the erosion probability p appears as a continuous variable. In fact, the way it is implemented is by regarding $\text{logit}(p)$ as a continuous (normal) variable with a very small variance. `EROSPREDICT` subsequently calculates the probability and the other quantities described in Section 4.

The features of the program are summarized below.

- Input: Values of the predictor variables of Type I
- Output (currently): Erosion risk given as:
 - Probability of erosion
 - Amount of erosion, given that there is erosion
 - The combined expectation
- Output (to come):
 - Median
 - Prediction interval

6 Improving Parameter Estimates using the Incomplete Data

About 60% of the observations contain no information about the type II variables and were therefore not used in establishing the models in Section 3. However, it is appealing to try to utilize all available information, i.e. the incomplete data D_M as well.

Recall that the structure of the models in Section 3 were established on the basis of the complete data D_C alone. The parameters of the models (symbolically written as θ) were also estimated in this process. Let $\hat{\theta}_C$ denote the estimate of the parameters based on the complete data.

6.1 Adopting a Bayesian Approach for Improving Parameter Estimates

There are several options for utilizing the incomplete data for obtaining updated (and hopefully improved) parameter estimates. One option is to put the problem into a Bayesian framework, see e.g. Gilks, Richardson and Spiegelhalter (1996).

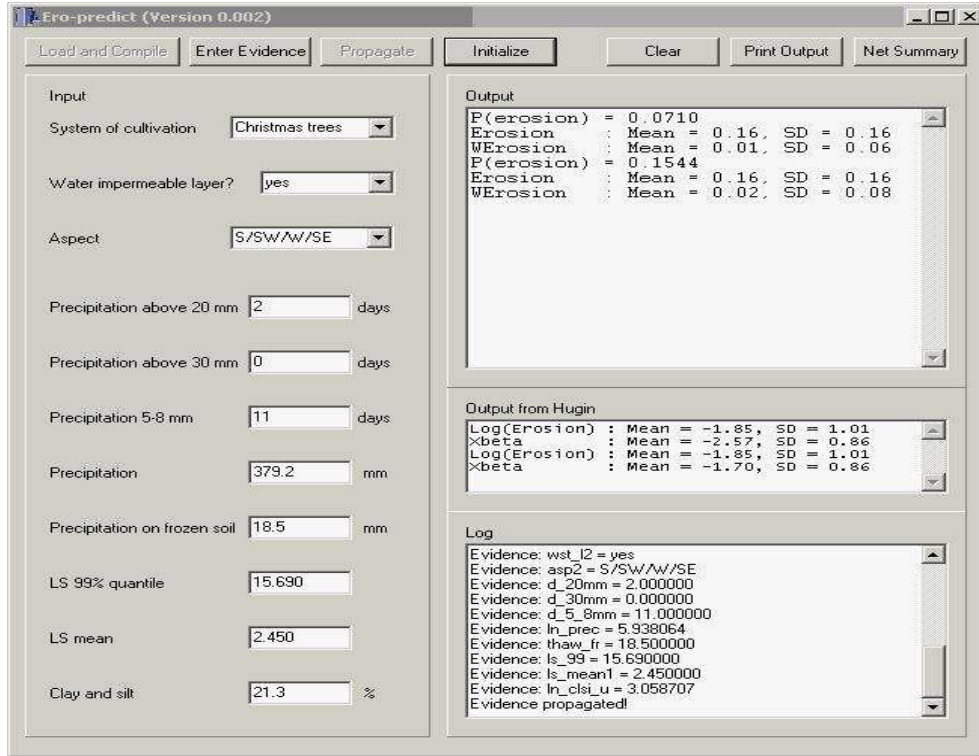


Figure 5: Graphical interface for the program ErosPredict. The program is also available for batch processing.

In this connection the parameters θ are regarded as random quantities with a distribution $\pi(\theta)$. The models in Section 3 when applied to the incomplete data D_M are symbolically written $P(D_M|\theta)$. Using Bayes formula, the posterior distribution of θ given the data D_M can then be obtained as

$$p(\theta|D_M) = \frac{P(D_M|\theta)\pi(\theta)}{\int P(D_M|\theta)\pi(\theta)d\theta}$$

These computations were made in the BUGS program, (Spiegelhalter, Thomas, Best and Gilks 1996).

We have then taken the mode (i.e. the value of θ for which $p(\theta|D_M)$ is highest as the updated parameter estimate $\hat{\theta}_M$. In doing so, the prior distribution $\pi(\theta)$ was for the mean value parameters taken to be independent normals with hyper parameters for mean and standard deviations as obtained from the analysis of the complete data. For the variance parameters, the prior distributions were assumed

to be Gamma, again with hyper parameters specified to match the estimates from the analysis of the complete data.

6.2 Evaluating the Predictive Ability

To evaluate the predictive abilities of the models, we have split the incomplete data D_M randomly into two data sets D_{M_1} and D_{M_2} . The model parameters θ were then update using only D_{M_1} giving an updated estimate $\hat{\theta}_{M_1}$.

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 \tilde{D}_{M_1} of D_{M_1} consisting of all cases with erosion and a random sample of the same size of cases without erosion. Hence \tilde{D}_{M_1} 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 $\tilde{\theta}_{M_1}$

The models were then used to predict the erosion in D_{M_2} using $\hat{\theta}_C$, $\hat{\theta}_{M_1}$ and $\tilde{\theta}_{M_1}$. Measures of predictive ability are discussed in Section 6.3 below.

6.3 Measures of Predictive Ability

Let N_1 be the number of observations with erosion and N_0 the number of observations without erosion in the prediction data set D_{M_2} . The scores below were used to evaluate the prediction of p_i and μ_i^y . Common to all scores is that a low value indicates a good predictive ability.

Brier score:

$$Brier_1 = \frac{1}{N_1} \sum_{i: \text{with erosion}} (1 - p_i)^2$$

$$Brier_0 = \frac{1}{N_0} \sum_{i: \text{without erosion}} p_i^2$$

Log score:

$$LogScore_1 = -\frac{1}{N_1} \sum_{i: \text{data with erosion}} \log p_i$$

$$LogScore_0 = -\frac{1}{N_0} \sum_{i: \text{data without erosion}} \log(1 - p_i)$$

Precision:

$$MSEP = \frac{1}{N} \sum_i (y_i - p_i \exp \mu_i^y)^2$$

7 Prediction Results

Table 4 below shows the prediction results for the data set D_{M_2} for the three parameter estimates. The general conclusion is that when updating the parameters using D_{M_1} , the model get slightly better in predicting cases with erosion, while performance in predicting cases without erosion is marginally worse.

When updating the parameter estimates from the balanced data \tilde{D}_{M_1} 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.

With regard to the expected amount of erosion an improvement is also noted when utilizing the incomplete data.

Table 4: Prediction results for three different sets of model parameters.

Score	$\hat{\theta}_C$	$\hat{\theta}_{M_1}$	$\tilde{\theta}_{M_1}$
<i>Brier</i> ₁	0.47	0.43	0.23
<i>Brier</i> ₀	0.07	0.08	0.18
<i>LogScore</i> ₁	1.33	1.16	0.64
<i>LogScore</i> ₀	0.26	0.28	0.56
<i>MSPE</i>	93.83	93.22	92.19

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.

References

- Andersen, S., Olesen, K., Jensen, F. and Jensen, F. (1989). HUGIN — a shell for building Bayesian belief universes for expert systems, *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI)*, Detroit, MI.
- Gilks, W., Richardson, S. and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.
- Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996). *BUGS Manual 0.50*, MRC Biostatistics Unit Cambridge.