This paper was peer-reviewed for scientific content.
Pages 979-993. In: D.E. Stott, R.H. Mohtar and G.C. Steinhardt (eds). 2001. Sustaining the Global Farm. Selected papers from the 10th International Soil Conservation Organization Meeting held May 24-29, 1999 at Purdue University and the USDA-ARS National Soil Erosion Research Laboratory.

A Non-parametric Analysis of Qualitative and Quantitative Data for Erosion Modeling: A Case Study for Ethiopia

B.G.J.S. Sonneveld, M.A. Keyzer and P.J. Albersen*

ABSTRACT

The objectives of this paper are twofold. First, it compares the discriminatory power of qualitative expert judgements with actual soil losses to express class boundaries in physically measured, quantitative terms. Secondly, it investigates the properties of a postulated functional relationship between soil loss and readily available explanatory variables on both, their reliability of fit and behaviour. The study uses quantitative soil erosion data of runoff plots of the Soil Conservation Research Project in Ethiopia. Qualitative expert judgements on the state of erosion for the same runoff plots were obtained through a questionnaire. The study applies a non-parametric technique that uses a flexible method of curve fitting. The first exercise applies this technique to determine the quantitative boundaries (soil losses) of qualitative classes. The results reveal a positive relationship between erosion hazard assessment by the expert and actual soil losses, however, experts tend to overestimate. In the second exercise, the mollifier program is used to visualize non-parametric estimates in 3-D graphs that show non-linear relationships and reliability of the estimates. The results indicate that soil loss should be modelled separately for annual crops and land use types with a permanent coverage. Further findings show that annual runoff has an almost linear relation with annual soil loss. An index derived from monthly rainfall data and the adjusted Cooks' method seems promising to represent the hydrological factor in the model. Most relations show a poor 'goodness of fit', which anticipates low correlation coefficients in future parametric, models and indicates that additional variables should be included.

INTRODUCTION

The detrimental effects of water erosion on soil productivity are particularly manifest in the least developed countries, where farmers are highly dependent on intrinsic land properties and unable to ameliorate soil fertility through application of purchased inputs. The highlands (above 1500 m) of Ethiopia, which carry among the highest population densities in Africa, are an important case in point. These highlands constitute 43 per cent of the country and are endowed with a high soil fertility that account for 95 per cent of the cultivated area. Here soil losses may reach annual levels of 200-300 ton per hectare (Hurni, 1993, Herweg and Stillhardt, 1999) affecting 50 per cent of the agricultural areas (UNEP, unpublished data) and 88 per cent out of a total population of 60 million people. Moreover, the fast grow rate of population (2.2 per cent annually; World Bank, 1998) causes a steady increase of the pressure on the land. Hence, there is an urgent need for policy interventions that arrest soil degradation and rehabilitate degraded areas. Since it is not possible to measure and experiment with soil erosion measures at every endangered spot in the country, spatial soil erosion models offer a vital tool in the design of these interventions. These models describe for every point on the geographical map the degree of soil erosion in its dependence on both biophysical conditions and actual land use practices and can be used to define options for sustainable land use.

The early soil erosion models consisted of relatively simple response functions that were calibrated to fit a limited number of statistical observations (e.g. USLE, SLEMSA). The current trend is towards replacing these by far more elaborate process based models (e.g. Morgan et al., 1992; Nearing, 1989; Yu et al. 1997). However, in case of Ethiopia and many other developing countries the application of these process based models is not a practical proposition in view of their large data requirements. Moreover, these models are apparently not yet in an operational stage witness the often poor correlations between modelled and observed soil losses (e.g. De Roo et al., 1996; Bjorneberg, 1997; Bonari et al., 1996; Klik et al., 1997, Littleboy et al.,1996 Quinton, 1997). One is thus confronted with the paradoxical situation that much effort is being invested in the development of soil erosion models that will eventually not be applicable to the locations where they are most urgently needed. To address this problem, alternative, qualitative procedures for land hazard assessment have been designed (e.g. Desmet et al., 1995; Gachene, 1995; King et al., 1999) that are based on expert judgement and generate a relative ranking of the degradation status. Sonneveld and Albersen (1999) in turn include this information in an ordered logit model (as in Greene, 1991) that has the expert judgements as dependent variable and the soil, climate and land use characteristics as independent variables. This model was used to both test the consistency of expert judgements in relation to the explanatory factors and to reproduce a judgement corresponding to biophysical and land use conditions at sites for which no expert assessment is available. However, the ordered logit model has two basic limitations. It specifies the boundaries between ordered classes with a common judgement in an indirect way, as unobservable variables, and assumes a linear form for the effect of the explanatory variables.

*Centre for World Food Studies of the Vrije Universiteit (SOW-VU). De Boelelaan 1105 1081 HV, Amsterdam, The Netherlands. *Corresponding author: b.g.j.s.sonneveld@sow.econ.vu.nl

In this paper, both restrictions are being addressed. First, the discriminatory power of qualitative expert judgements is compared with actual soil losses. This enables us to express the class boundaries in physically measurable, quantitative terms. Secondly, the paper investigates the properties of a postulated functional relationship between different measurements of soil losses and a limited number of explanatory variables that are generally available in developing countries. The approach is to look via a flexible method of curve fitting for an expression of soil losses in combination with explanatory factors that yields a surface which is both sufficiently reliable in terms of fit, and sufficiently well behaved (e.g. linear or concave and smooth) to promise a successful mathematical formalization through an explicit parametric form. The flexible curve fitting is effectuated by the non-parametric technique of kernel density regression (e.g. Bierens, 1987). This technique allows for functional forms that follow the observed data closely, so as to reveal possible non-linearities. Associated with it are descriptive statistics on the likelihood density of information at every site, the 'fit' and the error probability of the slope of the function. We apply the mollifier program (Keyzer and Sonneveld, 1998) which, among others, shows kernel density regressions as 3D-graphs that map the dependent variable against the independent variable(s) for fixed values of other exogenous variables, while information on associated statistics is shown in colours or shading of the surface plot and a ground plane. This visual representation is especially practical to explore large data sets and to investigate the properties of relationships where, as in the erosion process, the factors at play are more or less known but little a priori information is available on the functional form to be adopted.

The study uses classified and continuous data on soil and land characteristics and continuous data on precipitation, rainfall erosivity, runoff and soil loss as obtained by the Soil Conservation Research Project (SCRP) in Ethiopia. Qualitative observations on erosion hazard are derived from a questionnaire that was completed by one national and one international soil erosion expert, both associated with the project.

The paper proceeds as follows. Section 2 describes the questionnaire and the compilation of the qualitative assessments as well as the data on explanatory variables. Section 3 briefly discusses the methodology of non-parametric analysis. Section 4 reports on the quantitative interpretation of expert judgements. Section 5 gives a step-wise introduction to the 3-D graphs as generated by the mollifier program and shows how it is used in the quest for a reliable and well behaved representation. Section 6 concludes.

Data sources

SCRP data. The SCRP is co-ordinated by the Centre for Development and Environment, University of Berne in association with the Ethiopian Ministry of Agriculture. The present study uses the data from 28 runoff plots located at seven research areas, six in Ethiopia and one in Eritrea (Fig. 1), that were collected by SCRP during the period 1982-1993. The runoff plots had dimensions of 2×15 square meters and were bounded by galvanised sheets to prevent access of runoff from adjacent terrain. The plots were implemented in farmers' fields and in this way made subject to their regular land management activities. Plots are selected to represent prevailing climate, soil and land characteristics of the research area.

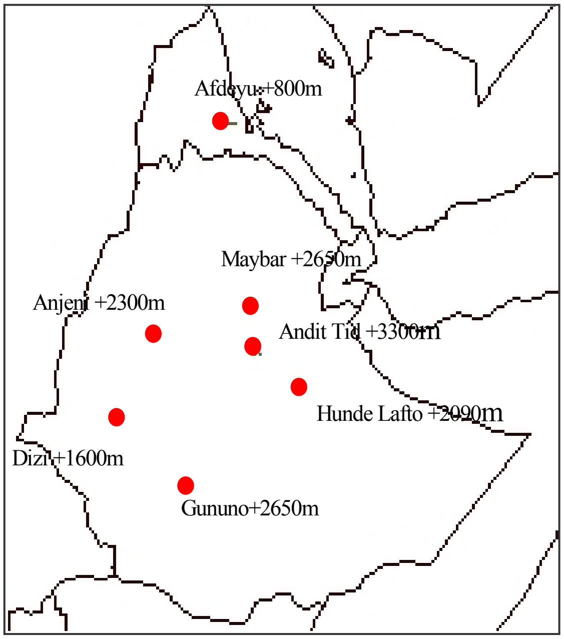

Figure 1 Location of SCRP research areas.

Qualitative data. Qualitative erosion assessments were obtained from one national and one international expert, involved in the SCRP who were asked to deliver their qualitative assessment of annual water erosion hazard for the 28 runoff plots under the land use types and land management in the period 1982-1993, on a scale of five (1=no erosion, 2=slight, 3=moderate, 4=severe, 5=extreme). The first erosion class refers to a situation in which erosion has tolerable levels. Classes 2 to 5 represent an increasing magnitude of the impact of water erosion on an ordinal scale. Thus, class 3 is more severe than the expert makes class 2, but the interpretation of differences in extent of the erosion only. The experts were asked not to consult the historical soil loss records that were registered by the SCRP. Other information conveyed in the questionnaire included: name of research area, plot number, soil type, annual rainfall, slope and land management.

Quantitative data. Quantitative data on erosion, land use, climatic, soil and land conditions were obtained as follows. For each plot, an erosion measurement was conducted in terms of runoff as well as soil loss while the land use information was collected through measurement of crop coverage, biomass and crop yield. For each research area, the climatic characteristics (rainfall, rainfall erosivity and temperature) were recorded and a detailed soil survey (app. 1:10 000) was done at the start of the experiments that provided data on soil and land characteristics of the runoff plots.

Table 1. Land use and C-factor[1].

Sole Cereals		Sole pulses/ potato		Associated crops		Perennials		Rangeland	
crop	C-factor	Crop	C-factor	crop	C-factor	Crop	C-factor	Grass	C-factor
barley	0.452	field pea	0.315	sorg/maiz/bean	0.250	coffee	0.210	grass	0.00945
maize	0.291	haricot bean	0.355	haricot b./barley	0.160	bushland	0.150	bush/gras	0.00100
niger seed	0.604	horse bean	0.246	maize/haricot b.	0.250				
sorghum	0.206	lin seed	0.483	barley/field pea	0.250				
teff	0.337	lentil	0.388	barley/horse b.	0.250				
wheat	0.477	sweet potato	0.350	Barley/lupine	0.250				
				emmerw./horseb	0.250				
				field pea/horseb.	0.250				
				gras/sorg/har. B.	0.250				
				hor.b./field p./ maize/sorgh.	0.250				
				horse b./field p.	0.250				
				horseb/emmerw.	0.250				
				maize/lentil	0.250				
				maize/sorgh./teff	0.250				
				sorghum/ sorghum	0.250				
				sorghum/har. B.	0.250				
				sorghum/potato	0.250				
				sweet pot./barley	0.250				
				teff/teff	0.250				
				wheat/wheat	0.250				
				barley/barley	0.250				
				maize/maize	0.250				
				sorghum/har.b.	0.250				
				sorghum/maize/ha r.b.	0.250				
				wheat/barley	0.100				

C-factors in black are calculated and C-factors in blue are based on assessments and published literature (Morgan, 1995; Lal, 1995) emmer. w.= emmer wheat, har.b.,= haricaot beans, horse (hor.)), b.= horse beans.

Limited data set. To construct the version of the erosion model, which is based on the limited data set, we use readily available data that are found in the regular natural resource databases. We also generate data from existing and already parameterized models.

Crop cover index. The crop cover plays a central role in the erosion process. To measure the average crop cover index we apply a model calculation rather than using the underlying statistical data. This enables us to take advantage of the structural information in our non-parametric analysis and to reduce the number of variables. More specifically, we compute the C-factor of the RUSLE model on the basis of the observed crop coverage, sub-surface and surface coverage and soil roughness according to the Renard et al. (1997) and data from the literature (Morgan, 1995, Lal, 1995). Table 1 shows the land use types that were cultivated in the SCRP plots and their average C-factor.

Hydrology. For the hydrological component of the erosion process three variables were compiled. First, we calculate the Modified Fournier Index (MFI: the sum of the squares of monthly rainfall divided by the total annual precipitation (Arnoldus, 1981)). The MFI seeks to measure the seasonal variability in rainfall erosivity. Secondly, we compute the R-factor of the (R)USLE model, which is based on a continuous rainfall registration and calculated based on the maximum 30-minute rainfall intensity and total amount of rainfall in one shower. Finally, we include the measured annual runoff.

Topography. A single continuous function for the slope gradient (Nearing, 1997) is applied to translate the influence of the topography on the soil erosion process. This function generates an "LS-factor". To translate this factor for rangeland conditions, we follow Renard et al. (1997).

Soils. Concerning soil data the following variables are selected: silt content, organic matter, phases, abrupt textural change, and drainage class. Theoretical evidence that these factors play an important role in the erosion process can be found in Morgan (1995) and Lal (1990).

The Mollifier program: 3D-visualization of kernel density regressions

This section provides some background on the non-parametric analysis by kernel density regression. A more detailed specification is given in annex I.

Mollifier mapping. The mollifier mapping is defined as the following stochastic model:

$$y = E(R(x+\varepsilon)) \qquad (1)$$

where y is the observed soil loss, x is a vector of explanatory variables and ε denotes measurements errors in x. The function R(x+ε) is the unknown erosion function, and the mollifier mapping is the expected value of this function. For an infinite sample of observations spread evenly over the domain of x, it would be possible to evaluate this expected value. However, in practice the value of y must be estimated given a finite sample of size S[1]. For this, one can use the Nadaraya-Watson kernel density estimator:

$$\tilde{y}(x) = \sum_s y^s P_s(x) \qquad (2)$$

where ys and x^s denote observations. Thus, the estimate is a probability weighted sample mean. The probabilities are computed on the basis of the distance of x^s from the given point x, attributing higher weight to nearby points. The probability is calculated on the basis a postulated density function (the kernel) for ε whose spread is controlled by the window size parameter θ. We suppose that all the elements of ε are independently and normally distributed. For small samples, a misspecification of this density will affect the estimate but this effect disappears as the sample size becomes larger.

Mollifier program. The mollifier program offers the possibility to exhibit the estimated $\tilde{y}(x)$ in 3-D graphs as a surface plot or blanket against two independent variables on, say, a 50×50 grid, while controlling for other explanatory variables by setting them, say, at their sample mean. In the default mode the program generates a colour shift in the surface plot to reflect the likelihood ratio of the observation density, which measures the number of observations on which the function evaluation is based at that point. The colours in a ground plane below the surface plot shows the probability of the actual y falling within a prescribed interval around the mollifier mapping, whose upper and lower bounds are specified as a percentage (default = 10) of the sample mean $\bar{y}$. However, the statistical information can be exchanged for other 'mollified' covariates to identify their location in the selected dimensions.

The mollifier assesses the partial derivative of the regression curve as well as a measure of reliability for it. For this, it calculates the first partial derivative to x_k at point x, where k represents an explanatory variable, at all data points.

$$\frac{\partial \tilde{y}(x^t)}{\partial x_k} = \sum_s \frac{\partial P_s(x^t)}{\partial x_k} y^s \qquad (3)$$

The mollifier program uses the band (or window) width as a control variable to specify the neighbourhood of x whose points affect the prediction of $\tilde{y}$. The user can vary the window size relative to a benchmark (optimum) level defined by:

$$\theta = \left(\frac{4}{n(d+2)}\right)^{\frac{1}{d+4}} \qquad (4)$$

with n being the number of observations and d being the number of exogenous variables (Silverman, 1986). If the averaging should emphasise nearby points, the window size should be small. The larger the window size, the tighter the blanket and the less it will follow the profile of observations. We will keep the window size at its benchmark level.

Quantifying the class boundaries of a qualitative assessment

Fig. 2 indicates how much actual soil loss corresponds to the qualitative assessment by experts, with the x-axis values 1 = 'no erosion', 2 = 'moderate erosion', .. 5 = 'very severe erosion'. As the figure shows, a wide range of soil losses can be observed for each of the qualitative classes, few observations belong to the classes 2 and 5, in classes 3 and 4 most observations lie in the lower range and, finally, the means by class are increasing, as could be expected.

In Fig. 3, the black line is the kernel density regression or mollifier curve for the five classes. This line is increasing, just like the class means of Fig. 2. The upper line is an estimate of the probability of a deviation by more than 10.7 units (i.e. 20 per cent of the sample average) from the mollifier curve. The probability of error increases steeply after class 1, due to the areas, which received a high rating but where, no actual soil loss was observed. Table 2 gives the class boundaries at midrange between the class values 1-5 of the individual experts and their combined assessments. The upper boundary of the first class of expert 1 is at eight units, which corresponds remarkably well with the often-assumed threshold values for sustainable development (Morgan, 1995). Expert II gives a value, which is somewhat

Table 2. Class boundaries of qualitative assessments

Class	Expert I	Expert II	Combined
No erosion	0-8	0-19	0-14
Slight	>8-32	>19-27	>14-28
Moderate	>32-75	>27-71	>28-74
Severe	>75-102	>71-134	>74-114
Very Severe	>102	>134	>114

[1] The minimum sample size for a relative mean square error $E(\tilde{y}-y)^2/y^2 \le 0.1$ are for 2 independent variables S=5; for 3 independent variables S = 67 for 4 independent variables S = 223 and for 5 independent variables s = 768 (Silverman, 1986). Note that these samples sizes hold for regression in the full dimensions of the independent variables, while the mollifier figures are based on two (visual) independent variables and conditioned values of other independent variables. Consequently, for mollifier pictures where the number of independent variables is larger than 2, the sample size S is always smaller to attain the accuracy indicate above.

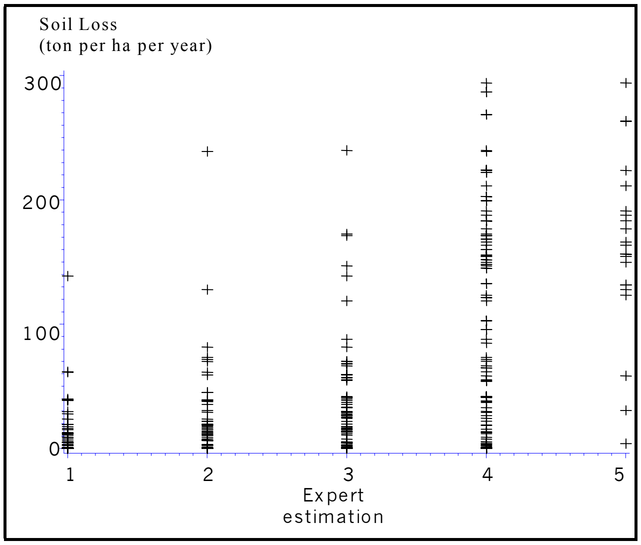

Figure 2. Measured soil loss by class

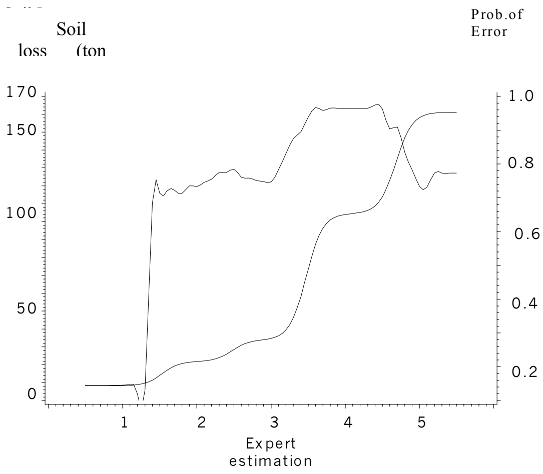

Figure 3. Kernel density regression of soil by class: mean value and probability of error

Table 3. Hit ratio between expert and observed classifications.

Expert

o		1	2	3	4	5	Total
b	1	74	39	36	22	1	172
s	2	8	11	15	6	1	40
e	3	7	14	27	24	2	74
r	4	0	1	2	6	0	9
v.	5	1	2	6	52	19	80
	Total	90	67	86	110	22	375

higher. Further, we notice that the upper thresholds of classes 2 and 3 are almost the same but for class four we observe a difference of 30 ton per ha per year.

Next, now the class boundaries have been estimated it becomes possible to compare the actual observation of the soil loss with the judgement of the expert. We will do this for the combined assessments of both experts and classify in table 3 their classifications against actual observations. The cells on the diagonal contain the observations that agreed 137 in total (or 37 per cent of the cases). In the 145 instances (38 per cent) above the diagonal the expert over-estimated the losses and in 93 instances (28 per cent) the converse was true. With respect to the size of the error it may be noted that the majority of the underestimations are one class lower than the observed soil loss class. We also notice that the hit ratio is high for class 1. Further we observe that the experts classified many cases higher than the class 1, whereas in fact the soil loss did not exceed its upper boundary. Class 4 has many underestimations but together classes 4 and 5 perform better with 189 correct classifications (50 per cent), 41 underestimations (11 per cent) and 145 overestimations (39 per cent).

Explaining soil erosion with a limited data set

This section presents results from kernel density regressions that seek to explain soil erosion based on a limited set of explanatory variables. Our criteria for eventually selecting a specification are: (a) *reliability*: probability of error in soil loss and probability of wrong sign for derivative, (b) *regularity*: monotonicity of the 3D-planes monotonic as well as concavity, convexity, or both (i.e. linearity); this eases subsequent parametric estimations, but more importantly, it suggests that the explanatory factors can indeed capture the fundamentals; in contrast, if the planes are bumpy, there are presumably unspecified factors at play which cause multiple changes in slope and curvature; and finally (c) *availability* of explanatory variables. The presentation starts with a stepwise introduction of the 3-D graphs as generated by the mollifier program, and then turns the search for a suitable specification.

Introducing the Mollifier Graphs

Scatter plot of rainfall erosivity (MFI) and topography factor

Fig. 4 is a three-dimensional scatter plot of the observed soil loss (ton per ha per year) against a rainfall erosivity index and a topography factor. The rainfall erosivity is represented by the Modified Fournier Index (MFI) while the influence of the topography on the erosion process is represented by the LS-factor that measures the influence of the slope gradient on the erosion process[2] The limitations of the presentation by such a scatter plot are evident: it is difficult to infer any relationship between the variables and it is not possible to control the relationship for other aspects such as soil factors and land use.

[2] Sensitivity tests showed that the estimated values of the dependent variable were robust for the C-factor values derived from the literature.

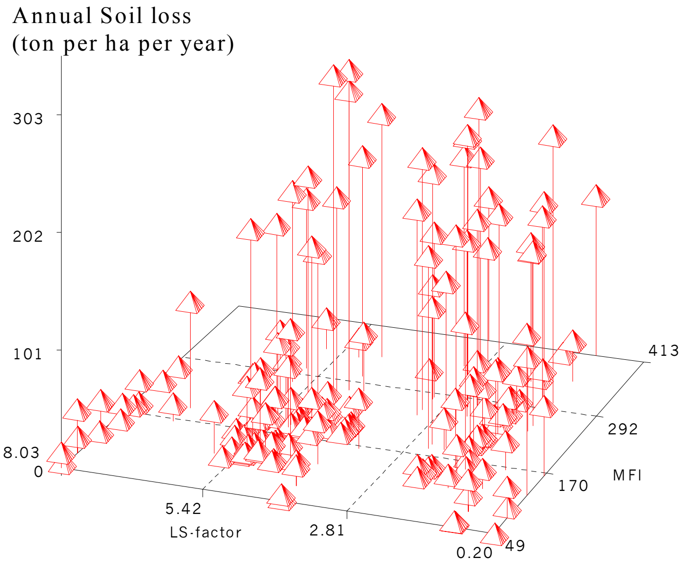

Figure 4. Scatter plot of soil loss against Modified Fournier Index (MFI) and topography.

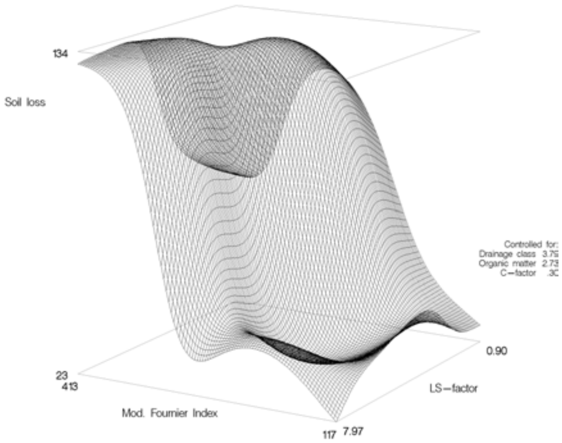

Figure 5. Soil loss against rainfall erosivity (MFI) and topography (LS-factor).

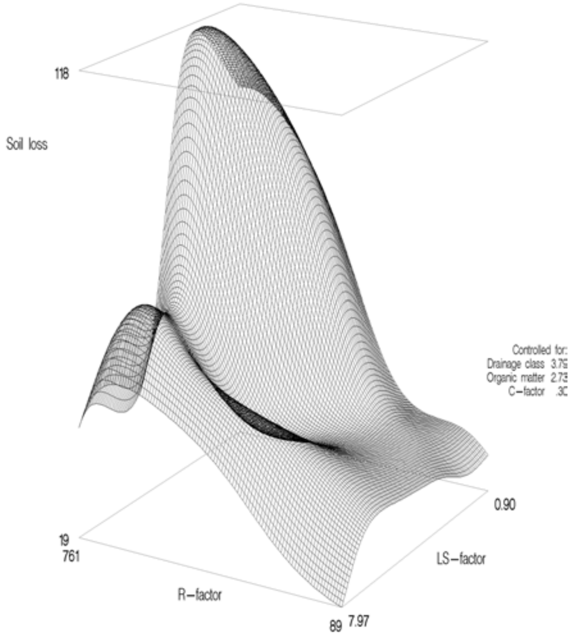

Figure 6. Soil loss against rainfall erosivity (R-factor) and (LS-factor).

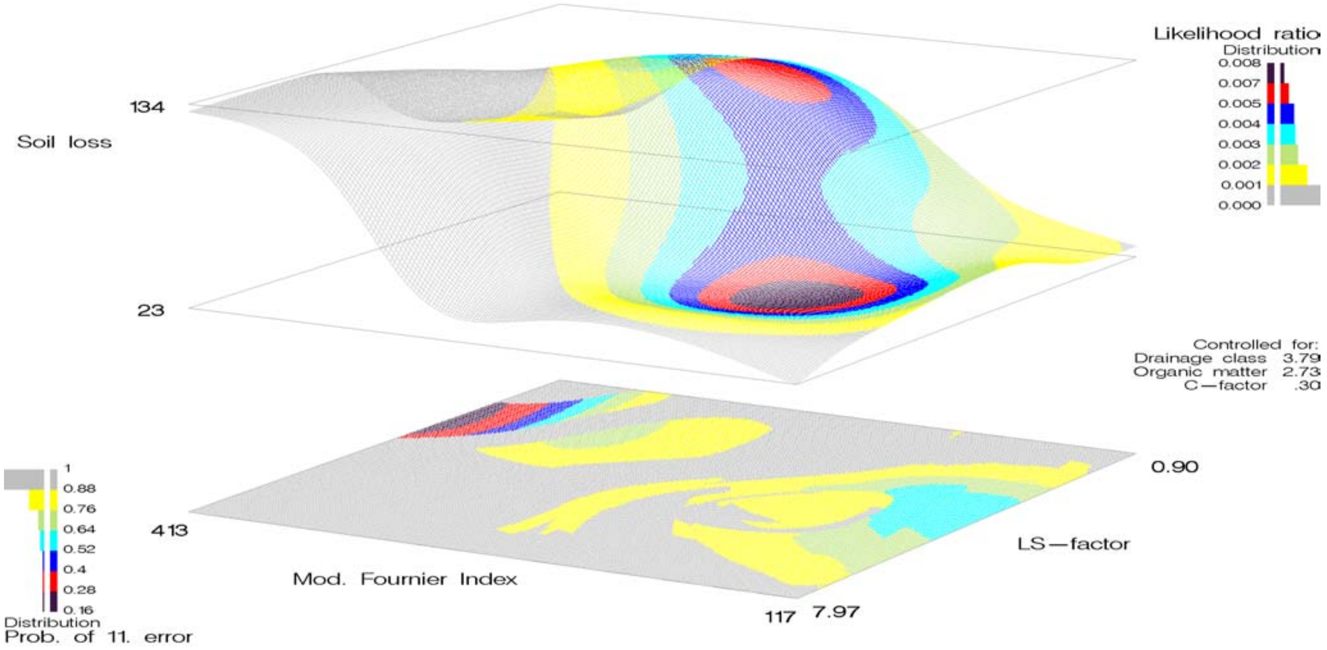

Figure 7. Soil loss against MFI and LS-factor. Covariates Likelihood ratio and probability of error.

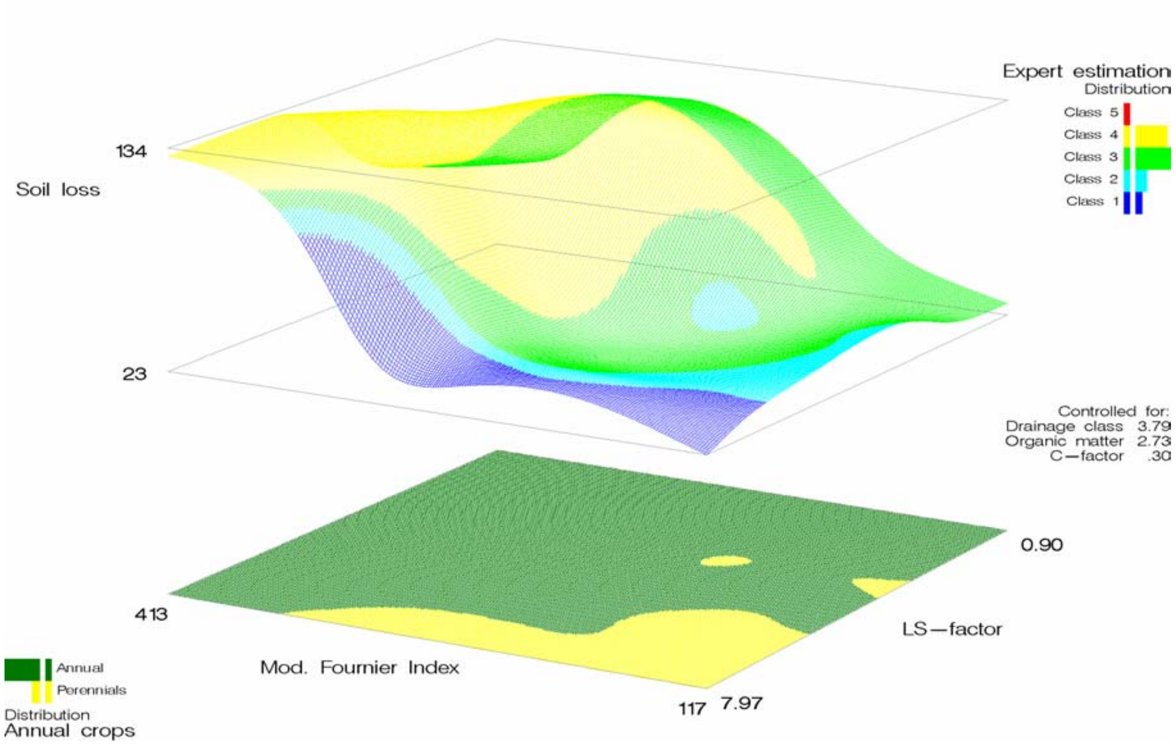

Figure 8. Annual soil loss values against MFI and LS-factor: Covariates expert classifications and land use groups.

Mollified surface plot of rainfall erosivity (MFI) and topography factor

Fig. 5 shows the surface plot of the estimated mollifier mapping with soil loss values regressed against topography and rainfall erosivity, while being conditioned (mean) values of two soil erodibility factors (organic matter and drainage) and land use coverage factor (C-factor of the USLE)[3] Notice that the figure has been rotated a 150 degrees from its point of origin. We see that for the lower to middle slope range, the soil loss increases more or less linearly at higher rainfall erosivity values but the curve drops for the lower slope values and forms a plateau for the higher ones. For the highest slopes, the relationship between erosivity and soil loss seems to be weak. The curve shows several bumps instead of the monotonic rise that could have been expected on theoretical grounds. Unexpected is also the reduction in soil loss for the highest slope values in the middle range of rainfall erosivity.

Replacement of rainfall erosivity by R-factor

The frail relationship between soil loss and its explanatory variables might in part be due to the use of the MFI instead of a more advanced and accurate variable such as the R-factor of the RUSLE model. However, as shown in Fig. 6, replacing the MFI by the R-factor does not make the relationship better behaved. This holds especially at higher R-factor values. The descending trend for higher LS-factors remains and the number of bumps stays large. Therefore, we return to the MFI as a measure of rainfall erosivity.

Complete mollifier picture

We add now to the mollifier curve of Fig. 5 descriptive statistics on the likelihood ratio of the observation density and on the probability of error (Fig. 7). The likelihood ratio is depicted through a colouring of the surface plot while the reliability of the estimate for a 20 per cent deviation (11 ton per ha per year) of the mean for the co-ordinate is reflected in the colouring of the ground plane. The legends of the likelihood ratio and reliability appear on the upper right and lower left side, respectively. The class boundaries for the colourings are found at the outside of the legend, while the histograms measure the percentage of total area in every class. It appears that the likelihood ratio of the density of observations is high at two places: at the higher range of the topography and lower rainfall erosivity values and at the lower topography values and the middle range of the erosivity values. This is where most observations are concentrated. In the area with high rainfall, erosivity and high slope gradients observations are relatively few. We also notice the scattered reliability pattern in the ground plane, with highest probability of error in the lowest reliability classes.

Land use and expert classification as covariates

The unexpected reversed effect of the topography deserves some more attention. In Fig. 8, we introduce the land use as a covariate in the plane to locate their appearance in relation to rainfall erosivity and topography. For this purpose the land use was subdivided into two groups with similar temporal and spatial development of the leaf area and, hence, resembling soil coverage features: annuals (sole cereals, sole pulses, associated annual crops) and perennials (coffee and grasses). The colour shift clearly depicts that perennials are cultivated at higher slope gradients and higher rainfall values while the annuals are cultivated in the middle and lower slope gradients Obviously, the coverage of perennials annuls the expected topography effect on soil erosion and the calculated C-factors do not compensate the estimation of expected soil losses. The expert classification is depicted as a covariate in the surface plot and follows the contour lines of soil loss values for the higher ordered classes.

Location of soils

As regards soil-related characteristics, it must be stressed that the soil surveys were conducted at the inception of the erosion trials and that therefore the soil data can be safely treated as explanatory factors since they are not the result of the recorded soil losses. For a first orientation, we show through the colouring of the ground plane in Fig. 10 the soils that were identified in the database. The prevalence of Luvisols, Nitisols, Phaeozems and Regosols is clear, while Lithosols and Andosols are next in importance. Yet we do not find any clear correspondence pattern between soil loss and soil types.

Relation with aggregate stability and organic matter

Typical soil characteristics that play an important role in the erosion process are aggregate stability of soils and organic matter content. The aggregate stability is a main determinant of the sensitivity to detachment and entrainment and the organic matter plays a crucial role in the structure formation of soils and increases the resistance against the dispersive forces of rainfall and runoff (Lal, 1987). In Fig. 11 we depict the aggregate stability as assessed in the field and the organic matter content as determined in the laboratory. Stability appears as a covariate in relation to rainfall erosivity and topography in the surface. Organic matter is now calculated as a covariate in the ground plane and was for the regression removed form the dependent variables to avoid a problem of endogeneity. The resulting patterns for covariates more or less confirm theoretical expectations. Soil losses are highest for the weakest aggregate stability and increase gradually as the organic matter content diminishes. However, soils with a strong aggregate stability classification also record high losses, while the relation with the moderate stability class is also not equivocal.

Replacement of MFI by annual runoff and introduction of silt fraction as covariate

Another important soil component related to the erodibility of the soil is the silt fraction (particle size 0.002-0.05 mm).

[3] Sensitivity tests showed that the estimated values of the dependent variable were robust for the C-factor values derived from the literature.

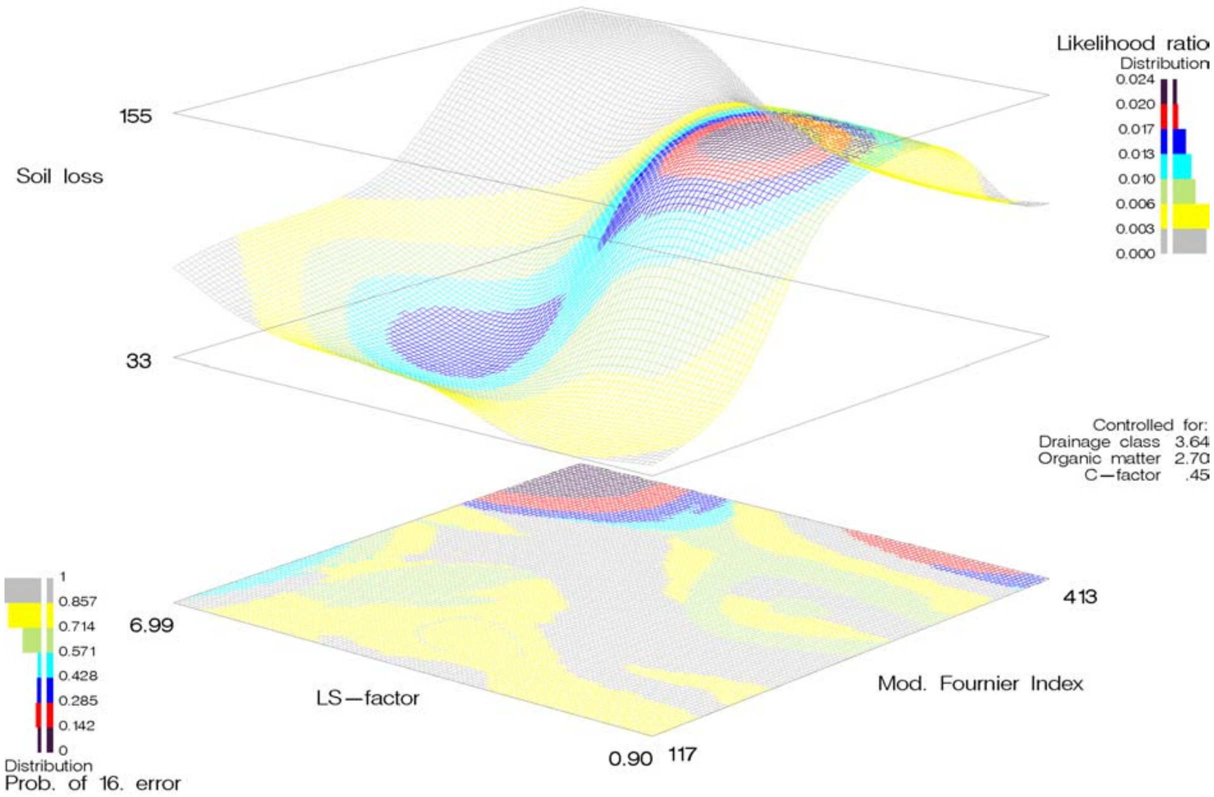

Figure 9. Soil loss values against MFI and LS-factor for annual crops. Covariates: likelihood ratio and probability of error

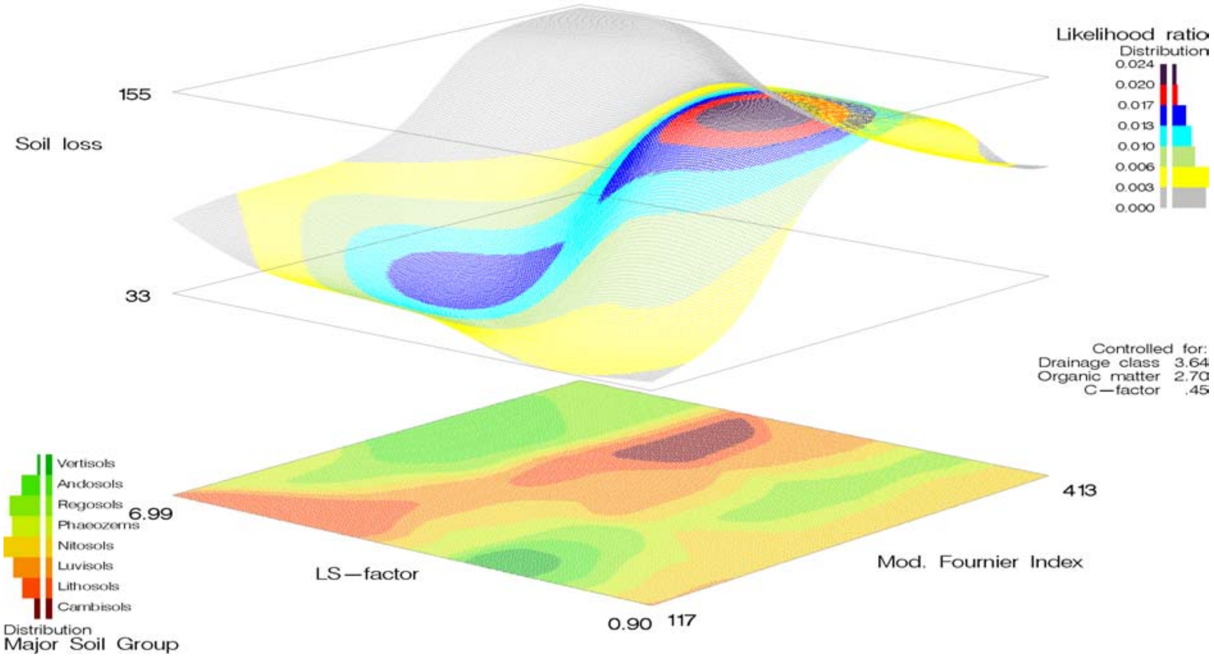

Figure 10. Annual soil loss values against MFI and LS-factor. Covariates: location of soils and likelihood ratio.

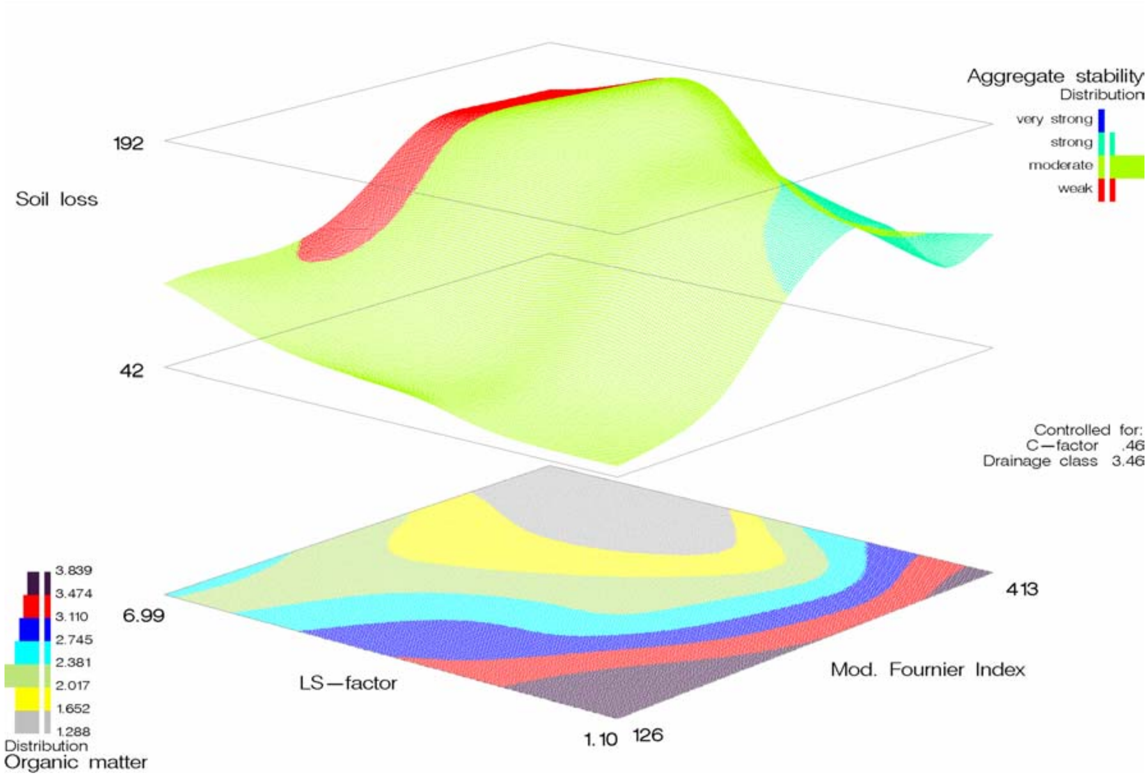

Figure 11. Annual soil loss values against MFI and LS-factor.
Covariates: aggregate stability and organic matter

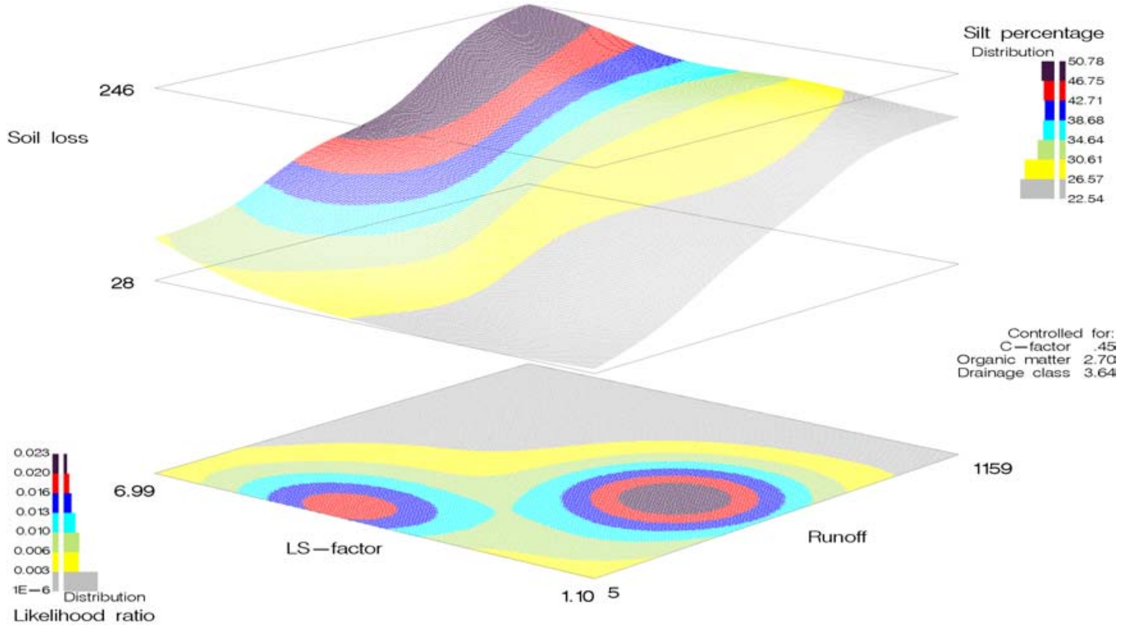

Figure 12. Annual soil loss values against annual run-off and LS-factor.
Covariates: silt percentage and likelihood ratio.

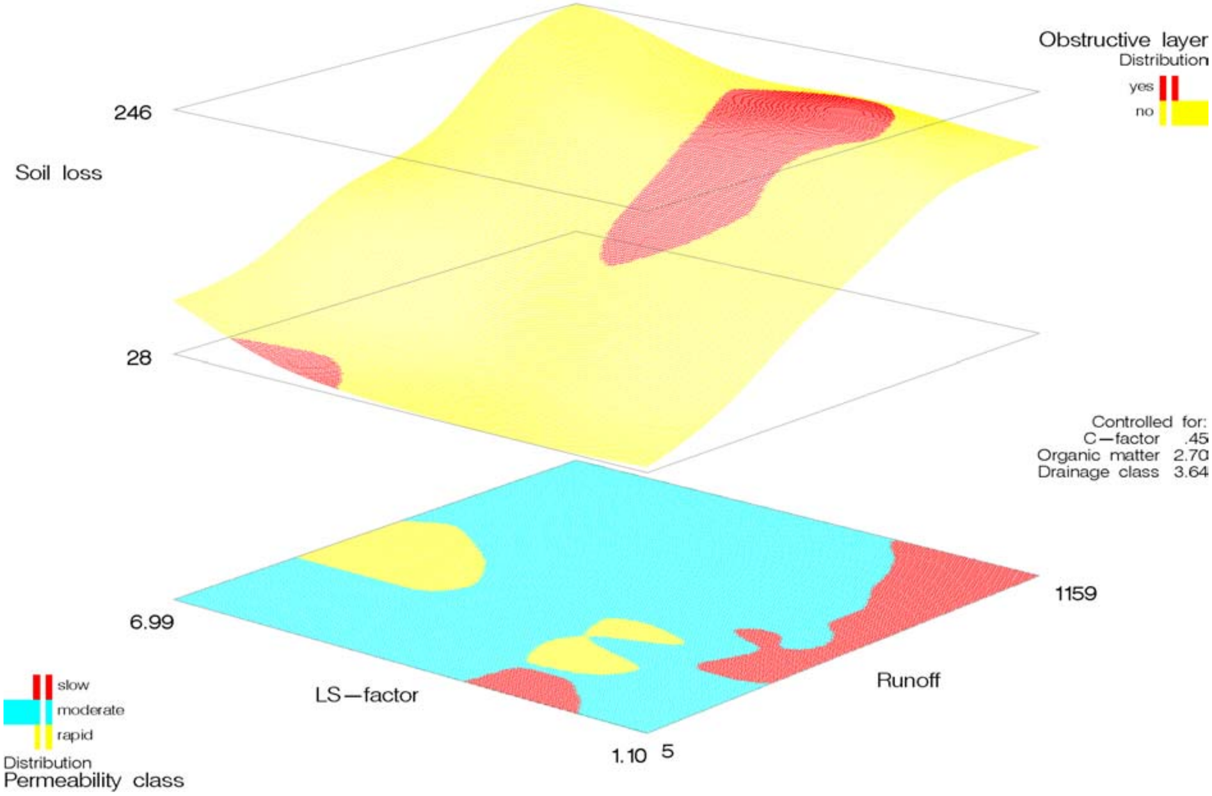

Figure 13. Annual soil loss values against annual run-off and LS-factor. Covariates: silt percentage and likelihood ratio.

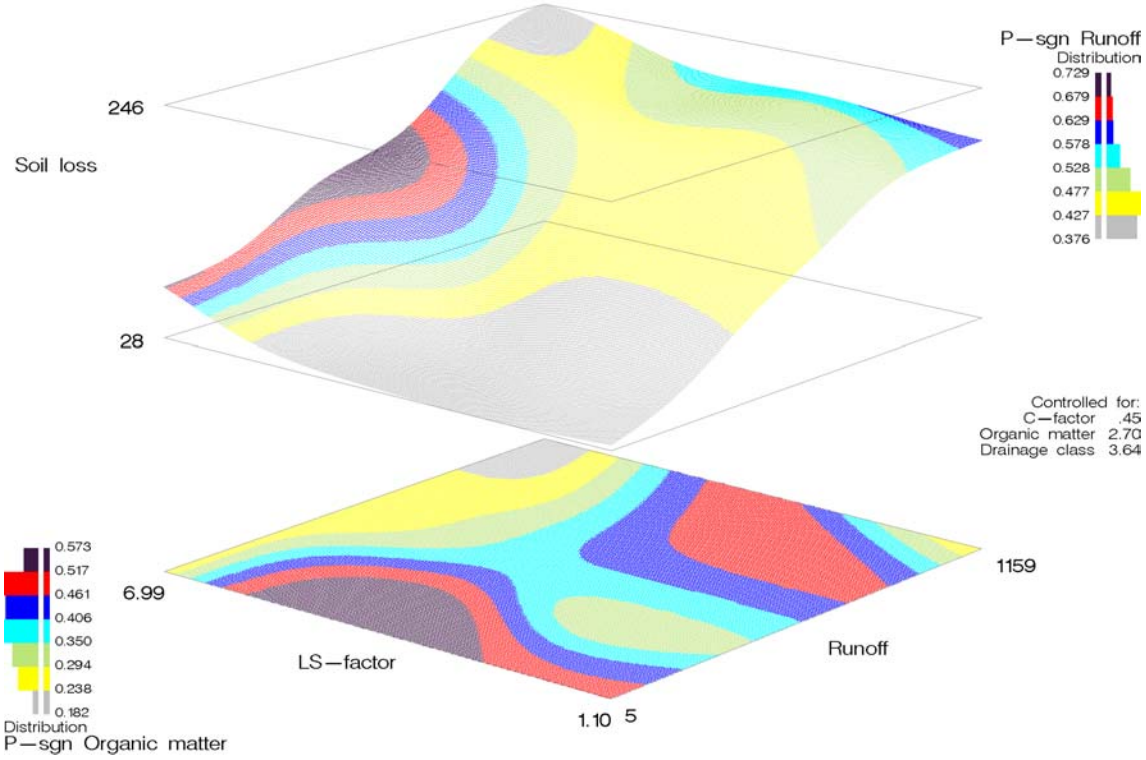

Figure 14. Annual soil loss values against annual run-off and LS-factor. Covariates: probability of wrong sign of 1st derivative for run-off and organic matter.

High percentages of silt makes the soil more erodible compared to soils with coarser of finer soil particles. The coarser (sand) are resistant to detachment because of their weight while the finest soil particles (clay) in combination with organic material withstand erosive forces because of their adhesive and chemical binding and formation of clods. Soils, which contain a lot of silt, like sandy loam or loamy sand textured, also have a greater tendency to seal. The fine silty particles block the pore spaces, obstruct water infiltration and elevate the runoff. Therefore, we introduce the silt fraction as a covariate in the surface plane in relation to the visual dimensions while the MFI is replaced by another hydrological component, the amount of annual runoff.

Fig. 12 shows an almost linear relationship between total annual runoff and soil loss for all slope ranges. The soil loss and runoff remains constant in the lower and middle slope range, and for the higher slope range the soil losses increases somewhat. The colour pattern of the silt content confirms its relation with soil erodibility. Soils with the highest silt content show the highest soil losses while the losses diminish gradually with the silt content.

Limited soil depth and drainage

Other soil factors that are likely to influence the runoff are limited soil depth and drainage. Soils with a limited depth have a restricted storage capacity and initiate overland flow earlier than deeper soils. We define in this study the soils with a limited depth (1) when they are classified as Lithosols, (2) soils that possess an Abrupt Textural Change (FAO, 1997) and (3) when they posses a Lithic or Petric phase within the upper 50 cm of the soils. Soil drainage in the database was given a qualitative classification (FAO, 1997) and later aggregated in three classes 'rapid' 'well' and 'poor'. Fig. 13 uses the same set of explicit and conditioning explanatory variables as the previous Fig. (12). It shows that only few soils in the sample possess an obstructive layer and that their correlation with the runoff is ambiguous. In addition, the qualitative classes for soil permeability show a weak correlation except for the highest runoff at the low slope ranges where high soil losses are recorded.

Reliability of slope direction

Next, we evaluate the reliability of the slope of the regression curve in Fig. 12 and 13 by plotting the probability of having a slope with an opposite sign as a covariate. We do this for the two factors: the runoff (surface plot) and organic matter content.

We notice in Fig. 14 that especially for the higher topography (LS) values the reliability of the slope sign of the runoff variable is low. The low reliability occurs around data points where the figure is somewhat bumpy and where it tends to descend. The reliability is much better elsewhere. For organic matter the slope sign has a higher reliability as can be seen by comparing the histogram on the left bottom with that of the upper right.

Runoff index for monthly precipitation

We now come to the final step of our exploration. As data on runoff are not commonly available, several procedures have been developed in the literature to estimate the runoff as a percentage of the rainfall. Here we calculate a runoff coefficient (CC) based on Cooks' method adjusted for African conditions (Hudson, 1986 p. 116), which only relies on readibly available data, i.e. on a broad categorization of land use types, soil type and drainage and slope. The CC was applied on monthly rainfall data and led to the following coefficient for yearly runoff (RI):

$$RI = \frac{\sum_{i=1}^{12} CC \times P_i^2}{\sum_{i=1}^{12} P_i}$$

where P_i is the monthly rainfall and the subscript, i denotes the month.

The results are shown in Fig. 15 where the RI is depicted as a covariate in the surface plot and the contour lines in the ground plane measure soil loss. The colour shift in the RI-classes appears to follow the contour lines on the surface plot except in its upper middle range. This suggests that this variable might be an appropriate predictor for the soil losses.

CONCLUSIONS

In this paper, we have applied non-parametric regression to conduct two separate exercises. The first is a quantitative interpretation of expert assessments that compares the qualitative but ordered classes of expert judgements with quantitative observations on soil losses. The second develops a functional form for estimating soil losses based on a limited set of data.

From the first exercise, we reveal a positive relationship between the erosion hazard assessment by the expert and the actual soil loss, though the reliability of this relationship becomes limited for higher classes, due to the wide range of observed soil losses. This possibly happens because experts tend to base their opinion on long term effects that would prevail under the existing conditions of rainfall, soil type, slope and land use, whereas annual soil losses might depend on a few showers in combination with a low soil coverage (Herweg and Stillhardt, 1999), which are not conveyed by the general data in the questionnaire. The analysis of the hit ratio shows that the experts give a reasonable assessment of the erosion risk hazard. It can even be classified as good if classes four and five are aggregated but experts tend to overestimate soil losses.

After a stepwise introduction of the mollifier methodology (Fig. 4-8), the second exercise proceeded in 5 steps. It was seen (Fig. 9) that soil loss should be modelled separately for annual crops and land use types with a more permanent coverage (grass and perennials). The MFI seems a better factor to represent the rainfall erosivity than the more advanced R-factor (Fig. 5, 6, 10), moreover it has the advantage that it can be composed from data that are readily available in Ethiopia. However, its surface plot shows several irregularities. Remarkably, the total annual runoff has an almost linear relation with annual soil loss (Fig. 12-14). The index derived from monthly rainfall data and the adjusted Cooks' method seems promising (Fig. 15) to represent the hydrological factor in the model and it is easily calculated with readily accessible data on monthly

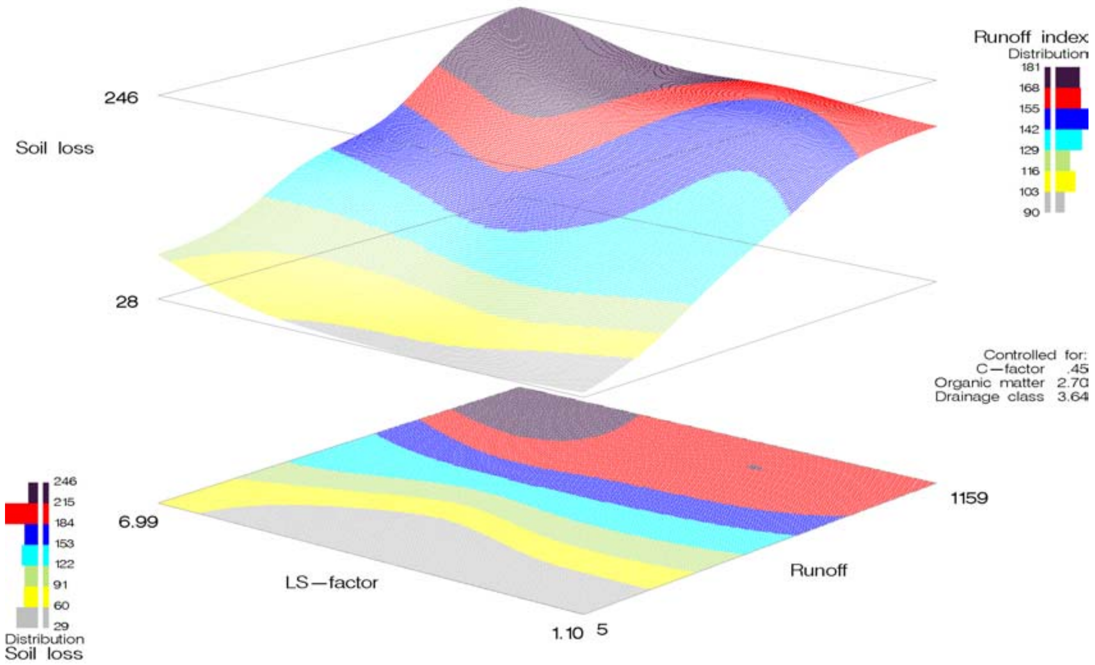

Figure 15 Annual soil loss values against annual run-off and LS-factor. Covariates: silt percentage and likelihood ratio

rainfall. The soil characteristics silt percentage and organic matter content showed (Fig. 11 and 12) a clear relationship with the estimated soil loss. We further noticed that observation densities around the highest values of the MFI, R-factor and runoff and the LS-factor are low and the visualized relationship in this area may therefore be less reliable. Also the poor ‘goodness of fit’ anticipates low correlation coefficients in future parametric models and indicates that additional variables should be included if a reliable model is to be obtained. This might particularly be the case for different land husbandry measures that were taken by the farmer and which are now included in a single C-factor. Another reason is the strong influence of extreme events in the erosion process that are not represented by the selected readily available data, which by definition excludes their high temporal resolution.

A disadvantage of the non-parametric method is that it is “weak on theory” in that the resulting regression curve is shaped according to the data and not according to imposed theoretical properties of functions. This may not always confirm the a priori’s of the modeller and experts. Therefore, the next step in this research will be to estimate a parametric model that uses (easily available) expert judgements and (scarce) real valued observations of soil loss as a dependent variable and a limited number of explanatory variables as independent variables.

ACKNOWLEDGEMENTS

The authors would like to thank Professor H. Hurni, dr. K. Herweg and B. Stillhardt M.Sc. of the Centre for Development and Environment (CDE), University of Berne, Switzerland, for allowing the use of the SCRP data set. The CDE also kindly collaborated in the distribution of the questionnaire. The comments of an anonymous referee are much appreciated.

REFERENCES

Arnoldus, H. 1980. An approximation of the Rainfall Factor in the Universal Soil Loss Equation. p 127-132. In: M. de Boodt and D. Gabriels (eds.). Assessment of Erosion. Wiley.

Bjorneberg, D.L., J.K. Aase and E.T.J. Trout. 1997. WEPP model erosion evaluation under furrow irrigation. American Society of Agricultural Engineers Paper No. 97-2115.

Bonari E, P. Barberi, M. Mazzoncini, N. Silvestri. 1996. Utilizzazione del modello "GLEAMS" per la simulazione del ruscellamento superficiale e dell'erosione da tecniche alternative di lavorazione del terreno nella collina toscana. Rivista di Agronomia 30:478-487.

Bierens, H.J. 1987. Kernel density estimations of regression functions. Advances in Econometrics 6, Cambridge University Press.

Desmet, P.J.J., G. Govers, D. Goosens. 1995. GIS-based simulation of erosion and deposition patterns In: J. Poesen and G. Govers (eds.). Experimental geomorphology and landscape ecosystem changes.

De Roo, A.P.J., C.G. Wesseling, N.H.D.T. Cremers, M.A. Verzandvoort, C.J. Ritsema and K. Oostindie. 1996.

LISEM-A physically based model to simulate run off and soil erosion in catchments: model structure. In: O. Slaymaker (ed.). Geomorphic hazards. John Wiley and sons Ltd.

Gachene, C.K.K. 1995. Evaluation and mapping of soil erosion susceptibility: an example from Kenya. Soil Use and Management. 11:1-4.

Greene, W.H. (1991) Econometric analysis. MacMillan, New York.

FAO (1977) Guidelines for soil profile description. FAO, Rome, Italy.

Hurni, H. 1993. Land degradation, famines and resource scenarios in Ethiopia. In World Soil Erosion and Conservation, ed. D. Pimentel, pp. 27-62. Cambridge University Press, Cambridge.

Hudson, N. 1986. Soil Conservation. B T Batsford Limited. London. U.K.

Herweg, K and B. Stillhardt. 1999. The variability of soil erosion in the Highlands of Ethiopian and Eritrea. Research Report 42. Centre for development and Environment. University of Berne.

Keyzer, M.A. 1996. Estimation of real-valued models form discrete and limited dependent observations: a programming approach. In: Proceedings of the Annual Conference of the Indian Econometric Society (March 21-23, 1996). Indian Statistical Institute.

Keyzer, M.A. and B.G.J.S. Sonneveld 1998. Using the mollifier method to characterize datasets and models: the case of the Universal Soil Loss Equation. ITC Journal 1997-3/4:263-272.

King, D., D.M. Fox, J. Daroussin, Y. le Bissonnais, and V. Danneels. 1998. Upscaling a simple erosion model from small areas to a large region. In: Soil and water quality at different scales. In: P.A. Finke and J. Bouma (eds.). Proceedings of workshop, Wageningen, Netherlands, 7-9 August, 1996. Nutrient Cycling in Agroecosystems. 50:1-3:143-149.

Klik A, B. Hebel, A. Zartl, and J. Rosner. 1997. Measured vs. WEPP simulated runoff and erosion from differently tilled plots. American Society of Agricultural Engineers Paper No. 97-2120.

Lal, R. 1990. Soil Erosion in the Tropics. Principles and Management. New York: McGraw-Hill Inc.

Lal, R. 1995. Sustainable management of soil resources in the humid tropics. United Nations University Press, Tokyo.

Littleboy, M., A.L. Cogle, G.D. Smith, D.F. Yule, and K.P.C. Rao. 1996. Soil management and production of Alfisols in the semi-arid tropics. I. Modelling the effects of soil management on runoff and erosion. Australian Journal of Soil Research. 34:

Renard, K.G., G.R. Foster, G.A. Weesies, D.K. McCool, and D.C. Yoder. Predicting soil erosion by water: A guide to conservation planning with the Revised Universal Soil Loss Equation (RUSLE). Agriculture Handbook No. 703. USDA-ARS.

Morgan, R.P.C., J.N. Quentin and R.J. Rickson. 1992. EUROSEM: Documentation Manual. Silsoe College, Silsoe, U.K.

Morgan R.P.C. 1995. Soil erosion and conservation. Longman Group Ltd.

Nearing, M.A., G.R. Foster, L.J. Lane and S.C. Finkner. 1989. A process based model for USDA Water Erosion Prediction Project (WEPP) technology. Trans. ASAE 32: 1587-1593.

Nearing, M.A. 1997. A single continuous function for slope steepness influence on soil loss. Soil Science Soc. Am. J. 61: 917-919.

Quinton J.N. 1997. Reducing predictive uncertainty in model simulations: a comparison of two methods using the European Soil Erosion Model (EUROSEM). Catena 30:101-117.

Silverman, B.W. 1986. Density Estimation for Statistics and Data Analysis, Chapman and Hall

Sonneveld B.G.J.S. and P.J. Albersen. 1999. Water erosion assessment based on expert knowledge and limited information using an ordered logit model. J. Soil Water Conserv. 50:592-599.

World Bank. 1998. African Development Indicators 1998/1999. World Bank Washington, D.C., USA.

Yu, B., C.W. Rose, K.J. Coughland, and B. Fentie. 1997. Plot scale Runoff modelling for Soil Loss prediction: In. A new soil conservation methodology and application to cropping systems in tropical steeplands (eds. K.J. Coughland and C.W. Rose). ACIAR Technical Reports 40. Canberra, 1997.

Annex 1

Further background on the mollifier

Let us start the explanation of the mollifier method by considering a given data set S of real-valued observations indexed s, and partition it into a vector of a vector of n (bounded) endogenous variables y^S and a vector of m exogenous variables x^S from the bounded set X. The mollifier calculates a value y(x) at intermediate points x, thus creating a blanket that fills the gaps between the observations. The mollifier uses for its estimation a weighting function $w^S(x)$ that equals the probability P^S of y^S being the correct value of y(x). This means that errors have to be accounted for and relaxes the requirement of conventional interpolation methods to let the curve pass through the observations. The resulting specification will be:

$$\tilde{y}(x) = \Sigma \; y^s P_s(x) \tag{1}$$

This defines a non-parametric regression function, whose shape will depend on the postulated form of the probability function. For example, if y^S is a scalar and x^S a two-dimensional vector of ground co-ordinates, every observation s can be viewed as a pole of height y^S located at point x^S. The regression curve lays a “soft blanket” on these poles that absorbs the peaks of the highest poles (upward outliers) and remains above the lowest poles. The analytical form of the probability function $P^S(x)$ of this model can be obtained in various ways. Here we will apply the mollifier approach.

For a finite sample of size S, the value of this mollifier function (1) can be estimated by a Nadaraya-Watson estimate i.e. a weighted sample mean with window size θ as parameter:

$$\tilde{y}(x) = \sum_s y^s P_s(x) \tag{2a}$$

for

$$P_s(x) = \psi(x^s - x)/\theta \big/ \Psi^s(x) \text{ if } \Psi^s(x) > 0 \text{ and 0 otherwise} \tag{2b}$$

where

$$\Psi^s(x) = \sum_{s=1}^{S} \psi((x^s - x)/\theta) \tag{2c}$$

and where the density function $\psi(\varepsilon;\, \theta)$ has its mode at $\varepsilon=0$ and is such that for θ going to zero its support goes to zero.

In this approach, expression $\psi(x^s - x)$ in (2c) can be interpreted as the likelihood of x being associated to the observation s and $\Psi^S(x)$ the likelihood of x being associated to any of the observations in the sample. Hence, probability $P_s(x)$ is the probability of x being associated to observation s, conditional on its association to at least one observation in the sample and $\tilde{y}^s(x)$ is the expectation of the y^s-values associated with the sample. We also define the likelihood ratio

$$\Lambda(x) = \sum_{s=1}^{S} \psi((x^s - x)/\theta) \bigg/ \sum_{s=1}^{S} \psi(0) \tag{3}$$

as well as the probability $Q(x; a)$ of y falling outside a given range $a = \alpha\bar{y}$ around $\Psi^s(x)$, where $\bar{y}$ is the sample average:

$$Q(x;a) = \sum_{s \in S(x;a)} P_s(x) \text{, for } S(x;a) = \left\{ s \middle\| y^s - \tilde{y}(x \right| \geq a \right\} \tag{4}$$

This probability serves as a measure of fit.

The mollifier program also assesses the partial derivative of the regression curve as well as a measure of its reliability. For this, it calculates the first partial derivative to x_k at point x, where k represents an explanatory variable, at all data points.

$$\frac{\partial\tilde{y}(x^t)}{\partial x_k} = \sum_s \frac{\partial P_s(x^t)}{\partial x_k} y^s \tag{5}$$

since $\sum_s \frac{\partial P_s(x^t)}{\partial x} = 0$ we can write

$$\frac{\partial\tilde{y}(x^t)}{\partial x_k} = \sum_s \frac{\partial P_s(x^t)}{\partial x_k} (y^s - y^t) \tag{6}$$

where y^t refers to the t^{th} observation. As by definition, $\frac{\partial P_s(x^t)}{\partial x_k} = P_s \frac{\partial \ln P_s(x^t)}{\partial x_k}$, it follows that

$$\frac{\partial\tilde{y}(x^t)}{\partial x_k} = \sum_s P_s(x^t) \left[\frac{\partial \ln P_s(x^t)}{\partial x_k} (y^s - y^t) \right]. \tag{7}$$

Let us now rewrite and interpret the term in square brackets.

$$\frac{\partial \ln P_s(x^t)}{\partial x_k} = \frac{\partial \ln \psi_s(x^t)}{\partial x_k} - \sum_{h=1}^{S} P_s(x^t) \frac{\partial \ln \psi_s(x^t)}{\partial x_k} \tag{8}$$

Now for a density $\psi_s(x^t) = \psi \frac{(x^s - x^t)}{\theta}$ where ψ is a normal joint density with diagonal variance matrix and variance σ_k^2 around x^t it follows that

$$\frac{\partial \ln \psi_s(x^t)}{\partial x_k} = \frac{x_k^s - x_k^s}{\sigma_k^2} \tag{9}$$

Hence the term in square brackets can be rewritten as

$$\frac{\partial\tilde{y}(x^t)}{\partial x_k} = \sum_s P_s(x^t) \left[\xi_k^s \delta_k^s \right] \tag{10}$$

where $\delta_k^s = (y^s - y^t)$ and

$$\xi_k^s = \frac{x_k^s - x_k^t}{\sigma_k^2} - \sum_h P_h(x^t) \frac{(x_k^h - x_k^t)}{\sigma_k^2}.$$

In other words, the term in square brackets is the contribution of observation s to the slope.

For given x^t this enables us to define the probability of a positive sign for the slope as

$$P^+(x^t) = \sum_S P_s(x^t \,|\, \xi_k^s \delta_k^s \geq 0)$$

Hence the probability of a wrong sign can be calculated as

$$P^\#(x^t) = P^+(x^t) \text{ if } \frac{\partial\tilde{y}(x^t)}{\partial x} < 0 \text{, and}$$

$$1 - P^+(x^t) \text{, if } \frac{\partial\tilde{y}(x^t)}{\partial x} \geq 0$$