Using the mollifier method to characterize datasets and models: the case of the universal soil loss equation

Michiel A Keyzer[1] and Ben G J S Sonneveld[1]

ABSTRACT

The mollifier method is a numerical technique that has been applied widely in physics and chemistry to approximate mathematical functions of an irregular shape. In this paper, we propose to use it for non-parametric interpolation to characterize dataset and models. The basic idea is to lay a "soft blanket" over a profile of points generated by an empirical dataset or by a mathematical model. If the blanket is pulled "tightly", its surface will smoothen most of the irregularities of the underlying profile. This amounts to averaging out information over a wider window size (neighbourhood) around every point. The resulting mollifier estimate (blanket) can be given a three-dimensional graphical representation with a dependent variable mapped out against two independent variables, for fixed values of other independent variables. Compared with parametric methods such as spline regression or variorum estimation, the mollifier method has the important advantage that it gives a measure of statistical reliability at every point and that this measure does not depend on the fit at other points. This makes it possible to create a picture of the blanket that has a higher resolution (through a deeper colour or a darker shading) in regions where the mapping is more reliable. In addition, the introduction of a plane in the 3D presentation is used to depict the goodness of fit. As an illustration, we apply this technique to the dataset of the universal soil loss equation (USLE), the USLE itself, and the deviations between the two. The resulting graphs show the estimated values of annual soil loss, as well as the error of the model mapped against two explanatory variables of the USLE and for average values of non-visible exogenous variables. The following properties stand out: (1) The data are heavily concentrated in certain areas of the graphs, where the probability of a correct estimate is relatively high and the model error low. Applying the USLE here is relatively reliable. (2) Unexpected trends in the shape of the regression curve are shown, whereas the mapping that is based on data is perfectly foggy in this range.

Attempts at integrating biophysical and socio-economic analysis have often failed in the past. A major reason has been that it is not possible to achieve meaningful integration by studying biophysical processes in one box, or subsystem, and socio-economic processes in another, with a limited number of interrelationships linking the two subsystems. If this were possible, it would be relatively easy to demarcate fields of expertise, with a separate module for every field, and to develop the integrated system as little more than the sum total of its components. Unfortunately, the economic analysis cannot be conducted within a separate box. It is necessary to incorporate technologic relationships within the decision models of economic actors, usually the farmer or the government. Thus the economist will invite experts in biophysical sciences to supply technologic information. However, these attempts have faced severe difficulties in meeting the standards of the various professions involved. Technologic relationships are either insufficiently formalized and tested statistically, which makes them unacceptable to economists, or too distant from the underlying process models to be acceptable to agronomists and soil scientists. Moreover, the spatial detail is generally limited, which causes geographers to question the approach.

The present paper argues that combining non-parametric interpolation methods from statistics with the visual tools of cartography might help bridge the gap between these professions. The proposed approach is close to cartography in that it primarily relies on colourful visual display, but it also meets the demands of econometricians in that it uses an explicit mathematical formalization that has a clear statistical interpretation. Finally, it should appeal to representatives of the biophysical sciences because it allows the data to be closely followed without imposing the straitjacket of any functional form.

The paper describes the non-parametric model in an informal manner, relegating the technical discussion to the Appendix. We apply the proposed method to the dataset on erosion that was compiled by the National Soil Erosion Research Laboratory in West Lafayette, Indiana, to calibrate the universal soil loss equation (USLE). We will show the type of functional form that is being suggested by the data and comment on the pattern of deviations that exist between these data and the calibrated USLE equation.

MODELS OF NON-PARAMETRIC INTERPOLATION

SPATIAL INTERPOLATION

Geographic information systems (GIS) contain data about a given set of geographic locations. These data refer to particular variables generated from empirical observations or from particular calculations based on simple rules or complex process models. The GIS apply various kinds of spatial interpolation techniques to fill the gaps in space between the observations. This enables them to generate colourful maps in two or three dimensions.

For example, consider a given dataset S of real-valued observations indexed s, and partition it into a vector of n (bounded) endogenous variables y^s from the bounded set X. In a GIS, the vector will usually stand for the geographic coordinates latitude and longitude. The aim of spatial interpolation techniques is to calculate a value y(x) at intermediate points x, thus creating a blanket that fills the gaps between the observations. Interpolation methods usually define a weighting function $w^s(x)$ and compute:

$$y(x) = \Sigma_{s=1}^{S} y^s w^s(x) \qquad [1]$$

where the weighting function $w^s(x)$ is (1) continuous, (2)

[1] Centre for World Food Studies of the Free University (SOW-VU) Amsterdam, The Netherlands

non-negative, (3) summing to unity: $\Sigma_{s=1}^{S} w^s(x) = 1$ and (4) passing through the observations: $y(x) = y^s$. Clearly, the simplest form for the weighting function would be the arithmetic average $w^s(x) = 1/S$.

NON-PARAMETRIC INTERPOLATION MODELS

In principle, these three-dimensional maps describe a particular relief through a function that maps out the altitude as a function of the geographic coordinates, but a good relief map does more. It also describes the pattern of other variables through variations in colour, shading, patterns, etc. It is this capacity to represent more than three variables in a three-dimensional plot that we will seek to exploit. Clearly, these ideas readily extend to applications where the vector has a larger dimension, and also includes causal factors such as hill slope or rainfall. However, since it is not possible to visualize the function y(x) if the dimension of x exceeds two, we will have to keep m -2 components of x fixed at controlled values, say at the mean value. As long as the pair of x-variables that are not fixed represent longitude and latitude, we remain within the context of a GIS and depict the spatial variability of y. As soon as other variables are shown on the axes, the diagram shows relationships between other factors, and depicts technologic functions. However, GIS methods have the analytic limitation that the weighting functions are generally hidden in procedures over which the user only has limited control.

THE MOLLIFIER AS A STATISTICAL APPROACH TO NON-PARAMETRIC INTERPOLATION

Requiring the interpolated value to pass through the observations makes the procedure mechanical in the sense that it rules out the possibility of random errors. Alternatively, interpolation can be viewed as the calculation of an expected value in the statistical sense. The weighting function $w^s(x)$ will then be equal to the probability P^s of y^s being the correct value of y(x), but this means that errors have to be accounted for and hence that it is natural to give up the requirement that the interpolation curve should pass through the observation. The resulting specification will be:

$$y(x) = \Sigma_{s=1}^{S} y^s P^s(x) \qquad [2]$$

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 coordinates, every observation s can be viewed as a pole of height y^s located at point x^s. The regression curlays a “soft blanket” on these poles that absorbs the peaks of the highest poles (upward outliers) and remains above the lowest poles. The mollifier uses a control variable to vary the band (or window) width of the neighbourhood of x whose points affect the prediction of y. If the averaging emphasizes nearby points, the probability function is said to use a small window size. The larger the window size, the tighter the blanket. The analytic form of the probability function $P^s(x)$ of this model can be obtained in various ways. We will apply the mollifier approach, which is given in some detail in the Appendix. Here we only give an intuitive description of this approach.

Suppose that an aeroplane conducts surveillance flights at varying altitudes above a given territory. At randomly chosen time intervals s = 1, …, S, it makes perfectly accurate measurements y^s of outside conditions, such as temperature or atmospheric pressure. However, the aeroplane’s equipment is unable to measure accurately the spatial coordinates x^s (altitude, latitude, longitude) of the measurement. What is recorded as a measurement at x^s is, with likelihood $\psi(\varepsilon)$, a measurement at $x^s - \varepsilon$, where the likelihood function is known a priori (it depends on the characteristics of the equipment, not on the external conditions and not on the local conditions). Consequently, to any given point measurement y^s corresponds a continuum of values $x = x^s - \varepsilon$, and conversely we can for every given value x compute (1) the likelihood $\psi(x^s - x)$ of x being the value actually associated with y^s, (2) the likelihood $\Psi(x) = \Sigma_s \psi(x^s - x)$ of x being associated with any of the observations y^s, in addition we can define the likelihood ratio $\Psi(x)/\Psi_0$, for $\Psi_0 = S\ \psi(0)$ to obtain a measure of the observation density ranging between zero and unity and, finally, (3) the conditional likelihood, *ie*, ratio $\psi(x^s - x)/\Psi(x)$, which is the probability $P^s(x)$, from which we can calculate the expected value of y at x, *ie*, the regression function y(x). We notice that had there been no error in measurement, it would not have been possible to make any inference for intermediate points other than x^s. Thus, the randomization with a known distribution $\psi(\varepsilon)$ makes it possible to make the transition from a point value to a function.

Clearly, the resulting estimates depend on the accuracy of the aeroplane’s equipment. We assumed the likelihood distribution of the measurement error to be a known a priori. Thus, this non-parametric regression method replaces the assumption about the parametric form to be known by one where the likelihood density is known. The theoretical justification is that as the number of observations is increased and the window size reduced, the interpolated value will approach the true model, independently of the assumed form of this likelihood function (see Appendix).

Compared with parametric methods such as spline regression or variogram estimation, the mollifier method has the important advantage that it gives a measure of statistical reliability at every point and that it does not depend on the fit at other points. Indeed, for every point x, we calculate the expected value, the likelihood ratio to measure the availability of observations, and as a measure of fit, a probability $Q_\alpha(x)$ of y falling within an α-percent range around y(x). These statistics will be depicted in the same diagram through a colouring or a shading of surface plots

A NON-PARAMETRIC INVESTIGATION OF THE UNIVERSAL SOIL LOSS EQUATION (USLE) DATASET

The dataset that we use for the analysis is a representative sample from the original USLE dataset. Risse *et al* [12] compiled the data to evaluate the model efficiency of the USLE. It comprises 1704 observations on annual soil loss, collected from 208 natural runoff plots at 22 experimental stations in the United States in the period 1930 to 1980. Each variable of the USLE equation is determined on the sites.

THE USLE

The USLE [16] is the most widely used water erosion

model to predict erosion hazards and to obtain plans for soil conservation measures. The mathematical form of the USLE is a multiplication of six factors that gives an assessment of soil loss:

$$A = RxKxLSxCxP \qquad [3]$$

where A is soil loss expressed in ton per ha per year; R is the rainfall erosivity, representing the destructive impact of raindrops on soil aggregates, and the runoff that serves as a transport module for soil particles; K is the soil erodibility which characterizes the vulnerability of soil types to water erosion; LS are the topographic factors, which can be subdivided into slope gradient and slope length, and which affect the volume and erosion causative speed of the runoff; C stands for soil coverage, the spatial and temporal distribution of leaf area and organic matter on top of the soil; finally, P is a value for soil conservation measures.

The USLE is popular for its easy applicability and low data requirements as compared with process-based models such as WEPP and EUROSEM. In 1978, its developers, Wischmeier and Smith, wrote the Agricultural Handbook 537, in which USLE factors can be derived from tables, charts and nomographs. Since then, the USLE has been applied in numerous studies to assess erosion hazard of the land. However, the equation has several shortcomings.

(1) The USLE is basically a statistical model, which makes it heavily dependent on the dataset that was used for its calibration. This is shown in the numerous studies where individual factors are recalibrated when the USLE is applied outside its calibration domain. For example, topography values were recalibrated by Hurni [6], Liu *et al* [7] and McCool *et al* [8]; and rainfall erosivity indexes adjusted for humid tropical areas were given by Hudson [5], Arnoldus [1] and Hargreaves [4]. A soil erodibility index for tropical soils is found in Vanelslande *et al* [14].

(2) It has a low (Nash-Suttcliff) model fit for annual soil loss estimations ($r^2 = 0.57$), although it leads to better results for long-term estimates ($r^2 = 0.75$) [12].

(3) The USLE consistently overestimates small soil losses and underestimates high soil losses (Nearing, in prep).

(4) A further disadvantage is the functional form, the multiplication of six factors, which easily leads to large error propagation if one or more factors are misspecified [15].

APPLICATION OF THE MOLLIFIER

The high dependence on data for its calibration and the poor fit make it worthwhile to apply the mollifier method to the original USLE dataset to review relationships between observed soil loss, the model error and explanatory variables. In this analysis, we will exhibit these relationships in 3D diagrams, where the non-parametric estimate (regression curve) of the endogenous variable is mapped out against two exogenous variables on a 50 x 50 grid, while the non-visible variables are conditioned on their mean. The 3D diagrams show the estimated values of the endogenous variable as surface plots (the blanket). The colours in the surface plot present the likelihood ratio of the observation density. The colours in the plane below the surface plot project the probability of y falling within the interval around y(x), the upper and lower bounds of which are given by 10 percent of the sample mean $\bar{y}$. Alternative window sizes are used to determine the width of the neighbourhood around x that affect the estimation of y.

We will consider the regression curves of three relationships. The first curve shows estimated values of soil loss mapped against rainfall erosivity and topography. We use it for a stepwise introduction of the regression curve with its statistical characteristics. The second regression curve maps estimates of soil loss against soil erodibility and soil coverage. Lastly, we map the estimated model error of the USLE against the rainfall erosivity and topography.

SOIL LOSS vs RAINFALL EROSIVITY AND TOPOGRAPHY

We will start our exercise with a scatter plot (Figure 1) of soil loss observations (z-axis) against rainfall erosivity (x-axis) and topography (y-axis). The scatter plot confronts us immediately with the limitations of this type of presentation: it is difficult to observe any clear relationship from these discrete data and we are not able to show the influence of other explanatory variables, soil erodibility, soil coverage, and the protection factor.

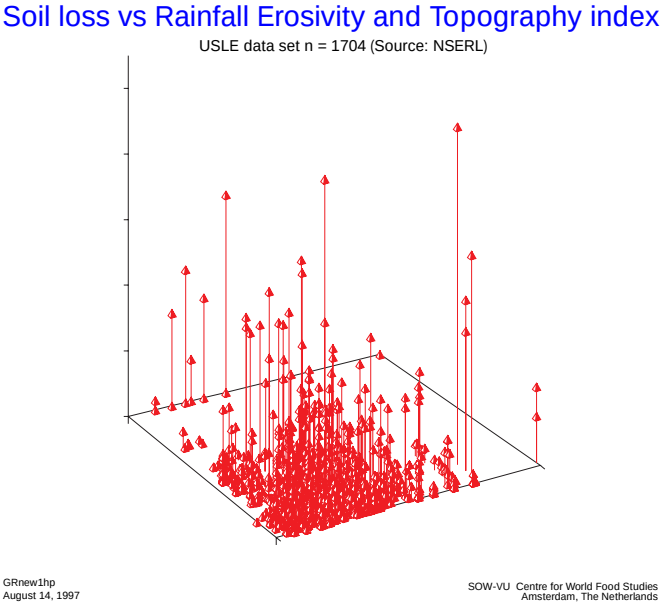

FIGURE 1

Figure 2 shows the graphical presentation of non-parametric estimates of soil loss mapped against rainfall erosivity and soil erodibility, while the exogenous variables soil erodibility, soil coverage and protection factor are conditioned on their mean.

In Figure 3, we exhibit the measure of observation density (the likelihood of the point x of the grid being associated with the observations by means of different shades on the surface plot. The shading goes from a dark (high observation density) to a light (low observation density) grey.

The probabilities of an accurate estimate are shown in the different colours of the plane in Figure 4. The legend bar on the lower left side of the figure indicates the colour shift from low to high probability. The colours for the observation density are shown in the surface plot and the corresponding legend appears at the upper right

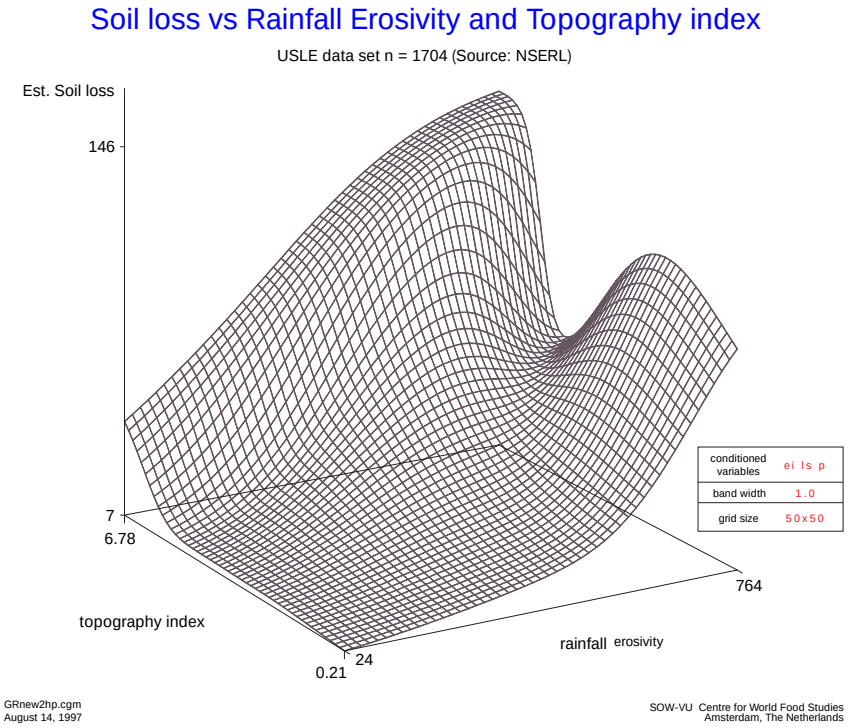

FIGURE 2

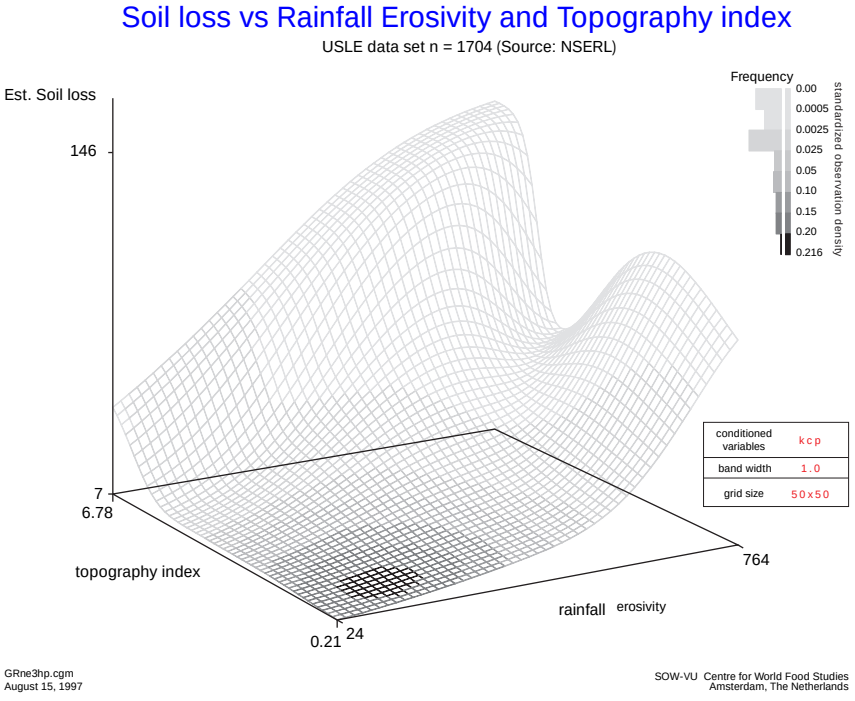

FIGURE 3

side of the graph. The values of the class boundaries appear next to the legends. The vertical histograms indicate the area distribution of the legend class in the surface plot and plane.

Figure 5 shows the estimates of soil loss (including observations density and probabilities of correct estimate) under a band width of 0.5. The reduced band width loosens the “blanket” of the mollifier, thereby making it adhere more closely to the original spikes formed by the data.

In Figure 6, the band width is reduced to 0.1 and this leads to an increasingly bumpy surface, as expected.

The regression curve in Figure 4 shows a slow linear increase in estimated soil loss for lower and middle values of rainfall erosivity and topography. For higher values, the soil loss increases exponentially but flattens out for their extremes. An unexpected “dip” is shown near the highest values of rainfall and topography, whereas the observation density is at its lowest, making this area of the graph extremely sensitive to the few observations in its vicinity. The observation density shown in Figure 3 (and 4) is heavily concentrated around the lower values of rainfall and topography, and drops rapidly for their middle and higher values. For a band width of 1,

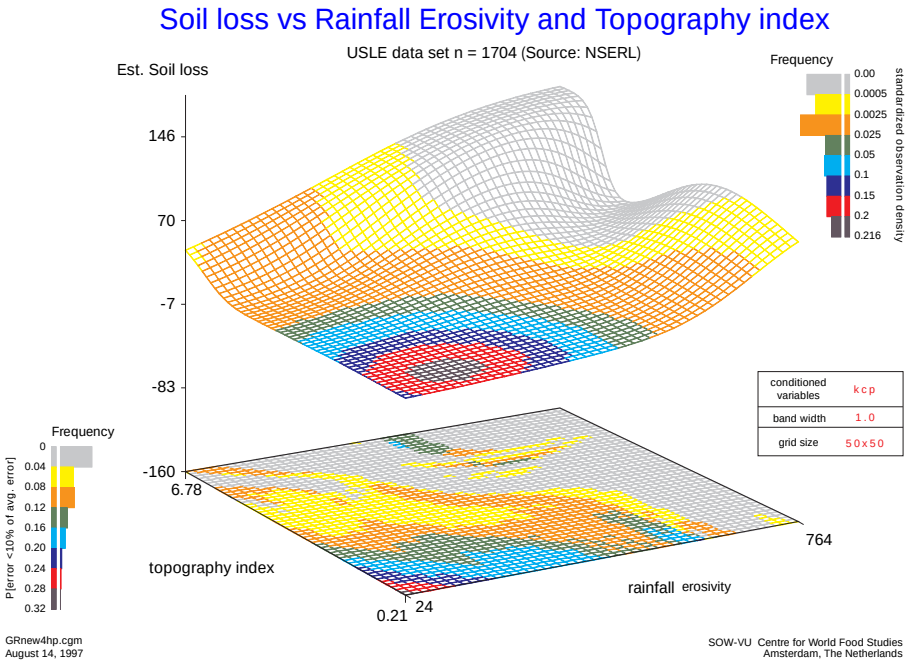

FIGURE 4

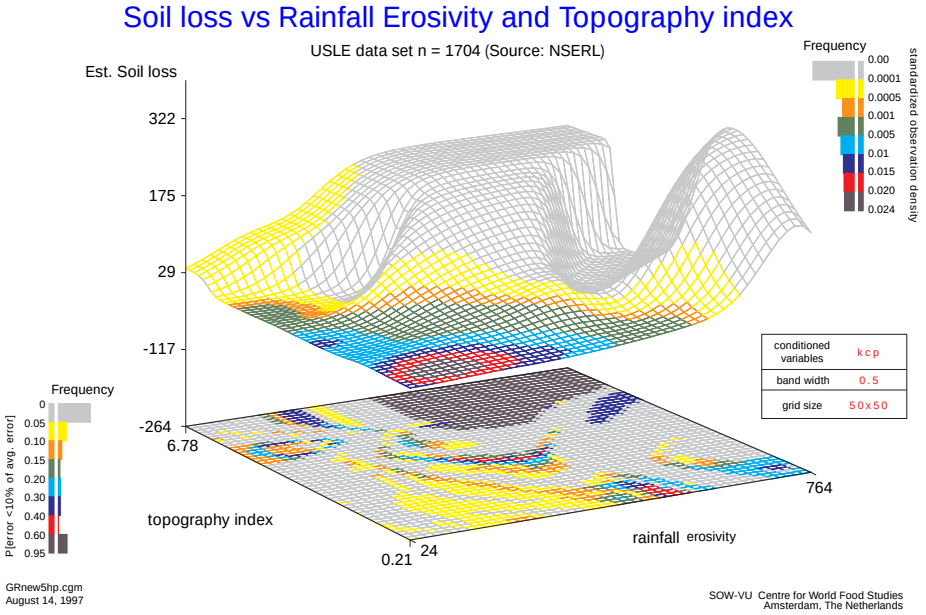

FIGURE 5

the probability of a correct estimate increases for smaller values of rainfall erosivity and topography, but has a rather low average value (6.2 percent). This indicates the limited capacity of the exogenous variables in explaining annual soil loss values. For reduced band widths of 0.5 (and 0.1), the probability increases in areas of low observation density because of the limited variability among the few observations that determine their value. The areas with a higher observation density show a greater variability and, consequently, a lower relative probability of an accurate estimate.

SOIL LOSS vs SOIL ERODIBILITY AND SOIL COVERAGE

Figure 7 shows the estimated values of soil loss mapped against the soil coverage factor and soil erodibility with fixed mean values for rainfall erosivity, topography and protection factor. Similar to the previous figures, the colours in the surface plot depict the observation density and the colours in the plane show the probability of a correct estimate; the band width is put at 1.

Figure 8 shows the same relationships with a band width of 0.5.

Both regression curves (Figures 7 and 8) give small soil losses for low values of soil erodibility and soil coverage factor. The soil loss remains low even for high values of the soil coverage factor in combination with low values of soil erodibility. This indicates that even bare soil can withstand soil detachment if it contains sufficient erosion-resistant properties. Soil loss increases exponentially for middle values of erodibility and soil coverage factor and reaches its maximum in a plateau.

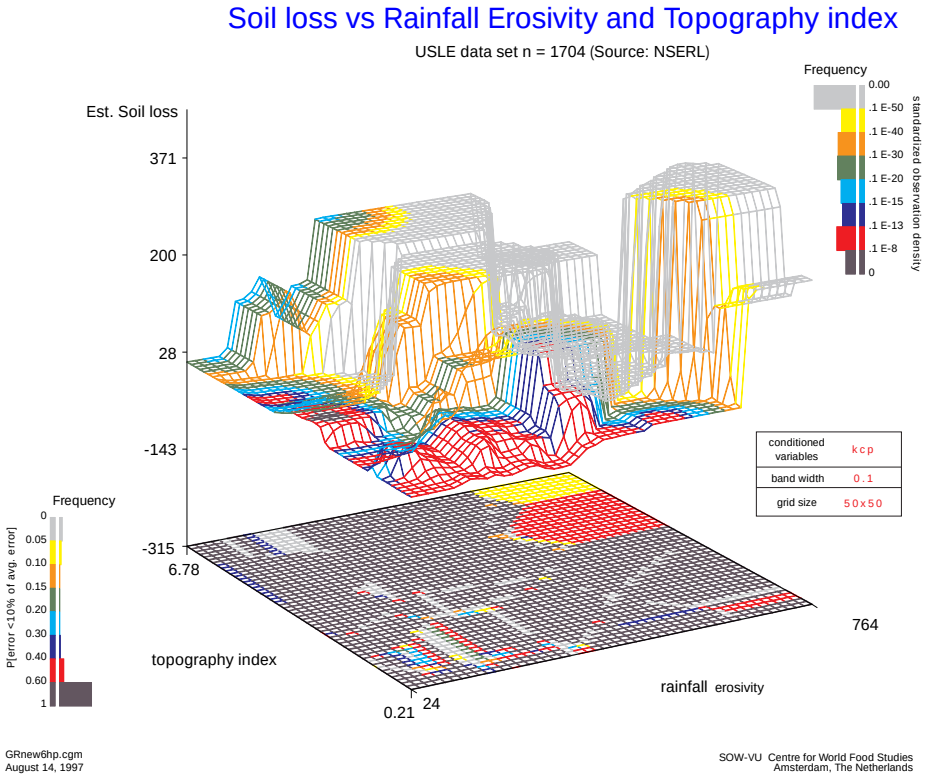

FIGURE 6

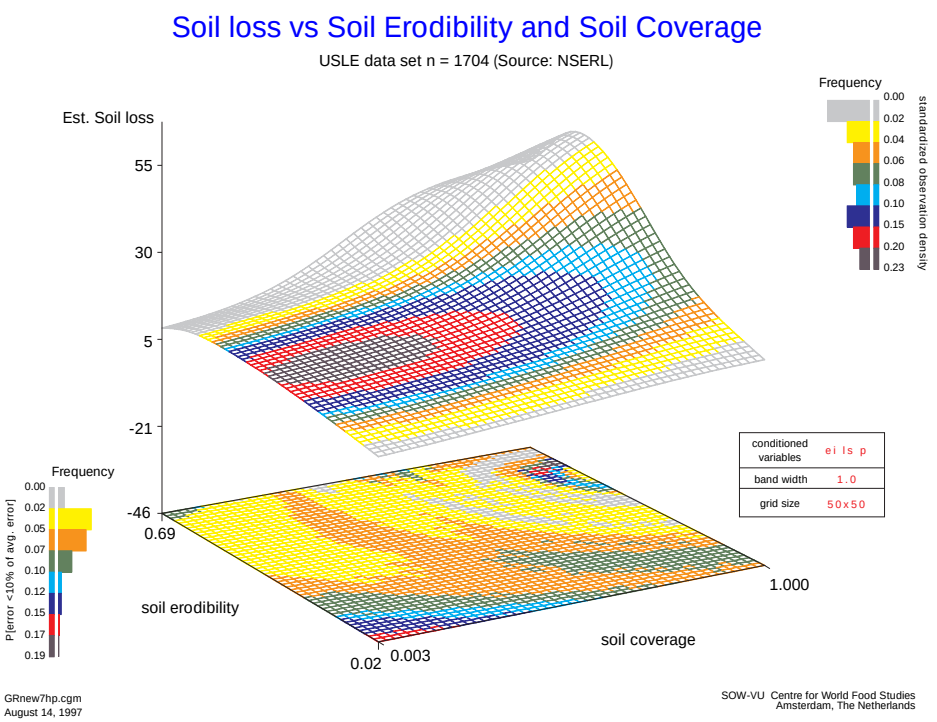

FIGURE 7

The regression curve shows, surprisingly, a decrease in soil loss for the highest erodibility values when compared with its middle values. The observation density of the coverage factor is well represented for its lower and highest values and somewhat less for the middle to higher range. The erodibility factor has few observations for its extremely low and high values. The probability of the correct estimate at a band width of 1.0 has a low average value (5.6 percent), confirming the low explanatory power of exogenous variables in accounting for annual soil loss. For a reduced band width (Figure 7), the probability increases considerably, especially in areas with few observations, but follows the earlier pattern.

MODEL ERROR, RAINFALL EROSIVITY AND TOPOGRAPHY

Figure 9 represents the estimated model error (soil loss observations - USLE estimations) against rainfall erosivity and topography factor. For the estimation, the soil erodibility, coverage factor and protection factor are

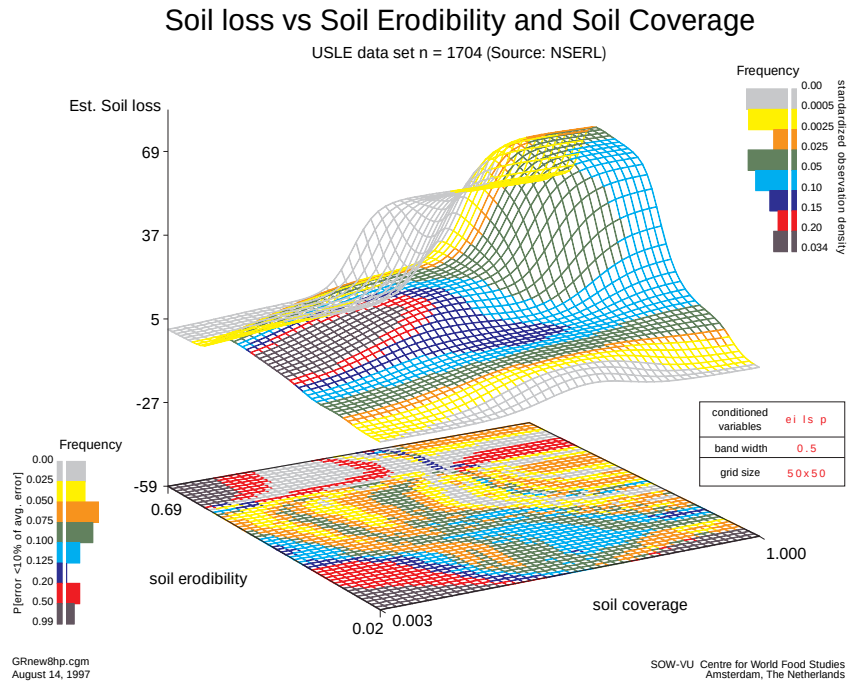

FIGURE 8

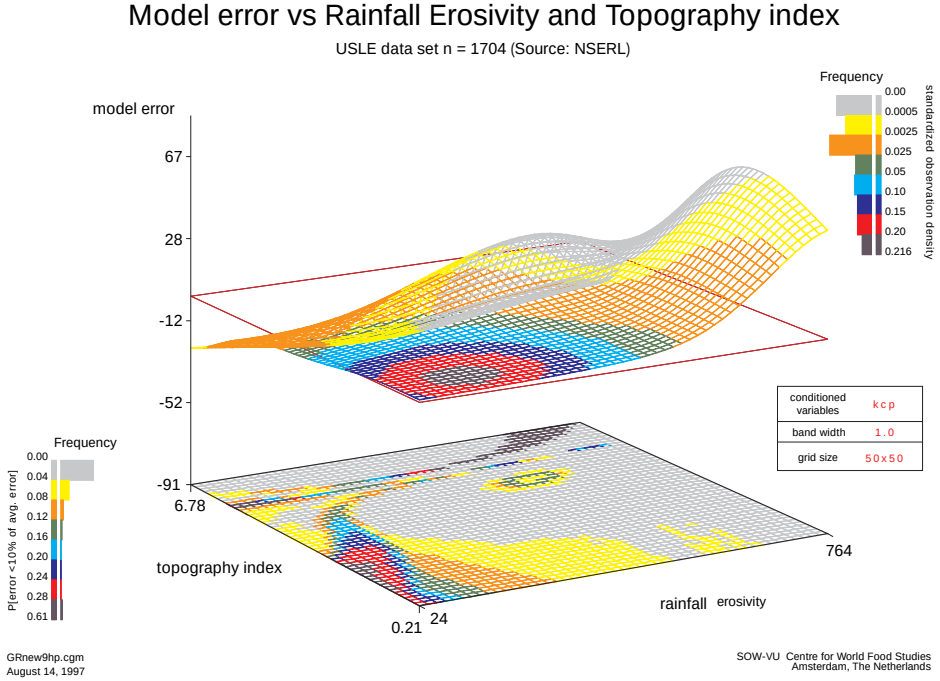

FIGURE 9

conditioned on their mean. The red line indicates the zero level of the model error.

Figure 10 shows the same relationships with a band width of half the original value. The graph has been rotated a 180 to facilitate the visualization of the curve at the extreme values of the rainfall erosivity and topography index.

The regression curve in Figure 9 shows that USLE calculations have a low model error for low rainfall erosivity and topography values, which part corresponds with the highest observation density. The USLE grossly underestimates (positive errors) soil loss for high values of the rainfall erosivity in combination with low values of topography. The USLE overestimates soil loss for higher values of topography. The large model deviations occur, as expected, in areas where observation density and probability of a correct estimate are low.

CONCLUSIONS

This paper has described how non-parametric techniques can be combined with cartographic skills to characterize large datasets and model results. The non-parametric approach has the advantage that no a priori model is imposed on the data. The visualization of the mollified values in multidimensional graphs shows the relationships between (conditioned and unconditioned) variables and uses colouring and planes to accommodate

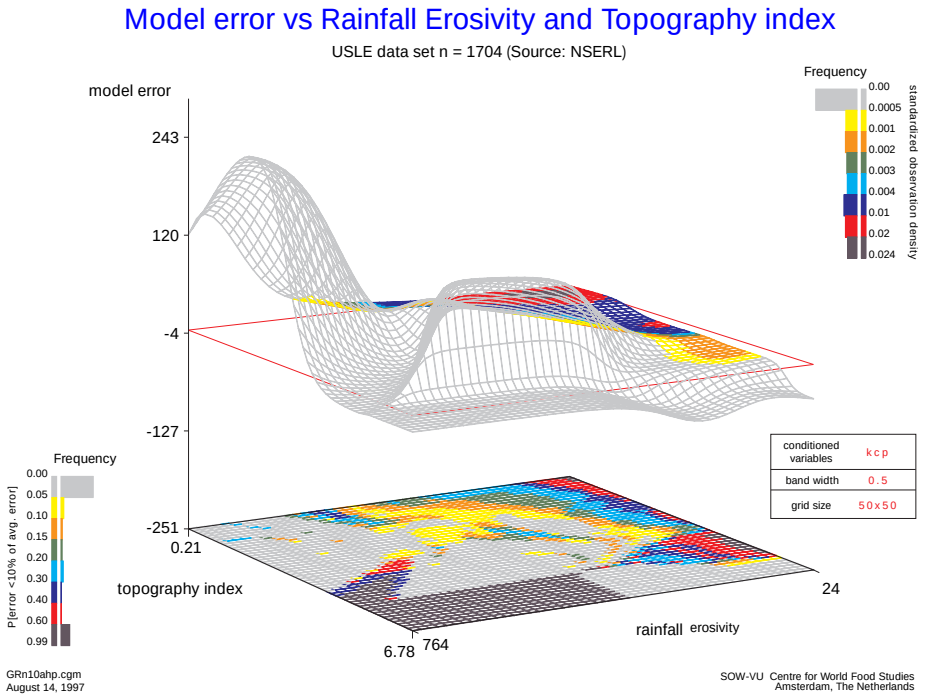

FIGURE 10

information on statistical properties of the estimated values. This allows non-linearities of relationships to be studied and, at the same time, the reliability of the estimated value to be evaluated. The application of the mollifier method to the USLE dataset shows that observation densities are low for the higher R and LS values and the lower and higher K values. Applying the USLE in these areas might lead to large errors, as is shown in the example with mollified model error value.

A disadvantage of the non-parametric method is that while the model derived in this manner could be highly flexible and elaborate, it will only reflect theoretical restrictions through the choice of variables and not through properties of the functions, precisely because the moulding of the model into a particular shape is almost exclusively driven by the data. In the context of policy modelling, this is often considered unsatisfactory and a parametric approach would eventually be preferable. In fact, the non-parametric model can be viewed as an intermediate product. Research is currently under way to fit a parametric model to the non-parametric function (Ermoliev and Keyzer, in prep). This has the advantage over direct estimation of the parametric model in that it becomes possible to ensure a good fit not only at points where data are available but also in between these points. Another line of ongoing research is to develop software that allows the visual representation to be generated in an interactive manner, allowing the user to "visit" his dataset and explore the underlying interrelationships in it—much like in an adventure game.

ACKNOWLEDGMENTS

The authors would like to thank the National Soil Erosion Research Laboratory, USDA-ARS, West Lafayette, United States, for the use of their dataset. P J Albersen (SOW-VU) wrote the SAS programs that were used in this paper; his contribution is gratefully acknowledged.

REFERENCES

1 Arnoldus, H M J. 1980. An approximation of the rainfall factor in the USLE. In: M de Boodt and D Gabriels (eds), Assessment of Water Erosion. John Wiley, Chicester.

2 Beirens, H J. 1987. Kernel estimations of regression functions. Advances in Econometrics 6, Cambridge Univ Pr.

3 Ermoliev, Y, V Norkin and R Wets. 1995. The minimization of semicontinuous functions: mollifier subgradients. SIAM J of Control and Optimization 33, pp 149-167.

4 Hargreaves, G H. 1981. Simplified methods for rainfall intensities. J of the Irrigation and Drainage Div, proc of the Amer Soco f Civil Engineers 107, IR3, pp 28-288.

5 Hudson, N. 1976. Soil Conservation. Batsford, London.

6 Hurni, H. 1985. Erosion-productivity-conservation systems in Ethiopia. In: Pla Sentis I, Soil Conservation and Productivity. Proc of 4th internatl conf on soil conservation. Maracay, Venezuela.

7 Liu, B Y, M A Nearing and L M Risse. 1994. Slope gradients effects on soil loss for steep slopes. Trans of ASAE 37, 6, pp 1835-1840.

8 McCool, D K, L C Brown, G R Foster, C K Mutchler and L D Meyer. 1987. Revised slope steepness factor for the Universal Soil Loss Equation. Trans of ASAE 30, 5, pp 1387-1396.

9 Nadaraya, E A. 1989. Nonparametric Estimation of Probability Densities and Regression Curves. Trans from Russian by S Klotz. Kluwer, Amsterdam.

10 Parzen, E. 1962. On estimation of a probability density function and the mode. Annals of Math Stat 33, pp 1065-1076.

11 Pflug, G. 1996 Optimization of Stochastic models. Kluwer, Amsterdam/Dordrecht.

12 Risse, L M, M A Nearing, A D Nicks and J M Laflen. 1993. Error assessment in the universal soil loss equation. Soil Sci Soc Am J 57, pp 825-833.

13 Sobolev, S L. 1988 Some Applications of Functional Analysis in Mathematical Physics (3rd edit). Nauka (in Russian), Moscow.

14 Vanelslande, A, R Rousseau, R Lal, D Gabriels and B S Ghuman. 1984. Testing the applicability of a soil erodibility nomogram for some tropical soils. In: D E Walling, S S D Foster and P Wurzel (eds), Challenges in African Hydrology and Water Resources. IAHS publ 144, pp 463-473.

15 Wischmeier, W H. 1976. Use and misuse of the universal soil loss equation. J of Soil and Water Conserv, Jan-Feb, 1976.

16 Wischmeier W H and D D Smith. 1978. Predicting Rainfall Erosion Losses, Agric Handbook 537, USDA, Washington DC.

RESUME

La méthode d'adaptation "mollifier" est une technique numérique qui a été largement appliquée en physique et chimie pour approcher des fonctions mathématiques de forme irrégulière. Dans cet article, nous proposons de l'utiliser pour une interpolation non-paramétrique pour caractériser des séries de données et des modèles. L'idée de base est d'étendre une "toile souple" sur un profil de points créés par une série empirique de points ou par un modèle mathématique. Si la toile est maintenue "tendue", sa surface va lisser la plupart des irrégularités du profil sous-jacent. Ceci revient à moyenner de l'information sur une fenêtre plus large (proximité) autour de chaque point. L'estimation résultant de cette "toile tendue" ("mollifier") peut être représentée graphiquement en trois dimensions avec une variable dépendante représentée par rapport à deux variables indépendantes, pour des valeurs fixes d'autres variables indépendantes. Comparé avec des méthodes paramétriques telles qu'une régression courbe (spline) ou une estimation des variations, la méthode "mollifier" a un avantage important, à savoir qu'elle donne une fiabilité statistique à chaque point et que cette mesure ne dépend pas de l'ajustement d'autres points. Ceci permet de créer une image de la toile avec une résolution plus élevée (en utilisant une couleur plus accentuée ou une ombre plus foncée) dans des régions où le levé est plus fiable. De plus, l'introduction d'un plan dans la représentation 3D est utilisée pour décrire la validité de l'ajustement. Comme illustration, nous appliquons cette technique à une série de données de l'équation universelle de perte de sol (USLE), à l'équation USLE elle-même et aux déviations entre les deux. Les graphiques qui en résultent montrent les valeurs estimées de pertes de sol annuelles, ainsi que l'erreur du modèle représenté par rapport à deux variables explicatives de l'équation USLE et pour les valeurs moyennes de variables exogènes non visibles. Il s'en dégage les propriétés suivantes: (1) Les données sont fortement concentrées dans certaines zones des graphiques où la probabilité d'une estimation correcte est relativement élevée et l'erreur modèle faible. L'application de l'équation USLE est ici relativement fiable. (2) Des tendances inattendues dans la forme de la courbe de régression sont montrées, alors que la représentation basée sur les données est très confuse dans cette région.

RESUMEN

El método de suavización (mollifier method) es una técnica numérica, que ha sido ampliamente aplicada en las ciencias físicas y químicas para aproximar funciones matemáticas de tipo irregular. En este artículo, se propone usar este método en la interpolación no-paramétrica para caracterizar series de datos y modelos. La idea fundamental es de colocar una "manta suave" sobre un perfil de puntos, los cuales han sido generados por una serie empírica de datos o por un modelo matemático. Si se hala "apretadamente" la manta, su superficie va a alisar la mayoría de las irregularidades del perfil subyacente. Esto conduce a promediar datos en una ventana más ancha (vecindario) alrededor de cada punto. El estimado de suavización resultante (la manta) puede ser representado gráficamente en tres dimensiones, con una variable dependiente cartografiada contra dos variables independientes, para valores fijos de otras variables independientes. Comparado con métodos paramétricos como la regresión de cuña (spline regression) o la estimación de variantes (variorum estimation), el método de suavización tiene la importante ventaja que da una medida de confiabilidad estadística en cada punto y que esta medida no depende del ajuste efectuado en otros puntos. Esto hace posible crear una imagen de la manta que tiene una resolución más alta (realzada mediante un color más fuerte o una sombra más oscura) en áreas donde el mapeo es más confiable. En adición, se utiliza la introducción de un plano en la representación tridimensional para mostrar la fineza del ajuste. Como ilustración, se aplica esta técnica a la serie de datos de la ecuación universal de pérdida de suelos (USLE), a la USLE misma, y a las desviaciones entre las dos. Los gráficos resultantes muestran los valores estimados de pérdida anual de suelos, así como el error del modelo mapeado contra dos variables explanatorias de la USLE y para los valores medios de variables exógenas no-visibles. Las siguientes propiedades se destacan: (1) Los datos se concentran fuertemente en ciertas áreas de los gráficos, donde la probabilidad de obtener un estimado correcto es relativamente alta y el error del modelo es bajo. Aquí, la aplicación de la USLE es relativamente confiable. (2) Aparecen tendencias inesperadas en la forma de la curva de regresión, mientras que el mapeo basado en datos es perfectamente nebuloso.

APPENDIX

Annex: the mollifier mapping

The function $P^s(x)$ can be derived in two equivalent ways either from kernel density estimation (Parzen, 1962, Bierens, 1987, Nadaraya, 1989) or from mollifier theory (see Sobolev, 1988, Ermoliev et al., 1995 and Pflug, 1996). Kernel density regression considers an unknown joint density function f(y, x) of bounded variance and seeks to characterise, on the basis of the empirical observations $\{y^s, x^s\}$, the regression function

$$R(x) \equiv E[y|x] \qquad \text{(A.1)}$$

Hence the basic model is

$$y = R(x) + \eta \qquad \text{(A.2)}$$

where η is an error with mean zero, bounded variance and unknown density. The Nadaraya-Watson Kernel density regression curve approximates R(x) by means of

$$R_s(x) = \sum_{s=1}^{S} y^s P_\theta^s(x) \qquad \text{(A.3)}$$ for

$$P_\theta^s(x) = \psi((x^s - x)/\theta)/\Psi_\theta^S(x), \text{ if } \Psi_\theta^S(x) > 0 \text{ and 0 otherwise} \qquad \text{(A.4)}$$

where

$$\Psi_\theta^S(x) = \sum_{s=1}^{S} \psi((x^s - x)/\theta) \qquad \text{(A.5)}$$

and window size (band-width) is set optimally according to some function

$$\theta = k(S), \qquad \text{(A.6)}$$

such that

$$k(S) \to 0 \quad \text{as } S \to \infty \quad \text{and}$$
$$(k(S))^m S \to \infty \quad \text{as } S \to \infty$$

The density function $\psi: R^m \to R_+$ is Borel measurable and satisfies: $\int \psi(\varepsilon)d\varepsilon = 1$; $\int \psi(\varepsilon)d\varepsilon < \infty$; $\lim_{\|\varepsilon\|^m \to \infty} \|\varepsilon\|\psi(\varepsilon) = 0$, where $\|\varepsilon\|$ is the Euclidean norm and $\sup_\varepsilon \psi(\varepsilon) < \infty$.

In this approach, expression $\psi((x^s - x)/\theta)$ in (A.5) can be interpreted as the likelihood of x being associated to the observation s and $\Psi_\theta^S(x)$ the likelihood of x being associated to any of the observations in the sample (summation is possible because of randomness of the sample). Hence probability $P_\theta^s(x)$ is the likelihood of x being associated to observation s, conditional on its association to at least one observation in the sample. Thus $y_\theta^S(x)$ is the

expectation of the y^s-values associated with the sample and is therefore subject to a sampling error on x^s.

The mollifier approach also uses (A.3) and (A.4) as estimator, but the underlying postulate about the relation between y and x is more specific. While the kernel density interpretation is common in econometrics, the mollifier is popular in optimization theory and is used to provide a "bird's-eye" view of models with stochastic variables, integer variables, and other complex features.

Thus both approaches yield the same finite sample estimation. For this paper we applied the mollifier because its terminology more closely suits the purpose of 3D-display. It considers a sample of size S of accurate observations y^s made at randomly selected points x^s drawn from a uniform distribution on the compact convex set X and assume that there is a measurement error in x^s. Let the likelihood of an error $\varepsilon^s = x^s - x$ (i.e. the likelihood of x being the correct co-ordinate associated to y^s) be characterised by a given mollifier density function $\psi(\varepsilon / \theta)$. The resulting stochastic model is:

$$y = R(x + \varepsilon) \tag{A.7}$$

Thus, unlike (A.2), the mollifier approach postulates a deterministic function $y(\cdot)$ and only accounts for error in associating an observation x^s to a given y^s. The mollifier mapping itself is defined as the expected value of $y(x + \varepsilon)$, given x and for $S \to \infty$:

$$y_\theta(x) = E\left[R(x + \varepsilon)\middle|x\right], \tag{A.8a}$$

$$= \int R(x + \varepsilon)\psi(\varepsilon / \theta)d\varepsilon \tag{A.8b}$$

$$= \lim_{s\to\infty} 1/S \sum_{s=1}^{S} R(x + \varepsilon)\psi(\varepsilon^s / \theta) \tag{A.8c}$$

and where the density function $\psi(\varepsilon / \theta)$ has Lebesgue measure, a mode at $\varepsilon = 0$, and for θ (the window size) going to zero, its support goes to zero.