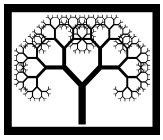
Paper 72



Proceedings of the Eighth International Conference on Engineering Computational Technology, B.H.V. Topping, (Editor), Civil-Comp Press, Stirlingshire, Scotland

Pedotransfer Functions Development by means of the Ensemble Data-Driven Methodology

M. Cisty, J. Bezak and J. Skalova
Department of Land and Water Resources Management
Faculty of Civil Engineering
Slovak University of Technology Bratislava, Slovak Republic

Abstract

Ensemble is one of the most widely used methods for improving the performance of data-driven regression models. Two ensemble methods, *i.e.* bagging and additive regression are applied in this paper for improving the performance of a single regression model for evaluating pedotransfer functions used in soil hydrology. The experimental results with data obtained from the Zahorska lowland in Slovakia indicated that mainly additive regression ensemble model showed improved performance over single neural networks and support vector regression.

Keywords: soil hydrology, pedotransfer function, artificial neural network, support vector machine, data driven model, ensemble.

1 Introduction

The water retention curve is one of the main soil hydraulic properties, which is used in simulating the water regime of soils. It represents the relationship between the water content and the soil's water potential (the potential energy of water per unit volume, which quantifies the tendency of water to move from one place to another). This curve is characteristic of different types of soil. It is used to predict a soil's water storage, the water supply to plants, and for other tasks in soil water modelling. In general, two categories of methods for evaluating a water retention curve can be distinguished: (1) direct measurement techniques in a laboratory or in the field and (2) methods that apply various regression models. However, despite the progress that has been achieved, the measurement techniques remain time consuming and costly. This is the reason why a relatively large number of works have appeared in the past which were devoted to determining the water retention curve from more

easily available soil properties such as particle size distribution, dry bulk density, organic C content, etc., e.g. [1,2,3]. In this context Bouma [4] introduced the term "pedotransfer function" (PTF), which he described as "translating data that we have (soil survey data) into data that we need (soil hydraulic data)." Tietje and Tapkenhinrichs [5] classified different types of PTF evaluations such as point estimation methods, parameter estimation methods, and semi-physical methods. In this paper we will focus on point estimation methods, which follow the direct approach by estimating the water content at predetermined pressure heads.

Besides the application of the standard regression methods for solving this task, data-driven techniques appeared in the scientific literature in the second half of the previous decade as a tool for solving regression tasks in developing PTFs. However, there is no overall best data-driven technique which could be used in building hydrology models, because their suitability depends on the details of the problem, the data structure, the input data used, etc. For this reason various data-driven techniques are analysed in this case study.

Artificial neural networks (ANNs) have become the tool of choice in the previous decade in developing PTFs, e.g., [6,7,8]. The above authors confirm that they received better results from ANN-based pedotransfer functions than from standard regression-based PTFs.

Recent developments in machine learning methods have forced the application of alternative data-driven methods in soil hydrology applications, e.g., support vector machines [9]. The same methodology was used in [10], which validated that the predictions by the SVM-based PTFs showed considerable improvement over ROSETTA, which is a well-established ANN-based methodology [11]. The foundations of support vector machines (SVMs) were developed by Vapnik [12] and are gaining in popularity due to their attractive features and promising empirical performance. Support vector machines have been established as a machine learning method with a promising ability to generalize, e.g., good performance with data which were not used in SVM model building. This means input data for which the modeller does not know the outputs, and it is one of the main aim of any modelling.

In recent years many researchers have investigated the technique of combining the predictions of multiple data-driven models to accomplish a single classification or regression task. The resulting model (hereafter referred to as an "ensemble" model) has often proved to be more accurate than any of the individual data driven-models making up an ensemble [13].

The objective of this work is to verify the declared advantages of ensemble learning against SVMs and ANNs in a case study, the aim of which is to develop PTFs for the Zahorská lowland in the Slovak Republic. The data used in this study were obtained from previous work [14] and various other measurements accomplished later in this area.

In the following part of the paper ("Methodology") the methods used in this study – ANN, SVM - are briefly explained, together with the ensemble methodologies used. Then the data acquisition and preparation is presented. In the "Results" part, the settings of the experimental computations are described in detail, and the "Conclusion" of the paper evaluates these experiments on the basis of the statistical indicators.

2 Materials and methods

2.1 Methods used to fit the pedotransfer functions (PTFs)

The first approach for modelling the PTFs used in this paper is the application of *artificial neural networks* (ANNs). This approach is mainly used for comparative purposes, e.g., for evaluating the possible gains obtained from the application of ensemble methodologies. This approach has been described in various previous works, and information about the subject can be found in [8,15,16], etc. Briefly summarized, a neural network consists of input, hidden and output layers, all containing "nodes" or "neurons." The number of nodes in the input layer and output layer correspond to the number of input and output variables of the model. So-called "learning" or "training" involves adjustment of the coefficients (i.e., the synaptic connections that exist between the neurons or weights), which are used for the transformation of the inputs to the outputs. The basic information about the application of an ANN to regression problems is available in the literature and is well known, so we will not provide a more detailed explanation here.

As was mentioned in the introductory section of this paper, also *support vector machines* (SVM) have been applied to PTF evaluations, and the SVMs in this work are also used for a comparison with the ensemble learning methodologies.

The basic idea behind the SVM regression of nonlinear functions is to project the input data by means of kernel functions into a higher dimensional space called the *feature space*, where a linear regression can be performed for an originally nonlinear problem which is to be solved. The results of the regression are then mapped back to the input space.

The next important concept in SVM methodology is to fully ignore small errors (by introducing the variable ε, which defines what the "small" error is) to make the regression task dependent on a smaller number of inputs than were given in the original task, which makes the methodology much more computationally treatable. These crucial vectors of the inputs are called the support vectors.

In an ε-SVM regression [12], the goal is to find a function *f(x)* that at most has an ε deviation from the actually obtained targets y_i (or *f(x)*) for the training data:

$$f(x) = w.\Phi(x) + b \quad w \in X, b \in R \tag{1}$$

where *f(x)* is the model's output, and input *x* is mapped into a feature space by a nonlinear function $\Phi(x)$ with the weight vector *w* and bias *b*.

The goal of a regression algorithm is to fit a flat function to the data points. "Flatness" means that one seeks a small w. One way to ensure this flatness is to minimize the norm, i.e. $||w||^2$. Thus, the regression problem can be written as a quadratic optimization problem:

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2}$$

$$\text{subject to: } y_i - (w.\Phi(x) + b) \le \varepsilon + \xi_i$$

$$(w.\Phi(x) + b) - yi \leq \varepsilon + \xi_i^*$$
$$\xi i, \xi i^* \geq 0$$

where ξ_i, ξ_i^* are slack variables that specify the upper and lower training errors, subject to an error tolerance ε (soft margin), and C is a positive constant that determines the degree of the penalized loss when a training error occurs. In Equation system (2), the first term of the objective function indicates the model's complexity, and the second term is the empirical risk. That is why this objective function simultaneously minimizes both the empirical risk and the model's complexity; the trade-off between these two goals is controlled by parameter C.

SVMs can be solved by transforming the above-described optimization problem into its dual form via a quadratic programming algorithm (utilizing Lagrange multipliers), and the solution to the quadratic programming is unique and optimal.

The radial basis function was chosen on a trial and error basis as the kernel function for this work (the function used to transform a nonlinear problem from an input space to a high dimensional space). This function has the following form:

$$K(x_i, x_j) = \exp(-\gamma^* || x_i - x_j||^2), \gamma>0 \quad (3)$$

The parameter γ of this kernel function, the tube size ε for the ε-insensitive loss function, and parameter C should be found, which the basic task is when SVMs are applied to any practical problem.

In this paper a comparison of SVM and ANN as established methodologies for evaluating pedotransfer functions are compared with ensemble learning methods. Many researchers have investigated the technique of combining the predictions of multiple data-driven models to produce a single model. The resulting model is usually more accurate than any of the individual models making up the ensemble. In practice, two basic conditions should be satisfied to achieve a good ensemble: accuracy and diversity [17].

Two ensemble methods are compared in this study: bagging and additive regression. The basic concepts behind these methodologies follow; the details of their implementation are described in the application part of the paper.

Bagging (short for "bootstrap aggregating") is one of the earliest ensemble learning algorithms [18]. It is relatively simple to implement, with a remarkably good performance. Bagging in its first step generates m training sets by sampling examples from the original training data with a replacement (a so-called bootstrap sample). Then the m regression models are trained with each of the training sets. Bootstrap aggregation or bagging averages these predictions, thereby reducing the final variance of the model's output and helping to avoid overfitting. Combining multiple models in such a way helps only when these models complement one another (every model is "specialized" to a different part of the input – output domain), and each one treats a reasonable percentage of the data correctly. Bagging is particularly useful when the available data is of a limited size, which is the case in this case study. To ensure that there are sufficient training samples in each subset, relatively large portions of the samples (75–100%) are drawn into each subset. This causes the individual training subsets to overlap significantly, with many of the

same instances appearing in most subsets and some instances appearing multiple times in a given subset [19]. Bagging usually consists of decision tree models, but it can be used with any type of model. In this work multilayer perceptron (MLP) was used, which satisfies the mentioned conditions.

Additive regression is another effective ensemble learning method, which uses a set of base learners to achieve greater predictive accuracy. Additive regression implements forward stage wise additive modelling. It starts with an empty ensemble and incorporates new members sequentially. At each stage the model that maximizes the predictive performance of the ensemble as a whole is added, without altering those already in the ensemble. The first regression model – for example, a MLP could be used – maps the input data to the outputs as usual. Then the residuals between the predicted and observed values are corrected by training a second model. Adding the predictions made by the second model to those of the first one yields fewer errors on the training data. The methodology continues with the next model, which learns to predict the residuals of the residuals, and so on [20].

2.2 Study area and data collection

The data used in this study were obtained from a previous work [14]. An area of the Zahorska lowland was selected for testing the methods described. A total of 226 soil samples was taken from various localities in this area.

The soil samples were air-dried and sieved for a physical analysis. A particle size analysis according to four grain categories was performed utilizing Cassagrande's methods. Category I means the percentages of the clay (diameter < 0.01 mm), category II - silt (0.01–0.05 mm), category III - fine sand (0.05–0.1 mm) and category IV - sand (0.1–2.0 mm). The dry bulk density, particle density, porosity and saturated hydraulic conductivity were also measured on the soil samples. The points of the drying branches of the WRCs for the pressure head values of -2.5, -56, -209, -558, -976 and -3060 cm were estimated using overpressure equipment (set for pF-determination with ceramic plates).

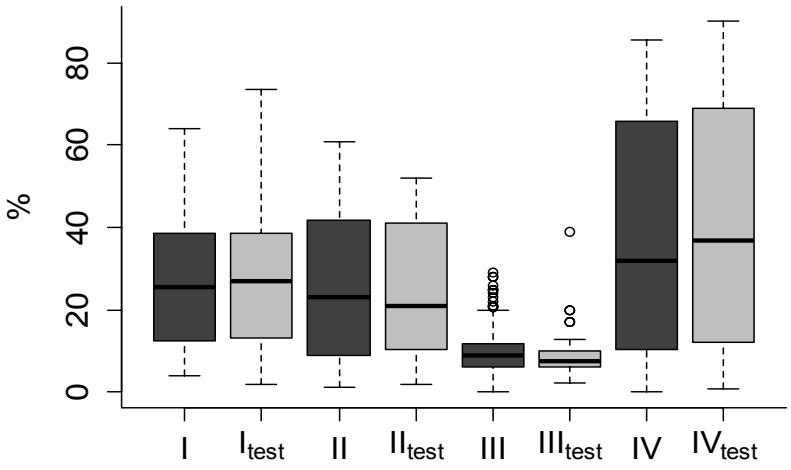


Figure 1. Comparison of grain categories (I, II, III, IV) in the training and testing data

A full database of the 226 samples and their properties were used for creating the input data for the modelling from which the training and testing subsets of the data were produced. The training data consist of 181 data samples and test data from 45 data samples. Statistically similar data should be in both data subsets; this condition is visualized by the boxplots on Figure 1. In this figure I, II, III and IV are grain categories in training set of data and the same identification with the subscript "test" is used for the test set. From this evaluation it can be seen that category III will probably have the lowest impact on the pedotransfer function evaluation, but it will be included in the input data, anyway.

3 Results

The first approach applied to determining the water retention curves in the presented work was the *neural networks* methodology (ANN). In this work a multilayer perceptron with 4, 5, and 6 neurons in the hidden layer was tested; an ANN with 5 neurons in the hidden layer was finally chosen for the final neural network model used in the comparisons (it has the best results). A neuron with a hyperbolic tangent activation function was used in the hidden layer and a linear activation function in the output layer. The Levenberg-Maquardt method was used in the context of the back propagation method. The networks were trained to compute the water content at the pressure head value h_w = -2.5, -56, -209, -558, -976, -3060 cm. The "hold-out" method was used for stopping the ANN to avoid overtraining, and this "hold-out" sample was 20% of the data in the training set.

Then the testing dataset was computed with the trained ANNs. The results with the regression coefficients are summarized in Table 1.

h_w [cm]	ANN – H4	ANN – H5	ANN – H6	SVM
-2.5	0.874	0.883	0.879	0.872
-56	0.846	0.857	0.849	0.872
-209	0.874	0.874	0.866	0.898
-558	0.866	0.872	0.873	0.896
-976	0.853	0.859	0.860	0.882
-3060	0.833	0.846	0.852	0.880

Table 1. Correlation of the model's results with the actual values of the PTFs. Three variants of the ANN with different hidden layer sizes and SVM are evaluated (h_w - pressure head, H4 – H6 is the number of neurons in the hidden layer).

For a comparison with the ensemble approach, the given regression problem was also solved using *support vector machines (SVM)*. The estimation of the practical steps of the SVM regression are as follows: 1) selecting a suitable kernel and the appropriate kernel's parameter (γ in Equation 3); 2) specifying the ε parameter (Equation 2); and 3) specifying the capacity C (Equation 2).

The radial basis function was chosen as the kernel function on a trial and error basis, and the cross-validation methodology with 10 folds was used for finding the mentioned parameters of the SVM model.

In the training phase, SVM models for computing the water content for the pressure head values of h_w = -2.5, -56, -209, -558, -976 and -3060 cm were created (on the basis of the particle size distribution). Then the testing dataset was computed with the models obtained, and the final results were summarized with the help of the regression coefficients in Table 2. The calculations of the SVM were performed using the LIBSVM library developed by Chang and Lin [21].

This paper proposes the application of *bagging* to obtain more robust and accurate predictions using ANN regression models. Bagging also helps to avoid overfitting. The training data were resampled using the bootstrap method to form five training sets of the same size as the original training data set (181 samples) from which the five ANN models were developed and combined to provide the predictions. The model selection process was not necessary for the neural networks used in the bagging since an ensemble of neural networks tends to cancel out each other's errors. Thus, we simply chose a large enough number of hidden units - around the number of input variables (five) in this experiment; the hyperbolic tangent activation function was used in the hidden layer and the linear function in the output layer. Each of the input vectors consists of five elements and includes four values of the particle distributions in the four classes and the value of the dry bulk density. A simple averaging of the five predictions from the ANN basic learners is used for the final prediction. The results with the regression coefficients are summarized in Table 1.

Secondly, we applied *additive regression ensemble* methodology, which uses the Gaussian process as a base learner. Gaussian processes have attracted significant interest in the data mining community. As a result of their good performance in practice, Gaussian process models have been applied to various applications, such as rehabilitation engineering [22], machining optimization [23] and digital terrain modelling [24 5]. Gaussian process prediction is also well known in the geostatistics field [25] where it is known as "kriging." A Gaussian process is specified by a mean and a covariance function. The mean is a function of x (which is often the zero function), and the covariance is a function C(x, x'), which expresses the expected covariance between the value of the function y at the points x and x'. Gaussian processes are kernel-based methods for solving regression or other pattern recognition problems - the term "kernel function" used in an SVM is approximately equivalent to the covariance function. In this work the Laplace kernel was chosen by trial and error:

$$K(x, x') = - \exp(-\gamma||x-x'||) \quad (4)$$

The setting of the parameters (length scale γ used in the kernel and the maximum number of basis vectors to be used) was optimized by tenfold cross-validation. Moreover, ten iterations of the additive regression were performed for the evolution of the pedotransfer functions. Additive regression starts with building a first prediction model for the original input and output data in the training set. The same variables were selected as in the other methods evaluated in this work. A base model

is built in each additive regression iteration. In the second and later iterations, the residuals of the labels are calculated, and the model is trained to predict them instead of the original output data. The trained meta model predicts the output variable (water content for the particular pressure head value h_w) by adding all the base model predictions. The results with the regression coefficients are summarized in Table 2.

h_w:	-2.5	-56	-209	-558	-976	-3060
Additive Regression	0.9020	0.9000	0.9080	0.9100	0.9080	0.8990
Bagging	0.8730	0.8680	0.8940	0.8930	0.8780	0.8850

Table 2. Correlation of the ensemble models results with the actual (measured) values of the PTFs

From the results expressed by the correlation coefficient in tables 1 and 2, it can be seen that the ensemble methods evaluated in this case study give better results than when individual learners are used solo (ANN, SVM).

	MLR	ANN	SVM	BAGG	AR
ME	-0.39	-0.53	-1.02	3.11	-1.04
MAE	4.21	3.75	3.31	4.88	2.99
MSE	30.40	25.98	21.44	31.03	18.34
RMSE	5.51	5.10	4.63	5.57	4.28
NRMSE	57.50	53.10	48.30	58.10	44.60
PBIAS	-1.80	-2.40	-4.70	14.30	-4.80
r	0.82	0.85	0.88	0.87	0.90
maxD	9.83	7.12	6.24	12.89	6.85
minD	-15.27	-15.02	-12.91	-11.01	-12.05

Table 3. Evaluation of the various models for prediction of the water content at h_w = 3060 cm by different statistics

A more detailed evaluation of the various data-driven methods applied in this work is presented in Table 3. For practical reasons (the limited extent of this paper) it is restricted only to an evaluation of the prediction of the water content for the pressure head value h_w = -3060 cm. The results for the other pressure heads are similar. In Table 3 the mean error (ME), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), normalized root mean square error (NRMSE), percent bias (PBIAS), correlation coefficient (r), maximal difference between the simulated and actual values (maxD) and the minimal difference between the simulated and actual values (minD) are evaluated. The names of the models in the heading of Table 3 are clear from their abbreviations. From this analysis it is evident that it is worthwhile to pay attention to the development and

choice of the proper regression model when evaluating the pedotransfer function, because it can be seen that a relatively big difference is between the effectiveness of the worst performing model (ANN) and the best model. The models are ordered in columns according to their quality. Of the two ensemble models evaluated (bagging, and additive regression), the best results were obtained by additive regression. In this evaluation bagging seems to work comparably to ANN and SVM with negligible improvement, and some statistics show that it works even worse. This could be due to the insufficient variability in the bagging model, which used five ANN models in the case investigated; moreover, these models do not offer very precise results.

4 Conclusions

This paper proposes and evaluates ensemble models for the development of pedotransfer functions for the point estimation of the soil-water content for six pressure head values h_w from the basic soil properties (particle-size distribution, bulk density). The ensemble data-driven models were compared to single data-driven models (artificial neural networks and support vector machines). The accuracy of the predictions was evaluated by the correlation coefficient between the measured and predicted parameter values and by other statistics. From the results obtained it was shown that for this task ensemble data-driven methods work better than single data-driven methods. The best regression method was obtained by the additive regression ensemble methodology composed of Gaussian process as base learner.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under Contract No. LPP-0319-09 and APVV-0139-10, and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/1044/11 and 1/0243/11.

References

[1] S.C. Gupta, W.E. Larson, “Estimating soil water retention characteristics from particle size distribution, organic matter percentage, and bulk density”, Water Resour. Res. 15, 1633-1635, 1979.

[2] W.J. Rawls, D.L. Brakensiek, K.E. Saxton, “Estimating soil water retention properties”, Trans. ASAE 25, 1316-1320, 1982.

[3] B. Minasny, A.B. McBratney, K.L. Bristow, ”Comparison of different approaches to the development of pedotransfer functions for water retention curves”, Geoderma, 93, 225–253, 1999.

[4] J. Bouma, “Using Soil Survey Data for Quantitative Land Evaluation”, Adv. Soil Sci., 9, 177–213, 1989.
[5] O. Tietje, M. Tapkenhinrichs, “Evaluation of pedo-transfer functions”, Soil Sci. Soc. Am. J, 57, 4. pp. 1088-1095, 1993.
[6] YA.A. Pachepsky, D.J. Timlin, G. Varallyay, ”Artificial neural networks to estimate soil water retention from easily measurable data”, Soil Sci. Soc. Am. J., 60, 727-733, 1996.
[7] S. Tamari, J.H.M. Wosten, J.C. Ruiz-Suarez, “Testing an artificial neural network for predicting soil hydraulic conductivity”, Soil Sci. Soc. Am. J. 60, 1732-1741, 1996.
[8] M. G. Schaap, F.J. Leij, M.Th. Van Genuchten, “Neural network analysis for hierarchical prediction of soil hydraulic properties”, Soil Sci. Soc. Am. J. 62, 847–855, 1998.
[9] M.G. Schaap, F.J. Leij, M. Th. van Genuchten, Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions, Journal of Hydrology, Volume 251, Issues 3–4, 1 October 2001, Pages 163-176, ISSN 0022-1694, 10.1016/S0022-1694(01)00466-8.
[10] K. Lamorski, Y. Pachepsky, C. Slawinski, R.T. Walczak: “Using support vector machines to develop pedotransfer functions for water retention of soils in Poland”, Soil Science Society of America Journal, 72 (2008), pp. 1243–1247, 2008.
[11] N.K.C. Twarakavi, J. Simunek, M.G. Schaap, “Development of Pedotransfer Functions for Estimation of Soil Hydraulic Parameters Using Support Vector Machines”, Soil Sci. Soc. Am. J., vol. 73, 1443–1452, 2009.
[12] V. Vapnik, ”The Nature of Statistical Learning Theory”, Springer, NY, 1995.
[13] L. Baker, D. Ellison, “The wisdom of crowds - ensembles and modules in environmental modelling”, Geoderma, 147 (1-2), 1-7, 2008.
[14] J. Skalová, ”Pedotransfer functions of the Záhorská nížina soils and their application to soil-water regime modelling”, Faculty of Civil Engineering STU Bratislava, 112 pp. (in Slovak), 2001.
[15] B. Minasny, A.B. McBratney, K.L. Bristow, ”Comparison of different approaches to the development of pedotransfer functions for water retention curves”, Geoderma, 93, 225–253, 1999.
[16] B. Minasny, A.B. McBratney, ”The neuro-m methods for fitting neural network parametric pedotransfer functions”, Soil Sci. Soc. Am. J., 66, 352–361, 2002.
[17] T. Windeatt, G. Ardeshir "An empirical comparison of pruning methods for ensemble classifiers", IDA2001, LNCS, vol. 2189, pp. 208–217, 2001.
[18] L. Breiman, "Bagging predictors", Machine Learning, 24 (2), 123–140, 1996.
[19] G. Wang, J. Hao, J. Ma, H. Jiang, ”A comparative assessment of ensemble learning for credit scoring”, Expert Systems with Applications, 38, 1, 223-230, 2011.
[20] I.H. Witten, E. Frank, M.A. Hall, “Data mining”, Morgan Kaufmann Publishers, 2011.
[21] C.Ch. Chang, C.J. Lin,”A library for support vector machines”, ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.

[22] R. Kamnik, J. Shi, R. Murray-Smith, T. Bajd, “Nonlinear modeling of FES-supported standing-up in paraplegia for selection of feedback sensors”, IEEE Trans. Neural Syst. Rehabil. Eng., 13, 40–52, 2005,
[23] J. Yuan, K.Wang, T. Yu, M. Fang, “Reliable multi-objective optimization of high- speed WEDM process based on Gaussian process regression”, Int. J. Mach. Tools Manuf., 48, 47–60, 2008.
[24] S. Vasudevan, F. Ramos, E. Nettleton, H. Durrant-Whyte, “Gaussian Process Modeling of Large Scale Terrain”, Journal of Field Robotics, vol. 26(10), 2009.
[25] N.A.C. Cressie, “Statistics for Spatial Data”, Wiley, New York, 1993.