Development of Liquefaction Index Prediction Equations from Post-Liquefaction CPT Data Using ANN and GEP

Sinjan Debnath[1][0000-0002-0148-8496] and Parbin Sultana[1][0000-0002-1241-4302]

[1] National Institute of Technology Silchar, Assam, India
sinjandebnath@gmail.com
parbinsultana@rediffmail.com

Abstract. In this paper, Artificial Neural Network (ANN) and Genetic Expression Programming (GEP) have been used to develop two Liquefaction Index (LI) equations which will be able to predict effectively whether a soil layer at any depth would liquefy or not in case of an earthquake. 226 post-liquefaction Cone Penetration Test (CPT) data (133 are liquefied cases and the rest 93 are non-liquefied cases) have been collected from published literature and using the collected data, an ANN and a GEP model have been built. From each developed model, a LI equation has been developed which uses CPT data of soil and Peak Ground Acceleration (PGA) as inputs and returns either 1 or 0 (1 means liquefaction may occur and 0 means liquefaction may not occur). A comparative study between both the models has also been conducted in this study.

Keywords: ANN, GEP, Liquefaction.

1 Introduction

Liquefaction is a phenomenon which can be observed in loose-saturated cohesionless soil deposit during earthquake shaking under undrained conditions. The loose cohesionless soil grains have a tendency to densify under the application of static and cyclic loading. During earthquake shaking, the loose saturated cohesionless soil grains tend to densify due to which the pore water pressure in the soil increases (under undrained conditions) and as a result, the effective stress in the soil decreases which in turn decreases the corresponding shear strength of the soil. As the effective stress in the soil completely reduces to zero, the soil totally loses its shear resistance and starts flowing like a fluid [1].

In this paper, two popular Machine Learning (ML) techniques – Artificial Neural Network (ANN) and Genetic Expression Programming (GEP) have been adopted to model empirical Liquefaction Index (LI) equations in order to predict future liquefaction cases using seismic, soil and Cone Penetration Test (CPT) parameters as inputs. The equations have been built and tested using a high-quality post liquefaction CPT database. The equations return 1/0 if, the soil layer is/not potentially liquefiable at a particular depth under seismic loading.

The developed equations can be used by any person having no prior knowledge of ML techniques. A comparative study amongst the developed empirical LI equations has been conducted and the best LI equation for the prediction of future liquefaction cases has been reported. Finally, sensitivity analysis has been carried out and the order of influence of the input parameters on the output parameter i.e., LI has been found out.

2 Database Compilation

A database has been compiled using 226 post liquefaction CPT records (133 liquefied and 93 non-liquefied cases) collected from [2]. The database contains CPT data of more than 52 sites and field observations of 6 different earthquakes, 4 in U.S. and 1 each in China and Taiwan respectively. Each record contains information about depth of the borehole, cone penetration resistance, friction factor, effective and total vertical stress, peak ground acceleration, moment magnitude of the earthquake and the corresponding liquefaction index (1 for liquefied cases and 0 for non-liquefied cases).

3 Model Inputs Selection

The database contains six input parameters which are boring depth (D), cone tip resistance (q_c), friction ratio (R_f), total stress at the boring depth (σ_v), effective stress at that same depth (σ_v') and peak horizontal ground acceleration (a_{max}).

Table 1. Co-efficient of correlation matrix.

	D	q_c	R_f	σ_v	σ_v'	a_{max}
D	1					
q_c	0.24	1				
R_f	0.35	-0.27	1			
σ_v	0.99	0.24	0.37	1		
σ_v'	0.92	0.25	0.31	0.92	1	
a_{max}	0.11	0.04	0.03	0.11	0.27	1

Now, so many inputs would make the models complex. In order to simplify the models, some highly correlated inputs can be eliminated [3]. To identify such inputs, a correlation analysis has been performed. Table 1 presents the correlation matrix. It can be observed in Table 1 that D has a high positive correlation with σ_v and σ_v'. Again, a high correlation exists between σ_v and σ_v'. As σ_v' covers the influence of the remaining two parameters (i.e., D and σ_v), only σ_v' has been selected instead of choosing all the three concerned parameters.

Therefore, input parameters finally considered for building the ML models are –

- Vertical effective stress at the boring depth (σ_v') (in kPa)
- Cone tip resistance (q_c) (in MPa)
- Peak horizontal ground acceleration (a_{max}) (in terms of g)
- Friction ratio (R_f) (in %)

4 Training and Testing Set

70% of the entire database (156 records) has been assigned to the training set and the remaining 30% (70 records) has been assigned to the testing set. The function of training set is to build up the ML models, to make the models learn the underlying patterns in the data while the function of testing set is to evaluate the trained models, calculate the error between the actual and predicted output and helps in optimizing the models. As suggested by Shahin et al. [4], statistical consistency has been maintained for both the sets, i.e., Mean and standard deviation of the data in both the sets have been kept as close as possible and the maximum and minimum values of all the parameters have been included in the training set in order to increase the range of interpolation.

Table 2. Statistical parameters of training and testing set.

Model parameters and dataset	**Mean**	**Standard deviation**	**Max**	**Min**	**Range**
q_c (MPa)					
Training set	5.745	4.069	25	0.9	24.1
Testing set	5.978	4.168	19.4	1.1	18.3
R_f					
Training set	1.231	1.064	5.2	0.1	5.1
Testing set	1.189	1.018	4.9	0.1	4.8
σ_v' (kPa)					
Training set	75.253	34.892	215.2	22.5	192.7
Testing set	73.303	33.469	161.6	23.9	137.7
a_{max} (g)					
Training set	0.297	0.147	0.8	0.08	0.72
Testing set	0.272	0.136	0.69	0.08	0.61

5 Development of the ANN Model

A two-layer feed forward ANN model has been built in MATLAB R2013b environment using tan-sigmoid transfer function in both hidden and output layers. Bayesian Regularization (BR) back-propagation learning algorithm has been used to train the ANN model.

5.1 Optimizing the Number of Hidden Nodes

In case of an ANN model, optimal number of hidden nodes is needed to be found out in order to obtain the optimum performance of the model. Hecht-Nielsen [5] suggested that for building an ANN model having ‘i’ number of inputs, maximum possible number of hidden nodes may be considered as (2i+1). As we have four input parameters in this case, several ANN models having number of hidden nodes starting from 1 to 9 have been created and the ANN model having the lowest Mean Squared Error (MSE) has been taken into account and the corresponding number of hidden nodes has been chosen as the optimal number of hidden nodes which is four in this case.

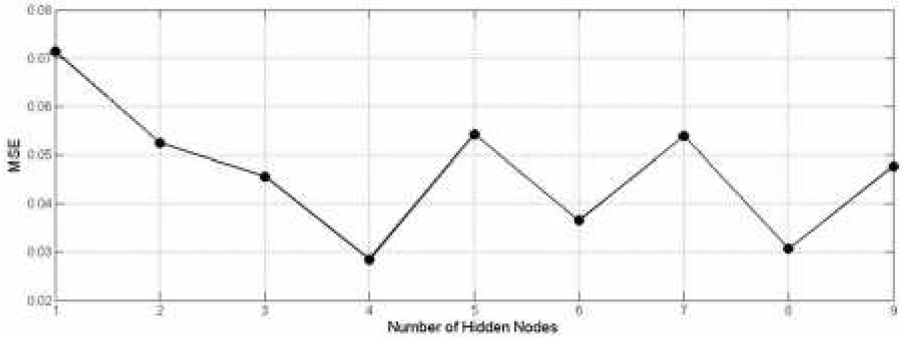


Fig. 1. Plot between MSE and number of hidden nodes.

5.2 Building the Optimum ANN Model

The optimum ANN model has been generated by using four hidden nodes. ANN architecture of the model has been provided for better understanding.

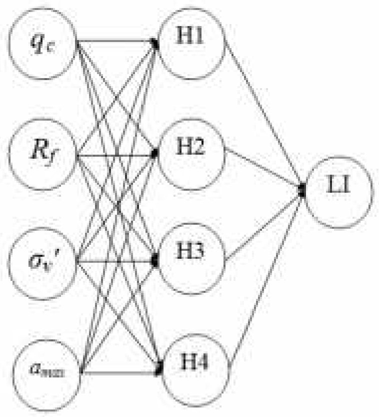


Fig. 2. Architecture of the ANN Model.

Table 3. Performance of the ANN Model.

	Training Set (156 cases)		**Prediction Rate (%)**
Liquefied cases	Actual	96	100
	Predicted by the ANN model	96	
Non-liquefied cases	Actual	60	95
	Predicted by the ANN model	57	
	Testing Set (70 cases)		**Prediction Rate (%)**
Liquefied cases	Actual	37	100
	Predicted by the ANN model	37	
Non-liquefied cases	Actual	33	87.88
	Predicted by the ANN model	29	

5.3 Development of LI Equation from the Optimum ANN Model

The mathematical expression as suggested by Goh et al. [6] incorporating all the independent input parameters and the dependent output parameter can be written as

$$Y_n = f_o\left\{b_o + \sum_{k=1}^{h}\left[w_k\, f_h\left(b_{hk} + \sum_{i=1}^{m} w_{ik}X_{ni}\right)\right]\right\} \tag{1}$$

Where,
Y_n = Predicted output parameter.
f_h & f_o = Tan-sigmoid transfer function for hidden and output layer respectively.
b_o = Output nodal bias.
w_k = Connection weight between the *k*th hidden node and the output node.
b_{hk} = Bias of the *k*th hidden node.
w_{ik} = Connection weight between the *i*th input node and the *k*th hidden node.
X_{ni} = *i*th input parameter.

Table 4. Weights and biases of the ANN model.

Hidden Node	**Input-Hidden Weight**				**Hidden-Output Weight**	**Bias**	
	q_c	R_f	σ_v'	a_{max}	**LI**	**Hidden**	**Output**
1	-5.433	-2.34	-0.91	2.692	21.267	-3.261	-8.69
2	5.71	-9.609	-2.659	2.383	-11.174	-0.405	
3	2.535	-4.33	-3.99	-1.471	15.192	-1.915	
4	-8.225	-6.036	14.718	12.681	13.09	7.933	

By substituting the values of weights and biases shown in Table 4 in Equation (1), the model equation for the prediction of LI has been developed. The following equations can be written to arrive at a correlation of the output parameter with the input parameters.

$$a = -5.433q_c - 2.34R_f - 0.91\sigma_v' + 2.692a_{max} - 3.261 \quad (2)$$

$$b = 5.71q_c - 9.609R_f - 2.659\sigma_v' + 2.383a_{max} - 0.405 \quad (3)$$

$$c = 2.535q_c - 4.33R_f - 3.99\sigma_v' - 1.471a_{max} - 1.915 \quad (4)$$

$$d = -8.225q_c - 6.036R_f + 14.718\sigma_v' + 12.681a_{max} + 7.933 \quad (5)$$

All the values of input parameters are needed to be normalized in the range of [-1,1] before substituting in the above Equations (2) to (5) using Equation (9) and the limits of the training data presented in Table 2.

$$x = -3.261\tanh(a) - 0.405\tanh(b) - 1.915\tanh(c) + 7.933\tanh(d) - 8.69 \quad (6)$$

$$LI\ (Normalized) = \tanh(x) \quad (7)$$

The Equation (7) has been de-normalized to Equation (8) using Equation (9) and the limits of LI i.e., [0,1].

$$LI = 0.5\{\tanh(x) + 1\} \quad (8)$$

If, $LI \geq 0.5$, Liquefaction Index should be considered as 1 which means the soil layer may liquefy at the concerned depth.
If, $LI < 0.5$, Liquefaction Index should be considered as 0 which means the soil layer may not liquefy at the concerned depth.

Normalization Equation. The equation for normalization is given as

$$X_n = \frac{2(X - X_{\min})}{X_{max} - X_{min}} - 1 \quad (9)$$

Where,
X_n = Normalized input parameter data in the range of [-1,1].
X = Actual given input parameter data.
X_{max} , X_{min} = Limits of the input parameter.

6 Development of the GEP Model

GEP is a supervised ML technique which is mainly used for regression and classification problems. GEP follows an evolutionary algorithm, it can learn and adapt by altering its shape, size and composition. GEP is related to Genetic Algorithm (GA) and Genetic Programming (GP). The concept of linear chromosome or individual which represents each data has been inherited from GA and the concept of expression trees of various shapes and sizes have been inherited from GP. In this study, the GEP model has been modelled using GeneXproTools 5.0.

The optimum GEP model has been obtained using four genes per chromosome (Each gene represents each attribute). Root Mean Squared Error (RMSE) has been used as the fitness function to select the fittest chromosomes. Addition function has been used to link up all the sub-programs. The optimum GEP model has been obtained by 1,00,000th generation. The model equation for the LI prediction has been obtained by adding up the mathematical expressions obtained by decoding the sub-expression trees.

Table 5. Performance of the GEP model.

	Training Set (156 cases)		**Prediction Rate (%)**
Liquefied cases	Actual	96	94.79
	Predicted by the GEP model	91	
Non-liquefied cases	Actual	60	88.33
	Predicted by the GEP model	53	
	Testing Set (70 cases)		**Prediction Rate (%)**
Liquefied cases	Actual	37	94.59
	Predicted by the GEP model	35	
Non-liquefied cases	Actual	33	87.88
	Predicted by the GEP model	29	

6.1 Expressing the Sub-Expression Trees

The four sub-expression trees (one sub-expression tree per gene) which have been obtained from the GEP model are as follows

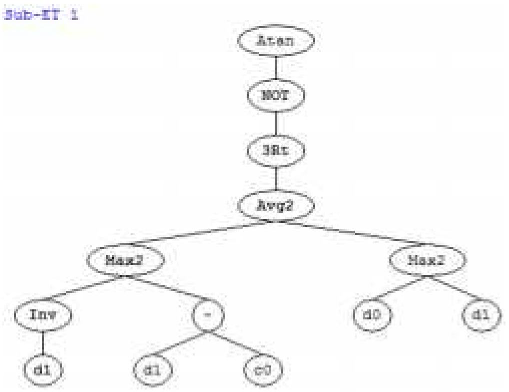


Fig. 3a. Sub-Expression Tree – 1.

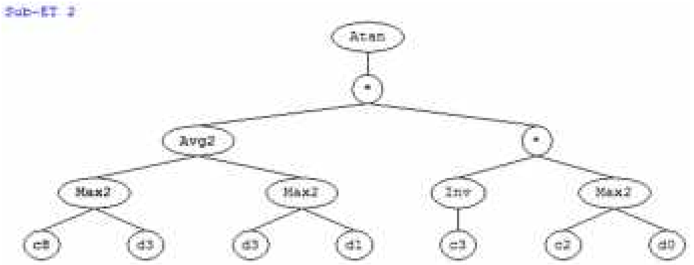


Fig. 3b. Sub-Expression Tree – 2.

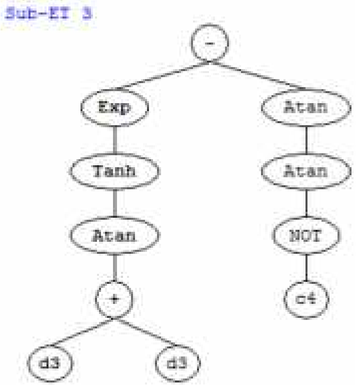


Fig. 3c. Sub-Expression Tree – 3.

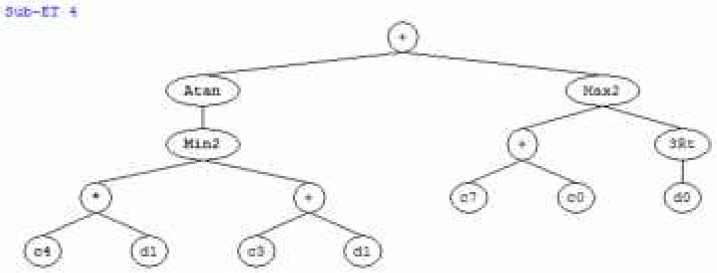


Fig. 3d. Sub-Expression Tree – 4.

Where,

$$\left.\begin{array}{l} \mathrm{d0} = \mathrm{q_c} \\ \mathrm{d1} = \mathrm{R_f} \\ \mathrm{d2} = \sigma'_{\mathrm{v}} \\ \mathrm{d3} = \mathrm{a_{max}} \end{array}\right\} \text{Input Parameters}$$

$$\left.\begin{aligned} \mathrm{G1-C0} &= 1.47\\ \mathrm{G2-C8} &= 0.21\\ \mathrm{G2-C3} &= -1.833\\ \mathrm{G2-C2} &= 4.453\\ \mathrm{G3-C4} &= -1.597\\ \mathrm{G4-C7} &= -3.424\\ \mathrm{G4-C0} &= 6.236\\ \mathrm{G4-C4} &= -2.558\\ \mathrm{G4-C3} &= -6.056 \end{aligned}\right\} \text{Constants obtained from the GEP model}$$

Mathematically expressing the sub-expression trees –

$$G1 = \tan^{-1}\left[1 - \left\{\frac{\max(R_f{}^{-1}, R_f - 1.47) + \max(q_c, R_f)}{2}\right\}^{\frac{1}{3}}\right] \tag{10}$$

$$G2 = \tan^{-1}\left[\left\{\frac{\max(0.21, a_{max}) + \max(a_{max}, R_f)}{2}\right\} * \{-0.545 * \max(4.453, q_c)\}\right] \tag{11}$$

$$G3 = e^{\tanh\{\tan^{-1}(2a_{max})\}} - 2.593 \tag{12}$$

$$G4 = \tan^{-1}\{\min(-2.558R_f, -6.056 + R_f)\} + \max\left(2.812, q_c^{\frac{1}{3}}\right) \tag{13}$$

$$LI = G1 + G2 + G3 + G4 \tag{14}$$

If, $LI \geq 0.5$, Liquefaction Index should be considered as 1 which means the soil layer may liquefy at the concerned depth.
If, $LI < 0.5$, Liquefaction Index should be considered as 0 which means the soil layer may not liquefy at the concerned depth.

7 Comparison with the Robertson and Wride (1998) Method

A comparative study has been carried out with the Robertson and Wride [7] method. Factor of safety against liquefaction ($\frac{CRR_{7.5}}{CSR_{7.5}}$) has been calculated for all the cases of both the sets to classify the liquefied and non-liquefied cases. $\mathrm{CRR}_{7.5}$ values have been calculated in accordance with the guidelines reported by the authors. $\mathrm{CSR}_{7.5}$ values have been calculated by following the guidelines reported by Youd et al. [8].

Table 5. Performance of the Robertson and Wride (1998) Method.

	Training Set (156 cases)		**Prediction Rate (%)**
Liquefied cases	Actual	96	83.33
	Predicted by the Robertson and Wride (1998) Method	80	
Non-liquefied cases	Actual	60	80
	Predicted by the Robertson and Wride (1998) Method	48	
	Testing Set (70 cases)		**Prediction Rate (%)**
Liquefied cases	Actual	37	83.78
	Predicted by the Robertson and Wride (1998) Method	31	
Non-liquefied cases	Actual	33	78.79
	Predicted by the Robertson and Wride (1998) Method	26	

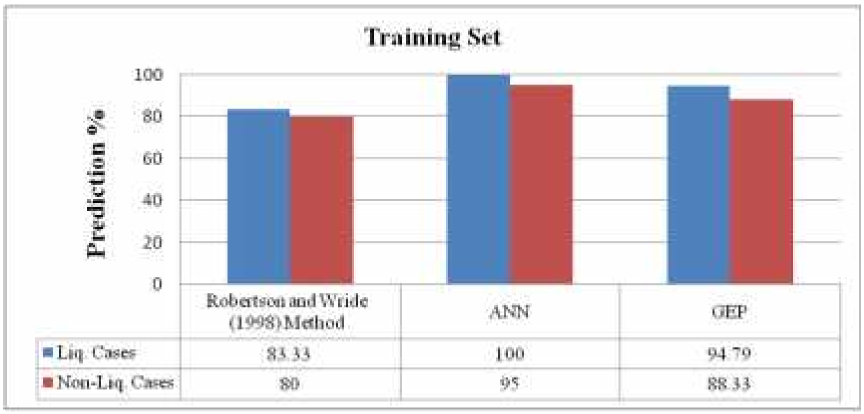
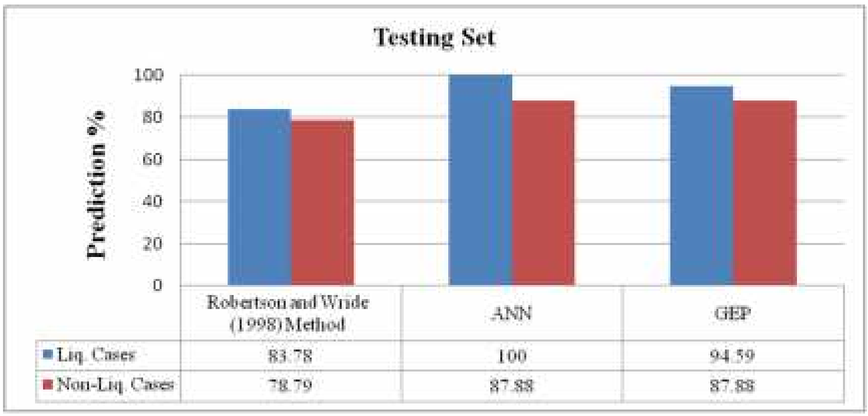


Fig. 4. Comparison of all the models (Training set).



	Robertson and Wride (1998) Method	ANN	GEP
🟦 Liq. Cases	83.78	100	94.59
🟥 Non-Liq. Cases	78.79	87.88	87.88

Fig. 5. Comparison of all the models (Testing set).

8 Sensitivity Analysis

Connection Weight (CW) approach has been adopted as the method for conducting sensitivity analysis for all the cases. In case of CW approach, algebraic operations are done on input-hidden weight matrix and hidden-output weight matrix of the optimum ANN model. After the operations, one gets the value of relative importance of each input parameter and order of importance of input parameters is done accordingly. The details of CW approach can be found in [9].

The order of influence of the input parameters on the predicted LI obtained after carrying out sensitivity analysis using CW approach –

PGA > Effective stress at the boring depth > Friction factor > Cone tip resistance

9 Conclusion

From Fig. 4 and 5, it can be observed that the ANN model has provided the optimum prediction performance; GEP model has also provided satisfactory prediction performance. Both the models – ANN and GEP have performed better than the conventional Robertson and Wride [7] method. Henceforth, it is recommended to use both the ANN and GEP models together with the conventional Robertson and Wride [7] method to be sure about the classification of liquefied and non-liquefied cases.

References

1. Marcuson III WF (1978) Definition of terms related to liquefaction. J Geotech Eng 104(9):1197–1200.
2. Juang CH, Yuan H, Lee DH, Lin PS (2003) Simplified Cone Penetration Test-based method for evaluating liquefaction resistance of soils. J Geotech Geoenviron Eng 129(1):66–80. doi:10.1061/(ASCE)1090-0241(2003)129:1(66)
3. Song Y, Gong J, Gao S, Wang D, Cui T, Li Y, Wei B, (2012) Susceptibility assessment of earthquake-induced landslides using Bayesian network: a case study in Beichuan, China. Computers & Geosciences 42(0):189–199. doi:10.1016/j.cageo.2011.09.011
4. Shahin MA, Holger RM, Jaksa MB (2004) Data division for developing neural networks applied to geotechnical engineering. J Comput Civ Eng 18(2):105–114. doi:10.1061/(ASCE)0887-3801(2004)18:2(105)
5. Hecht-Nielsen R (1988) Theory of the backpropagation neural network. Neural Networks 1(1) 445–445. doi:10.1016/0893-6080(88)90469-8
6. Goh ATC, Kulhawy FH, Chua CG (2005) Bayesian neural network analysis of undrained side resistance of drilled shafts. J Geotech Geoenviron Eng 131(1):84–93. doi:10.1061/(ASCE)1090-0241(2005)131:1(84)
7. Robertson P K, Wride CE (1998) Evaluating cyclic liquefaction potential using the cone penetration test. Can Geotech J 35(3):442–459. doi:10.1139/t98-017
8. Youd TL, Idriss IM, Andrus RD, Arango I, Castro G, Christian JT, Dobry R, Finn WDL, Harder LF, Jr. Hynes ME, Ishihara K, Koester JP, Liao SSC, Marcuson III WF, Martin, Mitchell JK, Moriwaki Y, Power MS, Robertson PK, Seed RB, Stokoe II KH (2001) Liquefaction resistance of soils: summary report from the 1996 NCEER and 1998 NCEER/NSF workshops on evaluation of liquefaction resistance of soils. J Geotech Geoenviron Eng 127(10) 817-833. doi:10.1061/(ASCE)1090-0241(2001)127:10(817)
9. Olden JD, Jackson DA (2002) Illuminating the “blackbox”: understanding variable contributions in artificial neural networks. Ecological Modelling 154(1-2):135–150. doi:10.1016/S0304-3800(02)00064-9