INTERNATIONAL JOURNAL FOR NUMERICAL AND ANALYTICAL METHODS IN GEOMECHANICS
Int. J. Numer. Anal. Meth. Geomech. 2010; **34**:755–770
Published online 18 September 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/nag.837

Site characterization model using least-square support vector machine and relevance vector machine based on corrected SPT data (N_c)

Pijush Samui*,†,‡ and T. G. Sitharam§

Department of Civil Engineering, Indian Institute of Science, Bangalore 560 012, India

SUMMARY

Statistical learning algorithms provide a viable framework for geotechnical engineering modeling. This paper describes two statistical learning algorithms applied for site characterization modeling based on standard penetration test (SPT) data. More than 2700 field SPT values (N) have been collected from 766 boreholes spread over an area of 220 sqkm area in Bangalore. To get N corrected value (N_c), N values have been corrected (N_c) for different parameters such as overburden stress, size of borehole, type of sampler, length of connecting rod, etc. In three-dimensional site characterization model, the function $N_c = N_c\ (X, Y, Z)$, where X, Y and Z are the coordinates of a point corresponding to N_c value, is to be approximated in which N_c value at any half-space point in Bangalore can be determined. The first algorithm uses least-square support vector machine (LSSVM), which is related to a ridge regression type of support vector machine. The second algorithm uses relevance vector machine (RVM), which combines the strengths of kernel-based methods and Bayesian theory to establish the relationships between a set of input vectors and a desired output. The paper also presents the comparative study between the developed LSSVM and RVM model for site characterization. Copyright © 2009 John Wiley & Sons, Ltd.

Received 23 December 2008; Revised 24 June 2009; Accepted 31 July 2009

KEY WORDS: site characterization; SPT; statistical learning algorithm; least-square support vector machine; relevance vector machine

1. INTRODUCTION

One of the first most important steps in geotechnical engineering is site characterization. The objective of site characterization is to provide sufficient, reliable information and data on the site condition to a level, compatible and consistent with the needs and requirements of the project.

[*] Correspondence to: Pijush Samui, Department of Civil Engineering, Indian Institute of Science, Bangalore 560 012, India.

[†] E-mail: pijush.phd@gmail.com

[‡] Research Scholar.

[§] Professor.

Geotechnical engineers use *in situ* tests for site characterization. It is never possible to know the detailed geotechnical properties at every location beneath an actual site because, in order to do so, one would need to sample and/or test the entire subsurface profile. Thus, one has to predict the geotechnical properties at any point of a site based on a limited number of tests. Prediction of geotechnical properties of a site is a difficult task due to uncertainty [1]. The goal in site characterization program is to estimate the subsurface soil properties, based on a limited number of tests. Conventionally, geotechnical engineer characterizes a site based on limited results and interprets them in terms of subsurface soil profiles. These profiles represent a simplified model of the geology, soil layers and their *in situ* soil properties. The task of establishing a subsurface profile requires generalization of soil properties based on limited test data.

Based on finite set of *in situ* data, in probabilistic site characterization, random field theory has been used by many researchers [2–14]. Geostatistics and kriging [15, 16] have also been used to model spatial variation of soil properties [17–21]. However, the literature on three-dimensional site characterizations using geostatistics is not available [22]. This is quite likely due to different reasons such as: (a) the practical difficulty in conducting soil investigation in different directions; (b) high anisotropy generally observed within horizontal and vertical planes; and (c) lack of proper model to describe the spatial variation of soil properties [22]. Random field and geostatistical methods have been applied in site characterization modeling with limited success [22]. Recently, artificial neural network (ANN) has been used for site characterization [22]. A major disadvantage of ANN models is that, unlike other statistical models, they provide no information about the relative importance of the various parameters [23]. In ANN, as the knowledge acquired during training is stored in an implicit manner, it is very difficult to come up with reasonable interpretation of the overall structure of the network [24]. This lead to the term 'black box' in which many researchers use while referring to ANNs behavior. In addition, ANN has some inherent drawbacks such as slow convergence speed, less generalizing performance, arriving at local minimum and overfitting problems.

In this paper, least-square support vector machine (LSSVM) and relevance vector machine (RVM) models have been used to characterize the site of Bangalore based on large data set of corrected standard penetration test (SPT) values (N_c). Relative to Vapnik's [25] support vector machine (SVM), the LSSVM can transform a quadratic programming problem into a linear programming problem, thus reducing the computational complexity. It is closely related to Gaussian processes and regularization networks. It requires solving a set of only linear equations (linear programming), which is much easier and computationally very simple. RVMs adopt a Bayesian extension of learning. RVMs allow computation of the prediction intervals taking uncertainties of both the parameters and the data [26]. The aim of the paper is as follows:

1. To investigate the feasibility of LSSVM and RVM model for the prediction of N_c at any point in the three-dimensional subsurface of Bangalore.
2. To perform a comparative study between developed LSSVM and RVM model.

2. SITE DESCRIPTION

Bangalore covers an area of over 220 sqkm and with large variation in ground reduced levels. It varies from 810 m in northeast part to 940 m in southwestern part of Bangalore. There were more than 400 lakes once upon a time, and more than 340 lakes are filed up due to erosion and encroachments for the construction of layouts and buildings. The population of greater Bangalore region is over 6 million and it is the fifth biggest city in India. It is situated at the latitude of 12°8′ North and longitude of 77°37′ East.

From geologically, most part of the Bangalore region consists of Gneiss complexes, which is formed due to several tectonic-thermal events with large influx of sialic material, and are believed to have occurred between 3400 and 3000 million years ago giving rise to an extensive group of gray gneisses designated as the 'older gneiss complex'. These gneisses act as the basement for a widespread belt of schist's. The younger group of gneissic rocks mostly of granodiomitic and granitia composition is found in the eastern part of the city, representing remobilized parts of an older crust with abundant additions of newer granite material, for which the name 'younger gneiss complex' has been given [27]. The soil is mostly a residual soil from granite gneiss due to weathering action.

3. GEOGRAPHIC INFORMATION SYSTEM (GIS) MODEL AND GEOTECHNICAL DATA

The Bangalore map forms the base layer for the development of GIS model (see Figure 1). The map entities have been developed in view of two aspects: first for locating the bore logs to the utmost accuracy on a scale of 1:20000 and second for the identification of bore logs by the end user. The digitized map has several layers of information. Some of the important layers considered are the boundaries (outer and administrative), highways, major roads, minor roads, streets, rail roads, water bodies, drains, ground contours and borehole locations. A large amount of geotechnical data consisting of 766 boreholes have been collated along with index and engineering properties of subsoil layers at different locations in Bangalore (location of boreholes is shown in Figure 1). In total, 766 bore log information has been entered into the database using a GIS with ARCINFO package. The latitudes and longitudes were confirmed using global positioning system stations at selected locations. In total, 2722 'N' values are available in the 766 boreholes in the three-dimensional GIS model. Distribution of collected boreholes in Bangalore is shown in Figure 2, indicating a very good distribution of the boreholes in each quadrant of Bangalore from the city center. Figure 1 depicts a grid of 1 km × 1 km within the corporate boundary of Bangalore along with outer boundary circumscribing the ring road. It gives a clear view of the spatial distribution of boreholes in the Bangalore region. An average of about four boreholes data is available within the grid of 1 km × 1 km.

Geotechnical data was collected from archives of Torsteel Research Foundation in India and Indian Institute of Science for geotechnical investigation carried out for several major projects in Bangalore. The data collected being for important projects are of very high quality in Bangalore during the years 1995–2003. The data in the model are on average to a depth of 30 m below the ground level. This bore log contains information about the depth, density of the soil, total stress, effective stress, fines content and N values and depth of ground water table. Figure 3 shows the typical bore logs along a section A–A. For typical soil profiles for the purpose of general identification of soil layers, the Bangalore map area is divided into four parts (four quadrants) in North–South and East–West directions. The typical soil profile in the Northwestern part of the Bangalore has three layers of soil deposition. The top layer contains brownish silty sand with clay up to 3 m, after which up to 6 m, medium dense to very dense silty sand is present. The third layer has weathered rock varying from 6 to 17 m depth and followed by hard rock. The Southwestern part contains red soil or reddish silty sand with gravel up to 1.7 m depth, yellowish clayey sand

DOI: 10.1002/nag

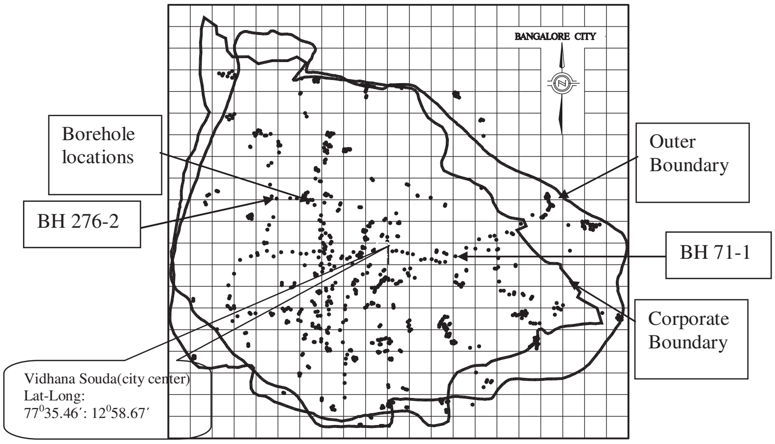

Figure 1. Borehole locations in Bangalore map (scale: 1:20000).

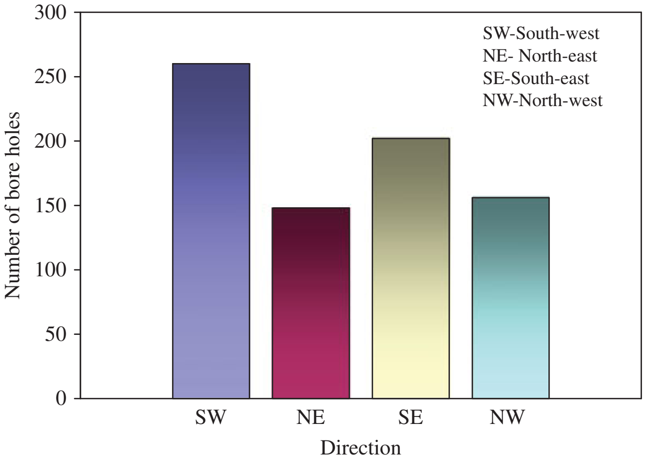

Figure 2. Distribution of boreholes in quadrants for Bangalore.

from 1.7 to 3.5 m, yellowish silty sand with clay from 3.5 to 8.5 m and hard rock below 8.5 m. The soil in the Southeastern part can be classified into four layers. The first layer up to 1.5 m contains brownish clayey sand, brownish clayey sand with gravel from 1.5 to 4 m, yellowish silty sand with gravel up to 5.5 m, different stages of weathered rock from 5.5 to 17.5 m and hard rock beneath. Northeastern side has four layer depositions, filled up soil to 1.5 m, reddish silty clay from 1.5 to 4.5 m, sandy clay up to 7.5 m, weathered rock form 7.5 to 18.5 m and hard rock below. The corrections for field N values (shown in Tables I and II) are applied for overburden pressures (CN), hammer energy (CE), borehole diameter (CB), presence or absence of liner (CS), rod length (CR) and correction for fines content (C_{fines}) as per the standard procedures existing in the literature [29–34].

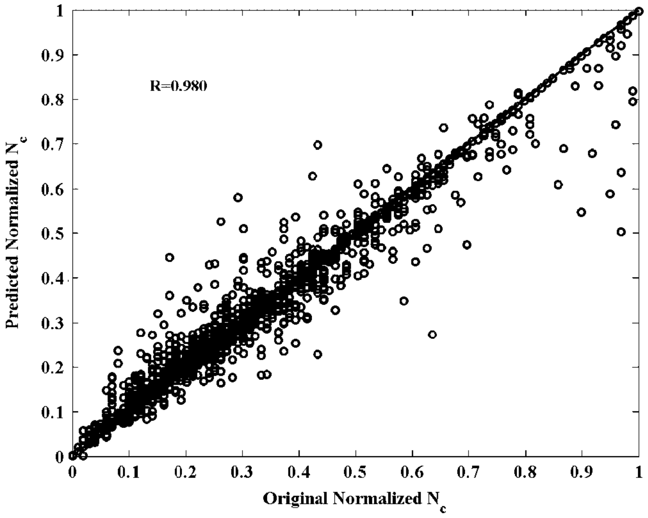

Figure 3. Performance of LSSVM model for training data set.

Table I. Different types of SPT corrections used to obtain N_c values [28].

Different types of correction	Correction factor
Correction for Hammer	
Donut Hammer	0.5–1.0
Safety Hammer	0.7–1.2
Automatic-trip Donut Hammer	0.8–1.3
Sampler correction	
Without liner	1.00
With liner: Dense sand, clay	0.80
Loose sand	0.90
Rod length correction	
length>10 m	1.0
6–10	0.95
4–6	0.85
0–4	0.75
Borehole diameter correction	
Hole diameter 60–120 mm	1.00
150 mm	1.05
200 mm	1.15
Correction for overburden pressure, $\sigma'_{v0}(C_N)$	$\frac{2.2}{\left(1.2+\frac{\sigma'_{v0}}{\mathrm{Pa}}\right)}$, where Pa = 100 kPa
Correction for fines content (C_{fines})	$1+0.004\mathrm{FC}+0.055\left(\frac{\mathrm{FC}}{N_{60}}\right)$, where FC is the percent fines content (percent dry weight finer than 0.074 mm) and N_{60} is the SPT value for 60% energy ratio

Table II. Typical calculation showing N_c values obtained from N values.

						Correction factor						
Depth (m)	SPT value (N_{field})	Unit weight (kN/m^3)	Total stress [T.S. (kN/m^2)]	Effective stress (kN/m^2)	C_N	Hammer effect	Borehole diameter	Rod length	Sample method	FC	C_{fines}	SPT value ($N_{\text{corrected}}$)
1.50	11	19.10	28.65	28.65	1.48	1	1.05	0.75	1	54.5	1.431	18
4.00	12	19.10	76.40	76.40	1.12	1	1.05	0.85	1	59	1.482	18
6.00	30	19.10	114.60	114.60	0.94	1	1.05	0.95	1	53.6	1.309	37
7.50	74	19.10	143.25	143.25	0.84	1	1.05	0.95	1	29.8	1.143	71

4. METHODOLOGY

In this paper, two models (LSSVM and RVM) have been adopted for the prediction of N_c in the subsurface of Bangalore. Brief descriptions of these two models are given as follows:

4.1. LSSVM model

LSSVM models are an alternate formulation of SVM regression [35] proposed by Suykens *et al.* [36]. Consider a given training set of N data points $\{x_k, y_k\}_{k=1}^{N}$ with input data $x_k \in R^N$ and output $y_k \in r$, where R^N is the N-dimensional vector space and r is the one-dimensional vector space. The three input variables used for the LSSVM for the prediction of N_c in the three-dimensional subsurfaces are X, Y, Z, where X, Y, Z are the coordinates of borehole in Bangalore. The output of the LSSVM model is N_c. Thus, in this study, $x=[X, Y, Z]$ and $y=N_c$. In feature space LSSVM models take the form

$$y(x)=w^{\mathrm{T}}\varphi(x)+b \tag{1}$$

where the nonlinear mapping $\varphi(\cdot)$ maps the input data into a higher-dimensional feature space; $w \in R^n$; $b \in r$; w is an an adjustable weight vector; b the scalar threshold. In LSSVM for function estimation, the following optimization problem is formulated:

$$\begin{aligned}&\text{Minimize:} && \frac{1}{2}w^{\mathrm{T}}w+\gamma\frac{1}{2}\sum_{k=1}^{N}e_k^2\\ &\text{Subjected to:} && y(x)=w^{\mathrm{T}}\varphi(x_k)+b+e_k, \quad k=1,\ldots,N\end{aligned} \tag{2}$$

where γ is the regularization parameter, determining the tradeoff between the fitting error minimization and smoothness, e_k is the error variable.

The Lagrangian $(L(w,b,e;\alpha))$ for the above optimization problem (2) is

$$L(w,b,e;\alpha)=\frac{1}{2}w^{\mathrm{T}}w+\gamma\frac{1}{2}\sum_{k=1}^{N}e_k^2-\sum_{k=1}^{N}\alpha_k\{y_k[w^{\mathrm{T}}\varphi(x_k)+b]-1+e_k\} \tag{3}$$

with α_k Lagrange multipliers. The conditions for optimality are given by

$$\begin{aligned}&\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{k=1}^{N}\alpha_k\varphi(x_k)\\ &\frac{\partial L}{\partial b}=0\Rightarrow\sum_{k=1}^{N}\alpha_k=0\\ &\frac{\partial L}{\partial e_k}=0\Rightarrow\alpha_k=\gamma e_k, \quad k=1,\ldots,N\\ &\frac{\partial L}{\partial \alpha_k}=0\Rightarrow w^{\mathrm{T}}\varphi(x_k)+b+e_k-y_k=0, \quad k=1,\ldots,N\end{aligned} \tag{4}$$

After the elimination of e_k and w, the solution is given by the following set of linear equations as:

$$\begin{bmatrix} 0 & 1^{\mathrm{T}} \\ 1 & \Omega+\gamma^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix}=\begin{bmatrix} 0 \\ y \end{bmatrix} \tag{5}$$

where $y=[y_1,\ldots,y_N]$, $1=[1,\ldots,1]$, $\alpha=[\alpha_1,\ldots,\alpha_N]$ and Mercer's theorem [35, 37] is applied within the Ω matrix, $\Omega=\varphi(x_k)^{\mathrm{T}}\varphi(x_l)=k(x_k,x_l)$, $k,l=1,\ldots,N$, where $k(x_k,x_l)$ is the kernel function. Choosing $\gamma>0$, ensures that the matrix

$$\Phi=\begin{bmatrix} 0 & 1^{\mathrm{T}} \\ 1 & \Omega+\gamma^{-1}I \end{bmatrix}$$

is invertible. Then the analytical of α and b is given by

$$\begin{bmatrix} b \\ \alpha \end{bmatrix}=\Phi^{-1}\begin{bmatrix} 0 \\ y \end{bmatrix} \tag{6}$$

The Gaussian kernel has been used in this analysis. The Gaussian kernel is given by

$$K(x_k,x_l)=\exp\left\{-\frac{\|x_k-x_l\|}{2\sigma^2}\right\}, \quad k,l=1,\ldots,N \tag{7}$$

where σ is the width of Gaussian kernel.

The resulting LSSVM model for the prediction then becomes

$$y(x)=\sum_{k=1}^{N}\alpha_k K(x,x_k)+b \tag{8}$$

The above-mentioned LSSVM methodology has been implemented to predict N_c in the three-dimensional subsurface of Bangalore. For SVM, each of the input variables (X, Y and Z) is first normalized with respect to their respective maximum value. The output variable N_c was also normalized with respect to the maximum N_c value. In LSSVM modeling, the data have been divided into two subsets: a training data set, to construct the model, and a testing data set to estimate the model performance. Thus, the N_c has been divided into training and testing data sets using sorting method to maintain statistical consistency. For our study only 30% of the total boreholes selected randomly are considered as testing data set, which consists of 493 N_c values. Gaussian kernel has been used in this analysis. The remaining 70% boreholes are considered as the training data set. The LSSVM program is constructed using MATLAB [38].

4.2. RVM model

The RVM, introduced by Tipping [38], is a sparse linear model. Let $D=\{(x_i,t_i), i=1,\ldots,N\}$ be a data set of observed values, where x_i is the input, t_i the output, $x_i\in R^d$ and $t_i\in R$. In this study, the input parameters for predicting N_c values are X, Y and Z. Thus, $x=[X,Y,Z]$. The output of the RVM model is N_c. Thus, $y=[N_c]$. One can express the output as the sum of an approximation vector $y=(y(x_1),\ldots,y(x_N))^{\mathrm{T}}$, and zero mean random error (noise) vector $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_N)^{\mathrm{T}}$, where $\varepsilon_n\sim\mathbf{N}(0,\sigma^2)$ and $\mathbf{N}(0,\sigma^2)$ is the normal distribution with mean 0 and variance σ^2. Thus, the output can be written as

$$t_n=y(x_n,\omega)+\varepsilon_n \tag{9}$$

where ω is the parameter vector. Assuming that

$$p(t_n|x) \sim \mathbf{N}(y(x_n), \sigma^2) \tag{10}$$

where $\mathbf{N}(y(x_n), \sigma^2)$ is the normal distribution with mean $y(x_n)$ and variance σ^2. $y(x)$ can be expressed as a linearly weighted sum of M nonlinear fixed basis function

$$\{\Phi_j(x)|j=1,\dots,M\}:\ y(x;\omega)=\sum_{i=1}^{M}\omega_i\Phi_i(x)=\mathbf{\Phi}\omega \tag{11}$$

The likelihood of the complete data set can be written as

$$p(t|w,\sigma^2)=(2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\|t-\Phi w\|^2\right\} \tag{12}$$

where $t=(t_1,\dots,t_N)^{\mathrm{T}}$, $\omega=(\omega_0,\dots,\omega_N)$ and

$$\Phi^{\mathrm{T}}=\begin{bmatrix} 1 & K(x_1,x_1) & K(x_1,x_2) & \dots & K(x_1,x_n) \\ 1 & K(x_1,x_2) & K(x_2,x_2) & \dots & K(x_2,x_n) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & K(x_n,x_1) & K(x_n,x_2) & \dots & K(x_n,x_n) \end{bmatrix}$$

where $k(x_i,x_n)$ is a kernel function.

To prevent overfitting, automatic relevance detection prior is set over the weights

$$p(w|\alpha)=\prod_{i=0}^{N}N(\omega_i|0,\alpha_i^{-1}) \tag{13}$$

where α is a hyperparameter vector that controls how far from zero each weight is allowed to deviate (36). Consequently, using Baye's rule, the posterior overall unknowns could be computed given the defined non-informative prior distribution:

$$p(w,\alpha,\sigma^2/t)=\frac{p(y/w,\alpha,\sigma^2)\cdot p(w,\alpha,\sigma)}{\int p(t/w,\alpha,\sigma^2)p(w,\alpha,\sigma^2)\,\mathrm{d}w\,\mathrm{d}\alpha\,\mathrm{d}\sigma^2} \tag{14}$$

Full analytical solution of this integral (6) is obdurate. Thus, the decomposition of the posterior according to $p(w,\alpha,\sigma^2/t)=p(w/t,\alpha,\sigma^2)p(\alpha,\sigma^2/t)$ is used to facilitate the solution [38]. The posterior distribution over the weights is thus given by:

$$p(w/t,\alpha,\sigma^2)=\frac{p(t/w,\sigma^2)\cdot p(w/\alpha)}{p(t/\alpha,\sigma^2)} \tag{15}$$

The resulting posterior distribution over the weights is the multivariate Gaussian distribution

$$p(w/t,\alpha,\sigma^2)=\mathbf{N}\left(\mu,\sum\right) \tag{16}$$

where the mean and the covariance are, respectively, given by:

$$\sum=(\sigma^{-2}\Phi^{\mathrm{T}}\Phi+A)^{-1} \tag{17}$$

$$\mu=\sigma^{-2}\sum\Phi^{\mathrm{T}}t \tag{18}$$

with diagonal $A=\mathrm{diag}(\alpha_0,\dots,\alpha_N)$.

DOI: 10.1002/nag

For uniform hyperpriors over α and σ^2, one needs to maximize the term $p(t/\alpha,\sigma^2)$:

$$p(t/\alpha,\sigma^2)=\int p(t/w,\sigma^2)p(w/\alpha)\,\mathrm{d}w=((2\pi)^{-N/2}/\sqrt{|\sigma^2+\Phi A^{-1}\Phi^{\mathrm{T}}|})$$
$$\times\exp\left\{-\frac{1}{2}y^{\mathrm{T}}(\sigma^2+\Phi A^{-1}\Phi^{\mathrm{T}})^{-1}y\right\} \qquad (19)$$

Maximization of this quantity is known as the type II maximum likelihood method [39, 40] or the 'evidence for hyper parameter' [41]. Hyper parameter estimation is carried out in iterative formulae, for example, gradient descent on the objective function [38]. The outcome of this optimization is that many elements of α go to infinity such that w will have only a few nonzero weights that will be considered as relevant vectors.

This study also investigates the feasibility above RVM model to predict N_c in three-dimensional subsurface of Bangalore. The training data set, testing data set, kernel and normalization technique for developed RVM model is the same as used in LSSVM model. The RVM model is constructed in MATLAB.

5. RESULTS AND DISCUSSION

For LSSVM training, the design value of γ and the width of Gaussian kernel (σ) have been chosen by trial-and-error approach. The design values of γ and σ are 200 and 4.6, respectively. A large γ assigns higher penalties to errors so that the LSSVM is trained to minimize the error with lower generalization, whereas a small γ assigns a fewer penalties to errors; this allows the minimization of margin with errors, thus higher generalization ability. If γ goes to infinitely large, SVM would not allow the occurrence of any error and result in a complex model, whereas when γ goes to zero, the result would tolerate a large amount of errors and the model would be less complex. A large σ indicates a stronger smoothing of Gaussian kernel. The performance of LSSVM model for training data set has been obtained by using the design value of γ and σ. Figure 3 shows the performance of training data set. The coefficient of correlation (R) for training data set is 0.980 and it is close to one. Thus, LSSVM model has captured the input–output relationship very well for training data set. Now, the performance of developed LSSVM model has been examined for testing data set. Figure 4 shows the performance of LSSVM model for testing data set. From Figure 4, it is clear that the LSSVM model has the capability to predict N_c value at any point in the three-dimensional subsurface in Bangalore. Figures 5 and 6 show three-dimensional and two-dimensional surface of N_c using LSSVM model, respectively. They depict the variability of N_c data with spatial coordinate as well as depth in the Bangalore.

For RVM model, the design value of σ has been chosen by trial-and-error approach during training. A large σ indicates a stronger smoothing of Gaussian kernel. The design value of σ is 1.5 and number of relevance vectors is 635. Figure 7 shows the performance training data set for RVM model. It could be observed that RVM model captures the input–output relation very well. Figure 8 shows the performance of testing data set for RVM model. The R value (0.982) of testing data set is close to one. Thus, the RVM model can be used to predict N_c value at any point in the three-dimensional subsurface of Bangalore. Figures 9 and 10 show the three-dimensional and two-dimensional surface of N_c using RVM model, respectively, that depicts the variability of N_c data with spatial coordinate as well as depth in the Bangalore.

DOI: 10.1002/nag

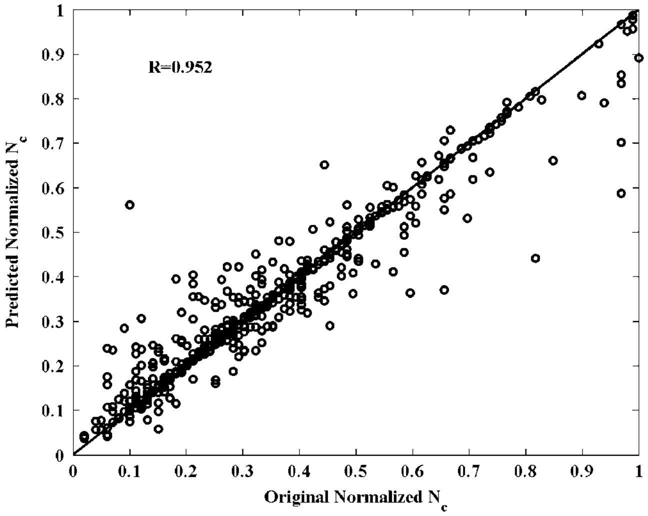

Figure 4. Performance of LSVM model for testing data set.

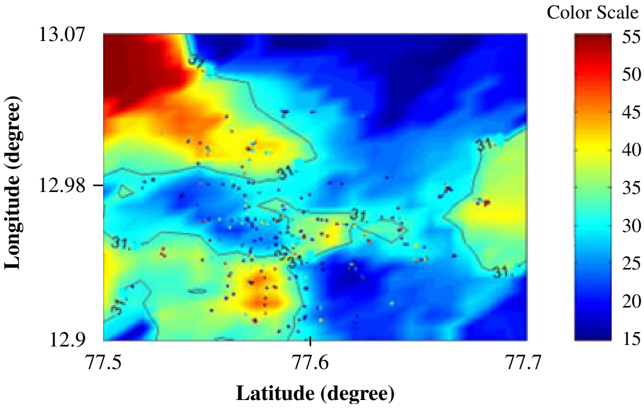

Figure 5. Two-dimensional surface of N_c on the plane of $z=1.5\,\text{m}$ using LSSVM model.

Two testing boreholes have been chosen in Bangalore for verifying the developed model results as a function of depth. For boreholes BH-71-1 and BH-276-2 (locations are shown in Figure 1), N_c values have been predicted by LSSVM as well as RVM models with depth. Figures 11 and 12 show that these N_c profiles with depth correspond to borhole nos BH 71-1 and BH 276-2, respectively. It can be seen that the performance of developed RVM model is slightly better than the LSSVM model. RVM model requires smaller tuning parameter (one parameter, i.e. σ) compared with LSSVM model (two parameters, i.e. γ and σ). There is a slight marginal reduction in the performance on the testing data set (i.e. there is a difference between machine performance on training and testing) for the LSSVM as well as RVM model. Thus, the developed LSSVM and RVM model has the capability to avoid overtraining. RVM model employs 28.48% of the training

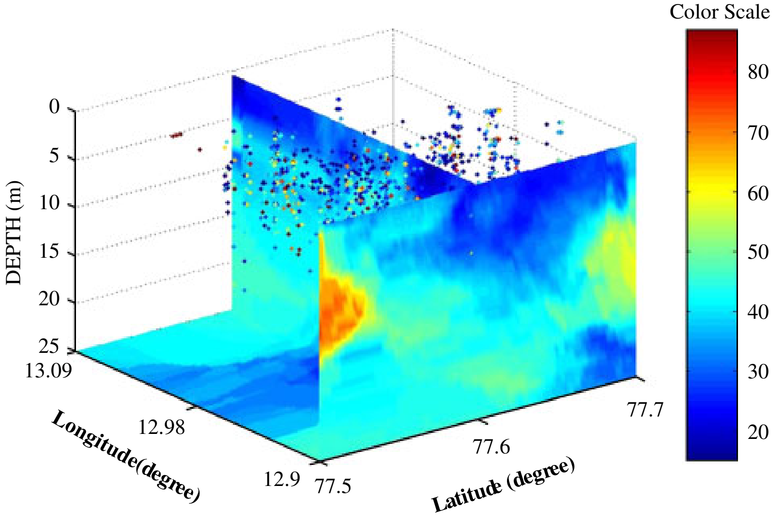

Figure 6. Three-dimensional surface of N_c using LSSVM model.

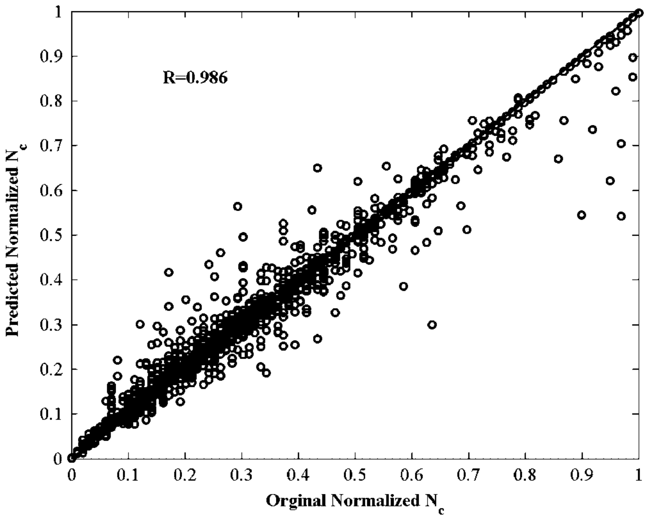

Figure 7. Performance of RVM model for training data set.

data as relevance vectors. The relevance vectors in the RVM model exhibit the essential features of the information content of the data thereby resulting in a sparse model in which most of the weights are equal to zero apart from the relevance vector data. The model thereby obtained is compact, computationally efficient and simple. In addition, the model also produces smooth functions. RVM model uses these relevance vectors for the final prediction, but LSSVM model uses all training data for final prediction. Therefore, LSSVM model does not exhibit sparse solution. It is well

DOI: 10.1002/nag

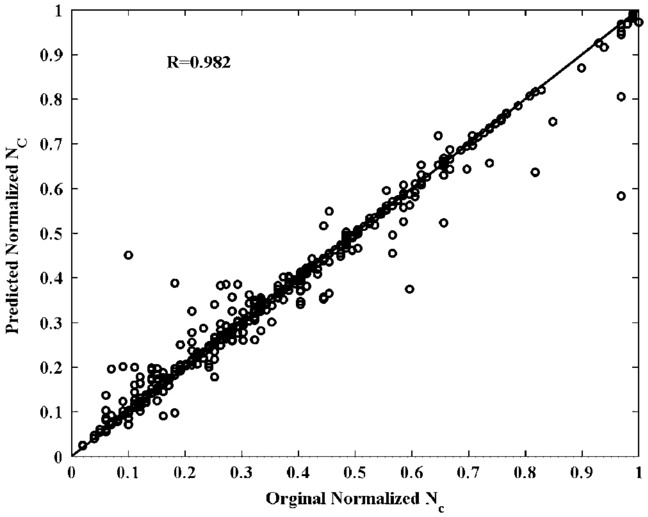

Figure 8. Performance of RVM model for testing data set.

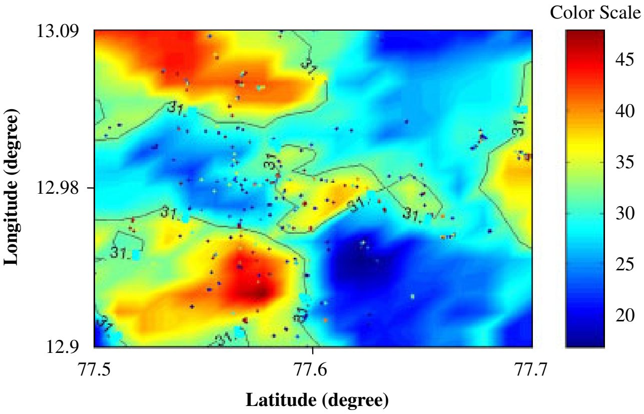

Figure 9. Two-dimensional surface of N_c on the plane of $z=1.5$ m using RVM model.

known that the efficiency of statistical learning algorithm depends on the quality and quantity of the data. Figures 3 and 7 depict the higher value of N_c, and the difference between the predicted and the actual values are very large. The models are built up by using training data set. It is quite expected that the developed models show same behavior (as in training data set) for testing data set. For this reason, Figures 4 and 8 also depict the higher value of N_c and the difference between the predicted the and actual values are very large.

DOI: 10.1002/nag

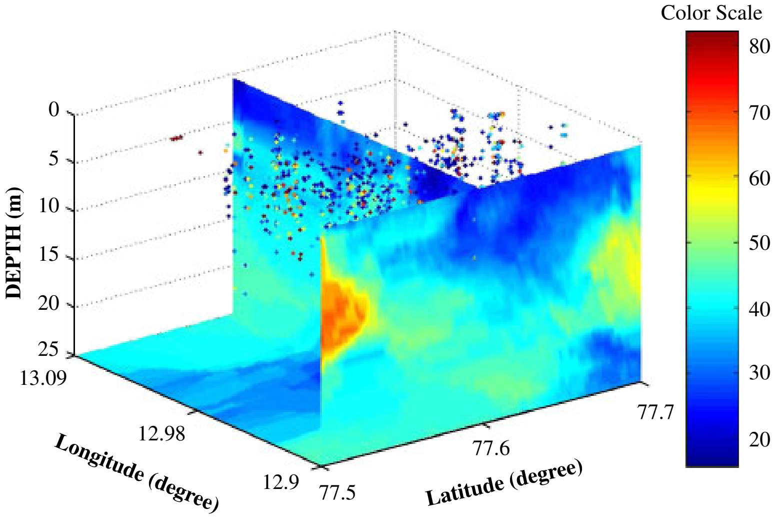

Figure 10. Three-dimensional surface of N_c using RVM model.

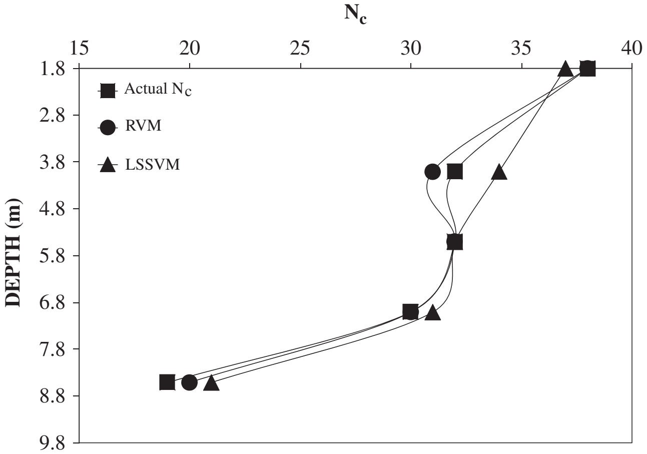

Figure 11. Comparison of N_c with predicted values at BH 276-2 by LSSVM and RVM.

6. CONCLUSION

This study investigates the feasibility of LSSVM and RVM model to predict N_c value at any point in the three-dimensional subsurface of Bangalore. Both models give promising result. The LSSVM was found to generalize well by setting the γ as 150 and σ value as 0.002. RVM model exploits only the set of observations that contain all the information necessary for defining the final decision surface. Once both the models are developed and trained, they require only a small

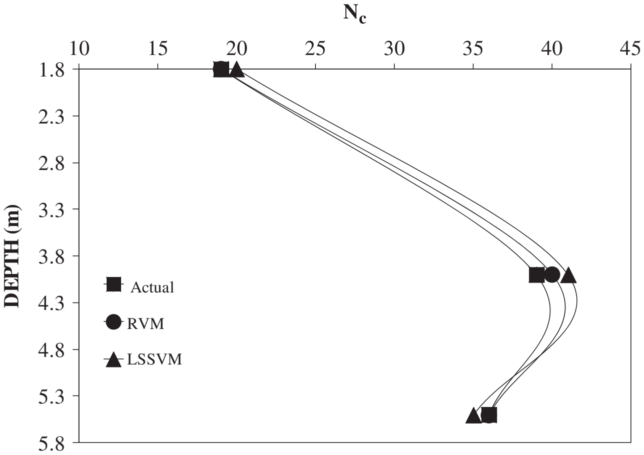

Figure 12. Comparison of N_c with predicted values at BH 71-1 by LSSVM and RVM.

amount of computational time and provide promising results. Comparison between the LSVM and RVM model indicates that the RVM model is superior to LSSVM model for predicting N_c values.

REFERENCES

1. Baecher GB. Geotechnical error analysis. *Transactions on the Research Record*, *No. 1105*, Washington, DC, 1986; 23–31.
2. Yaglom AM. *Theory of Stationary Random Functions*. Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1962.
3. Lumb P. Spatial variability of soil properties. *Proceedings of the Second International Conference on Application of Statistics and Probability in Soil and Structural Engineering*, Aachen, Germany, 1975; 397–421.
4. Vanmarcke EH. Probabilistic modeling of soil profiles. *Journal of Geotechnical Engineering* (ASCE) 1977; **102**(11):1247–1265.
5. Tang WH. Probabilistic evaluation of penetration resistance. *Journal of Geotechnical Engineering* (ASCE) 1979; **105**(GT10):1173–1191.
6. Wu TH, Wong K. Probabilistic soil exploration: a case history. *Journal of Geotechnical Engineering* (ASCE) 1981; **107**(GT12):1693–1711.
7. Asaoka A, Grivas DA. Spatial variability of the undrained strength of clays. *Journal of Geotechnical Engineering* (ASCE) 1982; **108**(5):743–745.
8. Vanmarcke EH. *Random Fields*: *Analysis and Synthesis*. The MIT Press: Cambridge, MA, 1983.
9. Baecher GB. On estimating auto-covariance of soil properties. *Specialty Conference on Probabilistic Mechanics and Structural Reliability*, vol. 110. ASCE: New York, 1984; **110**:214–218.
10. Kulatilake PHSW, Miller KM. A scheme for estimating the spatial variation of soil properties in three dimensions. *Proceedings of the Fifth International Conference on Application of Statistics and Probabilities in Soil and Structural Engineering*, Vancouver, BC, Canada, 1987; 669–677.
11. Kulatilake PHSW. Probabilistic potentiometric surface mapping. *Journal of Geotechnical Engineering* (ASCE) 1989; **115**(11):1569–1587.
12. Fenton GA. Random field characterization NGES data. *Paper Presented at the Workshop on Probabilistic Site Characterization at NGES*, Seattle, Washington, 1998.
13. Phoon KK, Kulhawy FH. Characterization of geotechnical variability. *Canadian Geotechnial Journal* 1999; **36**(4):612–624.
14. Uzielli M, Vannucchi G, Phoon KK. Random filed characterization of stress-normalized cone penetration testing parameters. *Géotechnique* 2005; **55**(1):3–20.
15. Matheron G. Principles of geostatistics. *Economic Geology* 1963; **58**:246–266.

16. Journel AG, Huijbregts CJ. *Mining Geostatistics*. Academic Press: New York, 1978.
17. Kulatilake PHSW, Ghosh A. An investigation into accuracy of spatial variation estimation using static cone penetrometer data. *Proceedings of the First International Symposium on Penetration Testing*, Orlando, FL, 1988; 815–821.
18. Kulatilake PHSW. Probabilistic potentiometric surface mapping. *Journal of Geotechnical Engineering* (ASCE) 1989; **115**(11):1569–1587.
19. Chiasson P, Lafleur J, Soulie M, Law KT. Characterizing spatial variability of clay by geostatistics. *Canadian Geotechnical Journal* 1995; **32**:1–10.
20. Soulie M, Montes P, Sivestri V. Modelling spatial variability of soil parameters. *Canadian Geotechnical Journal* 1990; **27**:617–630.
21. Degroot DJ. Analyzing spatial variability of in situ soil properties. *ASCE Proceedings of Uncertainty'96, Uncertainty in the Geologic Environment: From Theory to Practice*, vol. 58. ASCE Geotechnical Special Publications: New York, 1996; 210–238.
22. Juang CH, Jiang T, Christopher RA. Three-dimensional site characterization: neural network approach. *Géotechnique* 2001; **51**(9):799–809.
23. Park D, Rilett LR. Forecasting freeway link ravel times with a multi-layer feed forward neural network. *Computer Aided Civil and Znfa Structure Engineering* 1999; **14**:358–367.
24. Kecman V. *Leaming and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*. MIT Press: Cambridge, MA, London, England, 2001.
25. Vapnik VN. *The Nature of Statistical Learning Theory*. Springer: New York, 1995.
26. Tipping M. The relevance vector machine. In *Advances in Neural Information Processing Systems*, Solla S, Leen T, Muller KR (eds), vol. 12. MIT Press: Cambridge, MA, 2000; 652–658.
27. Radhakrishna BP, Vaidyanadhan R. *Geology of Karnataka*. Geological Society of India: Bangalore, 1997.
28. Bowles Joseph E. *Foundation Analysis and Design* (4th edn). McGraw-Hill: New York, 1988.
29. Schmertmann JH. Statics of SPT. *Journal of Geotechnical Engineering Division* (ASCE) 1979; **105**(5):665–670.
30. Riggs CO. American standard penetration test practice. *14th PSC*, vol. 124. ASCE: New York, 1986; 949–967.
31. Robertson PK, Wride CE. Evaluating cyclic liquefaction potential using the cone penetration test. *Canadian Geotechnical Journal* 1998; **35**(3):442–459.
32. Seed HB, Tokimatsu K, Harder LF, Chung R. Influence of SPT procedures in soil liquefaction resistance evaluation. *Journal of Geotechnical Engineering* (ASCE) 1985; **111**(12):861–878.
33. Finn L, Ventura C. Challenging issues in local microzonation. *Proceedings of the 5th International Conference on Seismic Zonation*, Nice, vol. II, 1995; 1554–1561.
34. Skempton AW. Standard penetration test procedures and the effects in sands of overburden pressure, relative density, particle size, aging and overconsolidation. *Géotechnique* 1986; **36**(3):425–447.
35. Vapnik N. *Statistical Learning Theory*. Wiley: New York, 1998.
36. Suykens JAK, De Barbanter J, Lukas L, Vandewalle J. Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputers* 2002; **8**(1–4):85–105.
37. Smola A, Scholkopf B. On a kernel based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica* 1998; **22**:211–231.
38. Tipping ME. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* 2001; **1**:211–244.
39. Berger JO. *Statistical Decision Theory and Bayesian Analysis* (2nd edn). Springer: New York, 1985.
40. Wahba G. A comparison of GCV and GML for choosing the smoothing parameters in the generalized spline-smoothing problem. *Annals of Statistics* 1985; **4**:1378–1402.
41. MacKay DJ. Bayesian methods for adaptive models. *Ph.D. Thesis*, Department of Computer and Neural Systems, California Institute of Technology, Pasadena, CA, 1992.

DOI: 10.1002/nag