# Model Development and Validation for Intelligent Data Collection for Lateral Spread Displacements

Thomas Oommen[1] and Laurie G. Baise[2]

**Abstract:** The geotechnical earthquake engineering community often adopts empirically derived models. Unfortunately, the community has not embraced the value of model validation, leaving practitioners with little information on the uncertainties present in a given model and the model's predictive capability. In this study, we present a machine learning technique known as support vector regression (SVR) together with rigorous validation for modeling lateral spread displacements and outline how this information can be used for identifying gaps in the data set. We demonstrate the approach using the free face lateral displacement data. The results illustrate that the SVR has relatively better predictive capability than the commonly used empirical relationship derived using multilinear regression. Moreover, the analysis of the SVR model and its support vectors helps in identifying gaps in the data and defining the scope for future data collection.

**DOI:** 10.1061/(ASCE)CP.1943-5487.0000050

**CE Database subject headings:** Earthquake engineering; Validation; Displacement; Data collection; Models.

**Author keywords:** Machine learning; Lateral spreading; Support vector machine; Geotechnical earthquake engineering; *K*-fold cross validation.

## Introduction

Geotechnical earthquake engineering is a field that involves significant uncertainty. For design, the engineer deals mostly with materials and geometries that nature provides and has to infer these natural conditions from limited and costly observations. The principal uncertainties in design have to do with the accuracy and completeness of limited data. As a result of these geotechnical data issues, empirically derived models are common in geotechnical earthquake engineering (e.g., ground motion prediction equations, probabilistic liquefaction potential equations, and lateral spread displacement regression equations). These empirical models take forms ranging from the most common statistical methods such as multilinear regression (MLR) (Bartlett and Youd 1995; Youd et al. 2002) to models based on machine learning such as Bayesian updating (Cetin et al. 2004; Moss et al. 2006) and artificial neural networks (ANNs) (Newmark 1965; Javadi et al. 2006). Although empirical models are common, the geotechnical earthquake engineering community has not embraced model validation. Therefore, there is very little information for practitioners to decide which model to use and what accuracy can be achieved using a particular model. Efforts to advance the field can be characterized by (1) data collection, (2) empirical model development, and (3) theoretical formulations (Towhata et al. 1992). The current approaches to advance the field are independent efforts. However, we believe that to close the current gap between observations and predictions we need a combined approach where there would be input between these independent efforts. We hypothesize that by employing machine learning models paired with model validation we can characterize the accuracy of current models used in practice, validate proposed models in a consistent framework, and set appropriate guidelines for data collection. In this way, we will work toward efficiently closing the current gap between observations and predictions in geotechnical earthquake engineering.

In order to demonstrate the proposed approach, we will examine empirical models for predicting lateral spread displacements. The amount of displacement due to lateral spreading depends on the physical and mechanical characteristics of the soil layers at a site, such as earthquake magnitude, distance from site to energy source, thickness of the critical layer, attenuation properties of the in situ soil, ground slope conditions, and water table depth (Towhata et al. 1992; Youd et al. 2002). Since several factors govern the liquefaction-induced lateral spreading, determination of its displacement is a complex geotechnical engineering problem that has been approached in many ways. The complexity of the problem and its impact on engineered structures have motivated researchers to model lateral spread displacement using analytical (Newmark 1965; Makdisi and Seed 1978; Towhata et al. 1992, Tokida et al. 1993; Kramer 1996; Franklin and Chang 1977), finite-element (Hamada et al. 1987; Yasuda et al. 1992; Gu et al. 1994), empirical (Youd and Perkins 1987; Bartlett and Youd 1995; Youd et al. 2002), semiempirical (Zhang et al. 2004), and machine learning approaches (Wang and Rahman 1999; Baziar and Ghorbani 2005; Javadi et al. 2006).

Currently, of all the different models, practitioners mostly use the empirical MLR model proposed by Bartlett and Youd (1992). Bartlett and Youd further updated and modified this approach in 1995 and 2002 (Bartlett and Youd 1995; Youd et al. 2002). Although these modifications have improved the ability to model lateral spread displacement compared to the initial approach pro-

[1]Dept. of Civil and Environmental Engineering, Tufts Univ., 113 Anderson Hall, Medford, MA 02155; presently, Dept. of Geological Engineering, Michigan Tech., Houghton, MI 49931 (corresponding author).

[2]Dept. of Civil and Environmental Engineering, Tufts Univ., 113 Anderson Hall, Medford, MA 02155.

Note. This manuscript was submitted on May 8, 2009; approved on December 11, 2009; published online on October 15, 2010. Discussion period open until April 1, 2011; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Computing in Civil Engineering***, Vol. 24, No. 6, November 1, 2010. ©ASCE, ISSN 0887-3801/2010/6-467–477/$25.00.

posed in 1992, Youd et al. (2002) recognized that better predictive capability is desirable in practice.

The current research on improving the predictive capability of lateral spread displacement has focused on improving the data quality (Zhang and Zhao 2005) or on the application of new modeling approaches such as machine learning (Wang and Rahman 1999; Baziar and Ghorbani 2005; Javadi et al. 2006). The existing models for lateral spread displacement lack a common and systematic form of validation, which prevents the recognition of improvement in model accuracy. Furthermore, no previous research has attempted to model lateral spread displacement using the machine learning approach: support vector regression (SVR). By using model validation, we show that SVR models offer better predictive capability over previous models, and provide a unique advantage in identifying what data are needed to improve the estimates of lateral spread displacement.

Machine learning algorithms have become popular in different fields due to their ability to make appropriate predictions that involve a high degree of complexity and nonlinearity. In geotechnical engineering, Wang and Rahman (1999) and Baziar and Ghorbani (2005) applied ANN to predict the lateral spread displacement and Hashash et al. (2003) used neural networks to update soil constitutive models using field performance data. Support vector machine (SVM) is a relatively new machine learning algorithm that has been proven to have several advantages over ANN such as absence of local minima and sparseness of the solution (Pal 2006; Misra et al. 2009). SVM utilizes the structural risk minimization principle, which has been shown to be superior to the empirical risk minimization principle employed by ANN (Gunn 1998). Researchers first developed SVM for classification problems; SVM for classification has been applied to geotechnical engineering in several instances (Pal 2006; Goh and Goh 2007; Sahoo et al. 2007; Oommen et al. 2008). Vapnik (1995) introduced an $\varepsilon$-insensitive loss function to extend SVM for regression problems, which is known as SVR. In SVR, $\varepsilon$ is the error limit, where any deviations $>\varepsilon$ are not acceptable.

In this study, we apply SVR models to the lateral spread displacement problem and outline how information gained from the SVR model can be used to inform data collection efforts. In machine learning, the data used to develop the model are called training data. When a model is developed using SVR, only the training instances that lie on the maximum margin (the support vectors) are required (Fig. 1), and these instances are known as support vectors. All other instances in the training data do not contribute to the final model. If instances likely to be support vectors can be identified in advance, this information may be used to intelligently plan the data collection efforts (Foody and Mathur 2004; Oommen et al. 2008). Intelligent data collection consists of sampling in the regions that may contain useful training instances, and avoiding those sample regions that will contribute insignificantly to the SVR, thereby reducing collection cost and improving predictability.

## Data Set

We use case histories of lateral spreading compiled by Youd et al. (2002). The data set contains 484 cases from 8 earthquakes, of which 371 cases are from Japan and 113 cases are from U.S. earthquakes. Based on the topographical conditions, Youd et al. (2002) compiled these data into two different groups: lateral spread toward a free face, and lateral spread down gentle ground slopes where a free face was absent. The free face lateral displacement has a higher maximum (max), mean, and standard deviation (SD) compared to the slope condition (Table 1). Out of the 484 cases compiled by Youd et al. (2002), we used 219 free face cases for this study. The researchers obtained the lateral spread displacement for the Japanese earthquakes from ground displacement vector maps developed from pre- and postearthquake aerial photographs (Hamada et al. 1986; Bartlett and Youd 1992). The estimated accuracies of these data are about $\pm 0.5 - \pm 0.75$ m. The researchers also obtained the displacement for U.S. earthquakes from a combination of photogrammetry, pre- and postearthquake ground surveys, reports of offsets of bridges, buildings, fences, canals, and other structures (Bartlett and Youd 1992). The accuracies of these data are highly variable and may range from $\pm 0.01 - \pm 0.5$ m (Bartlett and Youd 1992).

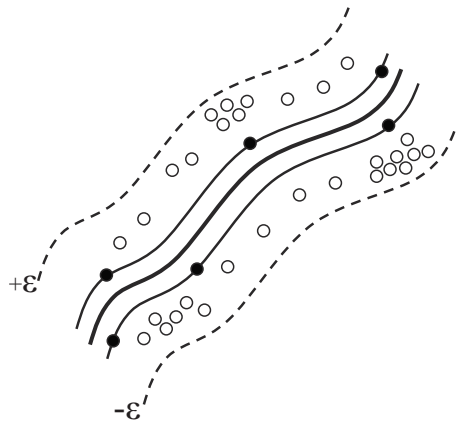


**Fig. 1.** Thick center line represents the optimal hyperplane (regression line), and the dashed outer lines represent the error margin; the filled dots that fall along the error margin are the support vectors; the open dots are the points that do not contribute to the definition of the regression line

## Methodology

In this study, we set a validation framework for existing models and then develop a SVR model as a comparison to the existing MLR model (Youd et al. 2002). The data used for the analysis are the lateral spread displacement for the free face condition. We structure the rest of the methodology section into four subsections. The first subsection explains the validation of models and how the different models can be reliably verified and compared to each other to evaluate their predictive performance. The middle two subsections briefly discuss the background, theory, and the inputs used for the two different models (empirical and SVR). The final subsection discusses how to identify data gaps using model validation techniques for intelligent data collection efforts.

### *Model Performance Evaluation*

In modeling, reliably ascertaining the predictive capability of the developed model is a major challenge. In Youd et al. (2002), the validation of the model used the same data that were used to develop the model (training data); therefore, the validation does not directly assess the predictive capability. Instead, the validation statistics evaluate the model fit (training fit) to the data. The model fit may be an influenced estimate of the predictive capability of the model as the uncertainty in the data set increases. On

**Table 1.** Statistics of Input and Output Data Used for Empirical and SVR Approaches

| Input and output parameters | Topography | Statistics | | | |
|---|---|---|---|---|---|
| | | Min | Max | Mean | SD |
| $M$=magnitude of earthquake | Free face | 6.4 | 9.2 | 7.22 | 0.52 |
| | Slope | 6.4 | 9.2 | 7.54 | 0.28 |
| $R$=distance from the site to energy source (km) | Free face | 0.5 | 100 | 16.65 | 11.41 |
| | Slope | 0.2 | 100 | 22.69 | 8.65 |
| $W$=free face ratio (%) | Free face | 1.64 | 56.8 | 10.5 | 9.14 |
| $S$=ground slope (%) | Slope | 0.05 | 11 | 0.92 | 1.41 |
| $T_{15}$=cumulative thickness of saturated granular layers (m) | Free face | 0.2 | 16.7 | 8.71 | 4.81 |
| | Slope | 0.7 | 19.7 | 6.65 | 3.77 |
| $F_{15}$=average fines content (%) | Free face | 1 | 70 | 16.83 | 13.42 |
| | Slope | 0 | 68 | 8.71 | 10.55 |
| $D50_{15}$=average mean grain size (mm) | Free face | 0.03 | 1.98 | 0.36 | 0.4 |
| | Slope | 0.06 | 12 | 0.39 | 0.78 |
| $D_H$=measured lateral displacement (m) | Free face | 0.01 | 10.16 | 2.62 | 2.28 |
| | Slope | 0.01 | 5.36 | 1.9 | 0.92 |

the other hand, previous researchers who have used machine learning algorithms for modeling lateral spread displacement (Wang and Rahman 1999; Baziar and Ghorbani 2005; Javadi et al. 2006) have used a single split of the data (testing data) for model assessment. When the data set is sufficiently large, a single split can be an unbiased estimate of the generalization error of the model. However, when the data set is small, model error can vary greatly depending on how the split is selected. It might be that we are just "lucky" or "unlucky" with a particular split and the generalization error estimate can be highly biased. When a single-split approach is used, usually researchers tend to try several random trials and present the split that had the least errors (i.e., Baziar and Ghorbani 2005). This results in presenting the lucky and therefore optimistically biased estimate of the predictive capability of the model.

Moreover, when presenting the results, the previous researchers who have used machine learning algorithms (Baziar and Ghorbani 2005; Javadi et al. 2006) have used a weighted average (based on the training and testing data size) of the training and testing performance of the model as the predictive capability of the model. In most cases, the sizes of the training data are much bigger than the testing data; therefore, when the performance of the model is evaluated using weighted average, the predictive performance from training data gets much higher weight than the performance on the testing data. In general, when nonlinear modeling techniques are used for empirical modeling and, in particular, when machine learning techniques are used, it is critical to make sure that the model is not overfit to the training data. In most data, two patterns exist: one from the real structure of the data and the other due to the presence of spurious noise. When we evaluate the nonlinear modeling technique for the optimal performance based on a training set, it simply strives to fit both the above patterns to achieve optimum performance. This can lead to the development of a model that has poor predictive capability.

Fig. 2 illustrates that averaging the training and testing errors can lead to a biased estimate of the model's final predictive capability. Factors such as the training and testing data size, the amount of overfitting in the training data, and the luck factor in the single-split testing performance contribute to this bias. For instance, we have a data set with two predictor variables that are nonlinearly related to the predicted data, as shown in Fig. 2. In the first case, a linear model is fitted to the data (Model A) [Fig. 2(a)], in the next case a nonlinear model with good generalization is fitted to the data (Model B) [Fig. 2(b)], and in the third case a nonlinear model is fitted by overfitting (Model C) [Fig. 2(c)]. In each case, we use 70% of the data to build the model, and the remaining 30% for testing. If the predictive capability of each of these models is compared using the training data and quantified

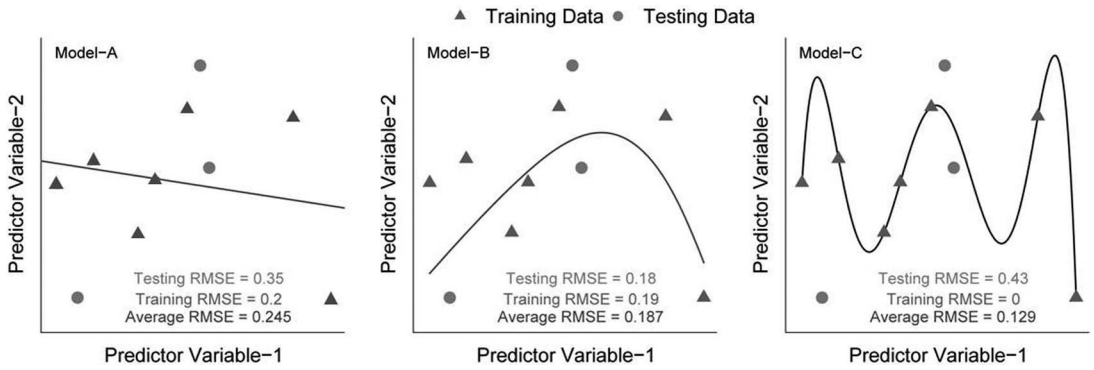


**Fig. 2.** Misrepresentation of the model performance due to averaging of training and testing errors: (a) Model A represents a linear model fit to the nonlinear data; (b) Model B represents a nonlinear model with good generalization fit to the nonlinear data; (c) Model C represents a nonlinear model overfit to the nonlinear data

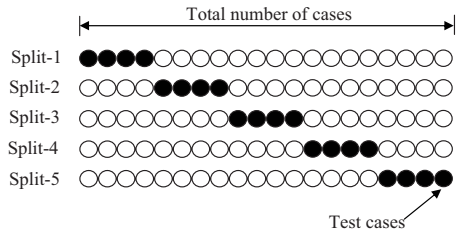


**Fig. 3.** *K*-fold cross validation where the original data are split into *K* approximately equal folds; for each *K* split, *K*−1-folds are used for training, and the remaining onefold is used for testing (filled dots represent the testing data, whereas the open dots represent the training data for each split)

using the root-mean-square error (RMSE) [definition is given later in the session, refer to Eq. (1)], we observe that Model C has the lowest RMSE. However, when the predictive capabilities of these models are evaluated using the testing data (data not used for model development), we observe that Model B has the least errors, whereas Model C has the highest errors. This illustrates that although Model C has the least error when evaluated using the training data, it has poor generalizability due to overfitting. If the errors evaluated using the training and testing data are averaged, we observe that even though Model C has poor predictive capability for future prediction, it has the least error. Therefore, averaging the training and testing errors provides a false sense of the predictive capability of a model, and may indicate that the overfit model has the best predictive capability. Machine learning algorithms stand a greater risk of overfitting data than other linear regression models.

*K*-fold cross validation offers a more reliable and robust approach for reporting the final predictive capability of a model than the single splits discussed above (Fig. 3) (Breiman and Spector 1992). For each of the *K* splits, we use *K*−1-folds for training and the remaining onefold for testing. The advantage of *K*-fold cross validation is that all the examples in the data set are eventually used for both training and testing. Finally, the generalizability of the model is estimated by calculating error estimates on the test cases of each of the *K* splits. Breiman and Spector (1992) demonstrated that a $K=10$- or $K=5$-fold cross validation works the best for linear regression. Therefore, in this study, we use a *K*-fold cross validation with $K=5$.

Most of the previous researchers who modeled the lateral spread displacement used the RMSE, and the coefficient of correlation ($r$) to quantify the model performance and error estimates

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \tag{1}$$

$$r = \left( \frac{n\sum_{i=1}^{n} O_i P_i - \sum_{i=1}^{n} O_i \sum_{i=1}^{n} P_i}{\sqrt{n\sum_{i=1}^{n} O_i^2 - \left(\sum_{i=1}^{n} O_i\right)^2}\sqrt{n\sum_{i=1}^{n} P_i^2 - \left(\sum_{i=1}^{n} P_i\right)^2}} \right) \times 100\% \tag{2}$$

where $O_i$=observed value; $P_i$=predicted value; and $n$=total number of observations.

Legates and McCabe (1999) illustrated that $r$ is insensitive to additive and proportional differences between the model prediction and observations [i.e., $r=100\%$ for any $P_i=(AO_i+B)$, where

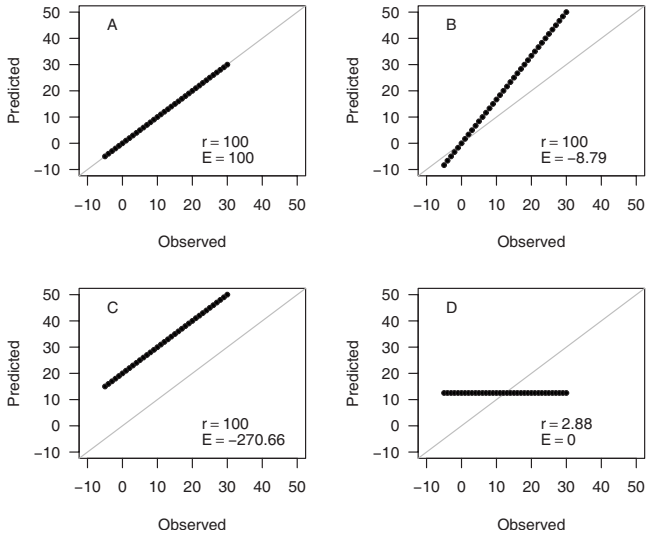


**Fig. 4.** (a) Observed and predicted values fall along the line of equality; (b) the values are underpredicted; (c) the values are overpredicted; and (d) the predicted value is close to the average of the observed value (the $r$ and $E$ values are presented as percentages)

$A$=any nonzero value and $B$=any value]. Fig. 4 demonstrates the insensitivity of $r$ to additive and proportional differences. In Fig. 4(a), $r=100\%$ and the model prediction is close to the line of equality. However, in Figs. 4(b and c), $r$ is again equal to 100% but the model is overpredicting in both cases. Because of these limitations, $r$ can indicate a model as a good predictor, even when it is not.

Nash and Sutcliffe (1970) defined the term coefficient of efficiency ($E$), which is a widely used goodness of fit measure in hydrology

$$E = \left( 1.0 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \right) \times 100\% \tag{3}$$

where $\bar{O}$=mean observed value; and $O_i$ and $P_i$=same as previously defined.

The $E$ varies from $-\infty$ to 100% with values closer to 100% indicating better agreement with the observed and predicted values. Figs. 4(a–c) illustrate that $E$ is sensitive to the additive and proportional differences between the model prediction and observations and can be a better performance measure of the predictive capability of the model than $r$. When the predicted value is close to the observed, the $E=100$ [Fig. 4(a)]. Unlike $r$, when the model overpredicts [Figs. 4(b and c)] the $E$ decreases. Legates and McCabe (1999) illustrated that when the sums of the mean squared difference between the predicted and observed data are as large as the variability in the observed data, then $E=0$. In other words, when the values predicted by the model are close to the mean of the observed data, $E=0$ [Fig. 4(d)]. Legates and McCabe (1999) also illustrated that when the sum of the mean squared difference exceeds the variability in the observed data, then $E<0$.

In the following sections, we compare different models using $E$. However, we have reported in the tables each of the three measures (RMSE, $r$, and $E$) so that the readers will be able to compare the models developed in this study to the lateral spread displacement models by previous researchers.

### Empirical Approach

In this study, we compare our results to the empirical approach developed by Youd et al. (2002) for lateral spread displacement. The equation of Youd et al. (2002) for the lateral spread displacement of the free face condition is

$$\log D_H = -16.713 + 1.532M - 1.406 \log R^* - 0.012R + 0.592 \log W + 0.540 \log T_{15} + 3.413 \log(100 - F_{15}) - 0.795 \log(D50_{15} + 0.1) \quad \text{(m)} \tag{4}$$

where $R$=nearest distance to the seismic energy source ($R^* = R + R_0, R_0 = 10^{(0.89M-5.64)}$); $M$=moment magnitude of the earthquake; $T_{15}$=cumulative thickness of saturated granular layers with corrected blow counts of standard penetration test less than 15; $F_{15}$=average fines content for granular materials included within $T_{15}$; $D50_{15}$=average mean size for granular materials with $T_{15}$; and $W$=free face ratio.

### SVR

In this paper, we provide a brief overview of the theoretical concepts of SVR. However, a detailed depiction of SVR is beyond the scope of this paper and can be obtained from Vapnik (1995), Kecman (2000), and Hastie and Friedman (2003).

For instance, we have a training data set $\{(x_{11}, x_{12}, \ldots, x_{1n}, y_1), \ldots, (x_{l1}, x_{l2}, \ldots, x_{ln}, y_l)\} \subset X \times R$, where $(x_{l1} \cdots x_{ln})$ represent the predictor variables and $y_l$ represents the observed lateral spread displacement at that location. The goal in SVR is to find a function $f(x)$ that has the most $\varepsilon$ deviation from the observed lateral displacements $y_l$ for all the training data, and at the same time, is as straight as possible to reduce the model complexity. In other words, what Vapnik (1995) introduced through the $\varepsilon$-insensitive loss function is that errors less than $\varepsilon$ are acceptable, but the loss function does not accept any deviation larger than this

$$f(x) = \langle w, x \rangle + b \quad \text{with} \quad w \in X, \quad b \in R \tag{5}$$

where $\langle w, x \rangle$ denotes the dot product in $X$. Straightness in Eq. (5) means a small value of $w$, and it can be obtained by minimizing the Euclidean norm, i.e., $\|w\|^2$. Thus the SVR problem can be formulated as follows:

$$\text{Minimize} \quad \tfrac{1}{2}\|w\|^2$$

$$\text{subject to} \quad \begin{cases} (\langle w, x_i \rangle + b) - y_i \leq \varepsilon \\ y_i - (\langle w, x_i \rangle + b) \leq \varepsilon \end{cases} \tag{6}$$

However, in some cases, it is not feasible to have a function $f(x)$ that is straight with errors less than $\varepsilon$. In order to deal with these infeasible situations a constant $C$ and slack variables $\xi_i^-$ and $\xi_i^+$ are introduced, which lead to the formulation [Eq. (7)] stated in Vapnik (1995)

$$\text{Minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} (\xi_i^- + \xi_i^+)$$

$$\text{subject to} \quad \begin{cases} (\langle w, x_i \rangle + b) - y_i & \leq \varepsilon + \xi_i^- \\ y_i - (\langle w, x_i \rangle + b) & \leq \varepsilon + \xi_i^+ \\ \xi_i^-, \xi_i^+ & \geq 0 \end{cases} \tag{7}$$

where $C$=prespecified term that controls the magnitude of the penalty associated with errors outside the error margin; and $\xi_i^-$ and $\xi_i^+$=slack variables, which measure the degree of misclassification of the upper and lower error margins (Fig. 1).

The constant $C > 0$ determines the trade-off between the straightness of the function and the amount to which deviations larger than $\varepsilon$ are tolerated (Scholkopf and Smola 2002). The use of Lagrange multipliers employing the Kharush-Kuhn-Tucker (KKT) method can solve the optimization problem in Eq. (7). The KKT method converts the inequality constraint into an equation of the form $h(x)=0$ by adding or subtracting slack variables and then solving the corresponding equality constrained quadratic optimization problem. Solution of the optimization problem results in a dual pair variable Lagrangian $L_d$ $(\alpha_i, \alpha_i^*)$, one for each of the training patterns. The pairs that result in nonzero $\alpha_i$ or $\alpha_i^*$ are termed as the support vectors. When the SVR model is developed (training), the support vectors define the hyperplane (regression line) (Fig. 1). Any data point in the training set that falls within the error margin does not contribute to the definition of the regression line.

In complex nonlinear problems, the original input space (predictor variable) is often nonlinearly related to the predicted variable (lateral spread displacement). This limits a linear formulation of the problem, as shown in Eq. (5). In SVR, this limitation is overcome by mapping the input space onto some higher dimensional space (feature space) using a nonlinear mapping function called kernel function. The advantage of the kernel function is that it enables us to implicitly work in a higher dimensional feature space and overcome the issues of dimensionality. Commonly used kernel functions include the linear, polynomial, Gaussian radial basis, and the sigmoid kernel functions. Keerthi and Lin (2003) showed that a linear kernel is a special case of the Gaussian radial basis kernel and that the sigmoid kernel behaves like a Gaussian radial basis kernel for certain parameters. In this study we use a Gaussian radial basis kernel function [Eq. (8)] because it is a more general function

$$K(x, x') = \exp\left(-\frac{\| x - x' \|}{2\gamma^2}\right) \tag{8}$$

In any SVR problem, when we use the Gaussian radial basis function as the kernel function, we have three parameters to optimize during training: the Gaussian radial basis function parameter $\gamma$, magnitude of penalty term $C$, and the width of the error margin $\varepsilon$. The Appendix provides a step by step guideline for the implementation of SVR and the optimization of the parameters. In modeling the lateral spread displacement using SVR, we use the same input parameters that have been used by Youd et al. (2002). The single output variable was the logarithm of the lateral spread displacement $\log(D_H)$.

### Identifying Data Gaps

Identifying gaps in data sets is extremely important for improving empirical models. The conventional empirical approaches use the entire training data for the development of the model, whereas the SVM uses a subset of the training data known as support vectors. Therefore, identifying the characteristics of these support vectors a priori can help in strategizing the data collection and in turn improving the empirical model using SVM.

In the SVM model, the support vectors define the shape of the regression line and they fall along the error insensitivity boundary (Fig. 1) during the training phase of the model development. The cases within the error insensitivity boundary and the regression line do not contribute to model development. Support vectors being points on the error insensitivity boundaries are the most

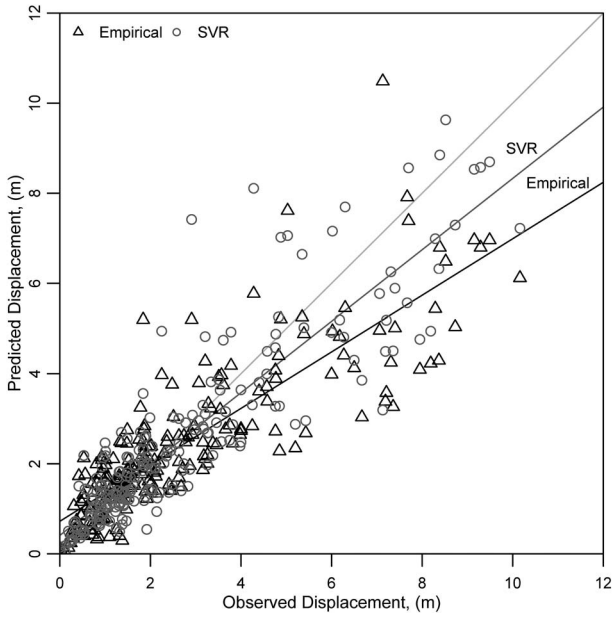


**Fig. 5.** Observed versus predicted lateral spread displacement using empirical and SVR approaches for free face cases; the lighter line is the line of equality, and the darker lines are the linear models of the observed and predicted values for the empirical and SVR models; the linear models are included to summarize the general trends in the predictions of the different models

distant away from the regression line, and thus have the highest error and uncertainty. Having the highest uncertainty, the region close to the support vectors requires further data collection to better constrain the empirical model.

In this study, we identify these regions by analyzing the range of the predictor variables and the quantity of support vectors in each of these ranges. The predictor variable can be divided into different ranges either based on its statistical distribution or based on equal intervals. In this study, we explore both the approaches to analyze their suitability in identifying the data gaps. In the case of equal intervals, the predictor variables except for $M$ are divided into ten equal intervals, whereas $M$ is divided into four considering the discrete nature of $M$ values compared to the other predictor variables. When dividing the data based on the statistical distribution each predictor variable was divided into six intervals. In each of these intervals, there were about 30+ data points except for $M$ and $R$ values. The $M$ and $R$ values were discrete with some values repeating over 100 times (e.g., out of the 176 data points 114 have an $M=7.5$). Ideally, we like to have a low percentage of support vectors from each range of the predictor variable, i.e., an indication that there is low uncertainty in these ranges. The specific range in the predictor variable that needs future data collection will be the one that has a high percentage of support vector or no data at all.

## Results and Discussion

We analyze the predictive capability of SVR and the empirical approach for the 219 free face cases using the scatter plot of the observed versus predicted values (Fig. 5) and the model performance measures (RMSE, $r$, and $E$) in Table 2. We find from Fig. 5 that the predicted values of lateral spread displacement using the SVR model are closer to the line of equality compared to the empirical approach. Table 2 shows that for all the five splits of free face cases and for all three performance measures, the SVR model exhibits relatively improved predictive capability compared to the empirical approach. The optimum values of $\gamma$, $C$, and $\varepsilon$, respectively, are 0.136, 9.0, and 0.162 for the SVR model. In the SVR model, roughly 60% of the training data are support vectors in each of the five splits. The large percentage of support vectors is an indicator of the complexity of the process that is being modeled and the high uncertainty in the data set. Combining the performance measures on the prediction of all the five splits of free face cases, we see that the SVR performs relatively better than the empirical approach. The SVR has a 4% improvement in $r$ and an 8% improvement in $E$ compared to the empirical approach.

We note that previous researchers (Wang and Rahman 1999; Baziar and Ghorbani 2005; Javadi et al. 2006) who used machine learning algorithms for predicting lateral spread displacement reported performance measures on a single split as the representative predictive capability of the method. Table 2 verifies that there is a considerable variation in the predictive performance measures (RMSE, $r$, and $E$) for each of the five splits as a result of the relatively small data set. The $E$ for the SVR varies from 60.68 to 85.99%. It is, therefore, likely that the previous researchers who reported a performance measure of only one of the splits have presented a biased measure of the predictive capability of the machine learning algorithm. Therefore, in the case of modeling small data sets for geotechnical engineering applications using machine learning algorithms, we recommend that a fivefold cross validation approach be used to reliably assess the predictive capability of the model.

We note that the predictive performance measures for the SVR model are based on the test data (not used for training the model), whereas in the case of empirical approach the entire data have been used to develop and test the regression equation. Therefore, in the case of the empirical approach, the following question still remains: "What is the predictive capability of the model for testing data?" This clearly illustrates that the SVR model has an

**Table 2.** Performance Measures of Empirical and SVR Approaches for Free Face Data

| | RMSE | | $r$ | | $E$ | |
|---|---|---|---|---|---|---|
| | Empirical | SVR | Empirical | SVR | Empirical | SVR |
| Split 1 | 1.05 | 0.77 | 84.21 | 91.02 | 68.57 | 82.82 |
| Split 2 | 1.43 | 0.98 | 90.94 | 94.50 | 70.37 | 85.99 |
| Split 3 | 1.30 | 1.23 | 82.07 | 86.34 | 65.54 | 69.42 |
| Split 4 | 1.28 | 1.27 | 78.68 | 78.99 | 60.17 | 60.68 |
| Split 5 | 1.21 | 0.94 | 88.75 | 93.40 | 75.97 | 85.46 |
| Summary | 1.26 | 1.06 | 84.77 | 88.80 | 69.42 | 78.50 |

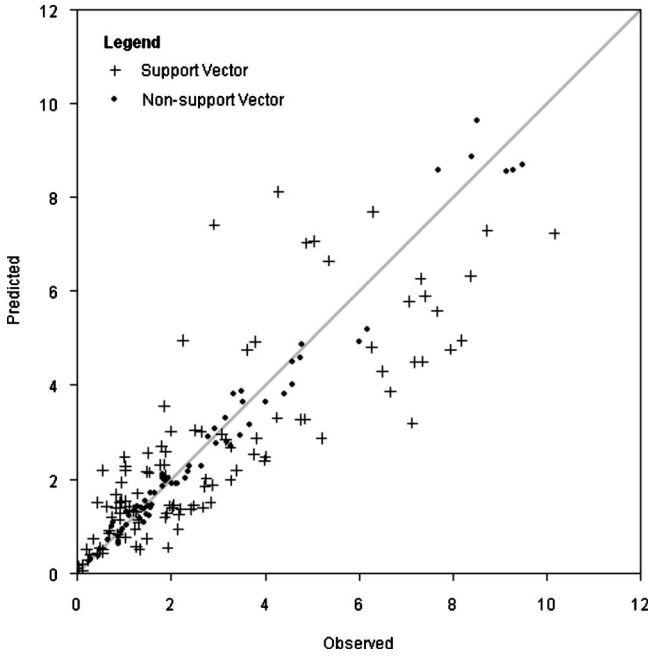


**Fig. 6.** Observed versus predicted lateral spread displacement for the training instances in Split 1 of free face cases; the plus sign indicates the support vector points and the filled dots indicate the nonsupport vector points; the diagonal line is the line of equality

improved predictive capability over the empirical approach for lateral spread displacement.

We use the concept of support vectors to analyze the data gaps in the data set for developing recommendations for the future data collection. Fig. 6 presents a scatter plot of the observed versus predicted lateral spread displacement for the training instances in Split 1 of free face cases. Split 1 has a total of 176 training instances, of which 107 are support vectors and the remaining 69 are nonsupport vectors. The instances that are support vectors and nonsupport vectors are differentiated with a plus sign and a filled dot, respectively. Fig. 6 illustrates that the nonsupport vector instances are closer to the line of equality than the support vector instances. The large numbers of support vectors (60.78%) indicate the higher uncertainty that exists in the data set and the resulting complexity in the model.

Fig. 7 presents the distribution of the support vectors in the range of the predictor variables defined in Table 1, divided into equal intervals. This figure helps us to understand the ranges in the predictor variables that need additional data collection in order to reduce the uncertainty in the data set and thus improve the predictive capability for lateral spread displacement. We observe that of the predictor variables, $T_{15}$ has the least uncertainty and lowest percentage of support vectors from the entire range of the data, whereas $R$ has the highest uncertainty. The $M$ values range from 6.4 to 9.2 with values from 7.8 to 8.5 being poorly sampled. The $R$ values range from 0.5 to 60 with values $>12.4$ being poorly sampled, which indicates the need for more data from regions with larger values of nearest distance to the seismic energy source. The $W$ values range from 1.64 to 55.68 with values $>23.26$ being poorly sampled with the exception of values from 39.47 to 44.87. The $T_{15}$ values are pretty well constrained through the entire range 0.5–16.7 with the exception of values from 8.6 to 10.22. The $F_{15}$ values are well constrained in the lower range from 1 to 33.5 with values $>33.5$ being poorly sampled. The $D_{50}$ values are well constrained in the lower and upper tails with values from 0.45 to 1.79 being poorly sampled.

Fig. 8 presents the distribution of the support vectors in the range of the predictor variables divided into six intervals based on the statistical distribution. We notice from Fig. 8 that most intervals have about 50–70% support vectors with few exceptions. This indicates that when the predictor variable is divided statistically most intervals have similar uncertainty in the data. The few exceptions with lower support vectors are observed in predictor variables $M$, $R$, and $D_{50}$ with values ranging from 6.5 to 6.6, 2.0 to 5.5, and 0.26 to 0.33, respectively. Similarly, few exceptions with higher support vectors are observed in predictor variables $M$ and $D_{50}$ with values ranging from 6.6 to 6.8 and 0.2 to 0.26. The lower and higher exceptions of support vectors observed in the predictor variables $M$ and $R$ are not of significant interest considering the discrete nature of these predictor variables, which caused the data to be nonuniformly distributed in the different intervals. Therefore, we excluded $M$ and $R$ from further quantitative analysis of the variation of support vectors in each interval. Fig. 9 provides a quantitative representation of the number of data points, amount of support vectors, and the percentage of support vectors in each interval of the predictor variables $W$, $T_{15}$, $F_{15}$, and $D_{50}$. We observe from Fig. 9 that the number/percentage of support vector in each interval is nearly uniform for predictor vari-

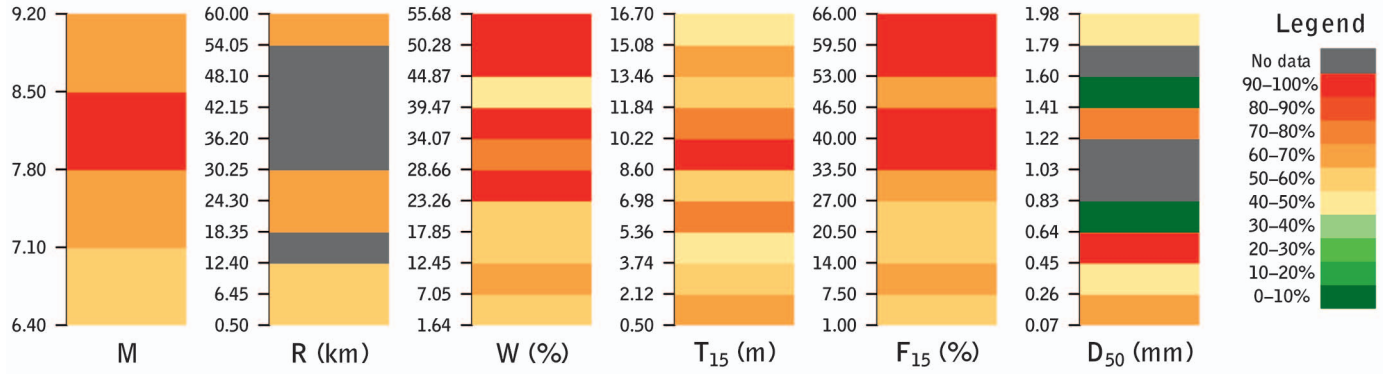


**Fig. 7.** (Color) Percentage of support vectors in the range of predictor variables for the lateral spread displacement of free face cases; the $y$ axis represents the range of the predictor variable and the colors represent the percent of support vector; a low percentage of support vector specifies that of all the instances available from the particular region only few instances are support vectors, which in turn indicate that there is less error and uncertainty in this region of the predictor variable

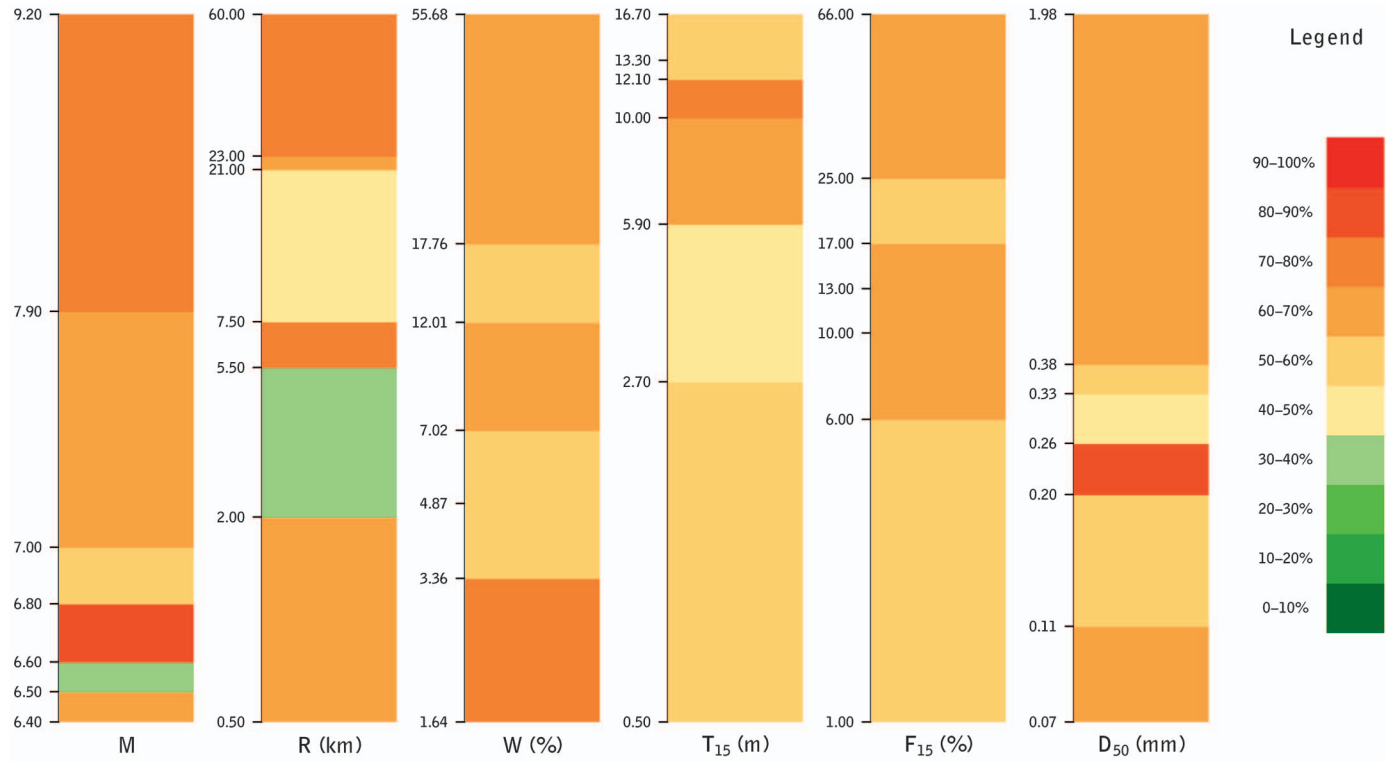


**Fig. 8.** (Color) Percentage of support vectors in the range of predictor variables divided based on the statistical distribution of the data for the lateral spread displacement of free face cases; each segment of the bar graph for $W$, $T_{15}$, $F_{15}$, and $D_{50}$ represents an approximately equal portion of the data (roughly 30 points); the segments in the $M$ and $R$ bar graphs represent approximately equal intervals but have some segments that are very highly sampled as a result of the discrete nature of the data set (e.g., $M=7.5$); the $y$ axis represents the range of the predictor variable and the colors represent the percent of support vector

ables $W$, $T_{15}$, and $F_{15}$, whereas in $D_{50}$ there is an anomalous increase in the percentage of support vectors from values 0.20 to 0.26. To investigate the cause of the anomalous increase in support vectors observed in $D_{50}$, we plotted the different predictor variables ($W$, $T_{15}$, $F_{15}$, and $D_{50}$) against the lateral spread displacement, as shown in Fig. 10. We observe from Fig. 10 that the displacement with regard to $D_{50}$ has a distinct pattern with values from 0.04 to 0.2 having low displacement, values from 0.2 to 0.38 having large variability in displacement, and values from 0.38 to 1.98 having much lower displacements. It is interesting to note that the anomalous increase in the percentage of support vectors for $D_{50}$ from values 0.20 to 0.26 is at the boundary between the low displacement and the large variability in displacement. This indicates that the increase in support vectors that we observe in $D_{50}$ from values 0.20 to 0.26 is due to the uncertainty at this boundary, and the data collection should particularly focus on this range to improve the model for lateral spread displacement.

A comparison of the approaches of dividing the data into equal intervals (Fig. 7) and dividing based on the statistical distribution (Fig. 8) indicates that the latter approach helps to make more definite conclusions about the data gap and meaningful recommendations for data collection.

In order to use the SVR model to predict the lateral spread displacement for a given site, the geotechnical practitioners would run the SVR algorithm using the training data and optimal model values. For the convenience of practitioners we have provided the code for the SVR algorithm and the training data set for the free face case at the following URL: http://ase.tufts.edu/cee/geohazards/peopleOommen.asp. The model values ($\gamma$, $C$, and $\varepsilon$) presented in this paper can be used for the optimal training of the models. The Appendix provides a step by step guideline for running the SVR algorithm.

## Summary and Conclusions

Using the 219 free face cases from the database developed by Youd et al. (2002) we illustrate a methodology for reliably verifying the applicability of a machine learning algorithm (SVR) for use with geotechnical engineering regression problems. We also demonstrate how proper validation combined with the concept of support vectors can help in developing intelligent future data collections efforts.

1. This study illustrates that SVR can be applied to model complex nonlinear geotechnical engineering problems such as the prediction of lateral spread displacement with improved predictive capability compared to the widely used empirical approach by Youd et al. (2002) using MLR.
2. The analysis of the support vectors in the SVR model provides a unique advantage over other common statistical empirical methods in identifying the gaps in the data and intelligently defining the scope for future data collection. In particular, the large numbers of support vectors (60.78%) indicate the high uncertainty that exists in the data set and the resulting complexity in the model.
3. In this study, we compare the approaches of dividing the data into equal intervals and dividing based on the statistical distribution to analyze the distribution of support vectors in the predictor variable. The latter approach of dividing the data

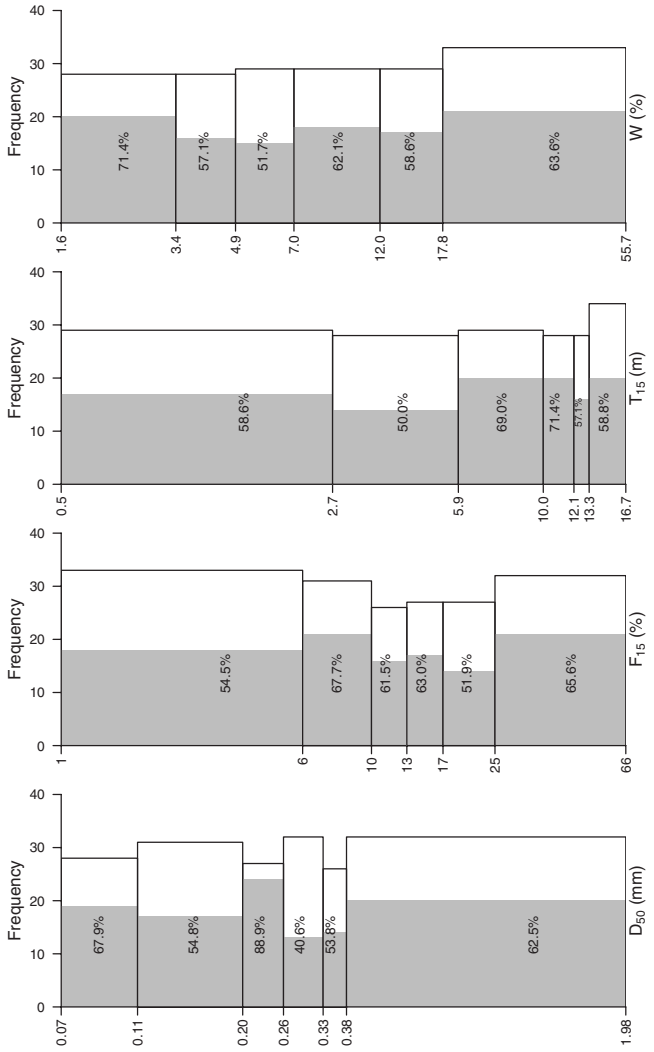


**Fig. 9.** Frequency of data, number of support vectors, and percentage of support vectors in each range of predictor variables divided based on the statistical distribution of the data for the lateral spread displacement of free face cases; the $x$ axis for the plots is scaled logarithmically for readability

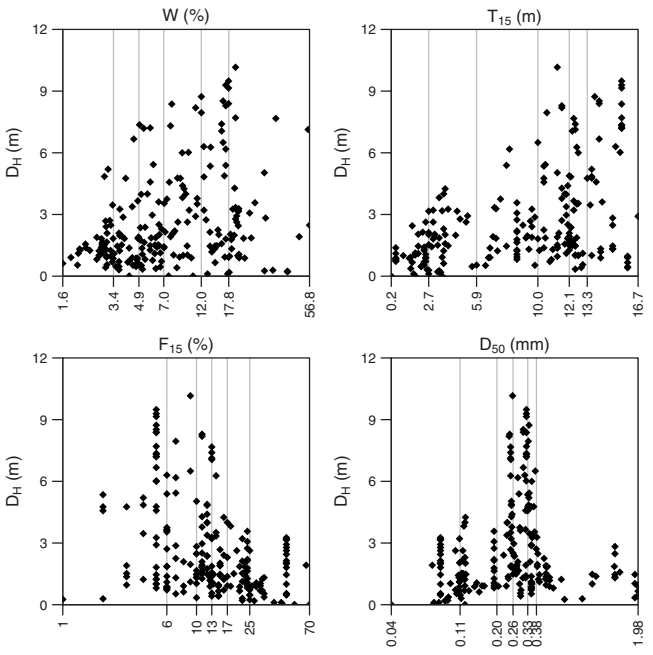


**Fig. 10.** Scatter plot of the predictor variables $W$, $T_{15}$, $F_{15}$, and $D_{50}$ against the lateral spread displacement $D_H$; the intervals on the $x$ axis correspond to the divisions based on the statistical distribution of the data presented in Figs. 8 and 9

based on the statistical distribution helps to make more definite conclusions about the data gap and meaningful recommendations for data collection.

4. Our analyses indicate the range of values from 0.2 to 0.26 of the predictor variable $D_{50}$ of particular interest for future data collection to improve the lateral spread displacement model.
5. Comparing model performance using the coefficient of correlation ($r$) can be misleading due to its insensitivity to the additive and proportional errors between the predicted and observed values; the coefficient of efficiency ($E$) is a more reliable measure of the model performance
6. The large variation (25%) in $E$ over the five splits of free face data indicates that using a single-split approach might misrepresent the predictive capability of the model, but instead the $K$-fold cross validation is a more robust and unbiased representation of the model predictability.

## Acknowledgments

The writers would like to thank the National Science Foundation (NSF) for supporting this work in part by a grant (Grant No. CMMI 0547190). Thanks also go to Eric M. Thompson, Eugene Morgan, and James Kaklamanos for the comments and suggestions on the draft. T.O. would also like to thank the Dean, College of Engineering, Tufts University, for the financial support for this work through the Deans Fellowship. The writers also wish to thank Professor T. L. Youd and S. F. Bartlett for providing access to the data set, which has been used in the study from their website.

## Appendix. Guidelines for SVR Implementation

In this study we used the SVR algorithm (Meyer 2006) available in the e1071 package version 1.5-17 (Dimitriadou et al. 2007) of the R programming language (R Team 2007). R is an open source programming language and software environment for statistical computing and graphics. Details on installing R can be obtained from the URL http://www.R-project.org. The guideline for SVR implementation described here is applicable for any regression problem.

1. The data are divided into training and testing data subsets. Training data are required to develop the model and the testing data are required to examine the performance of the model.
2. The data for training and testing are arranged in a matrix form in which each row represents an instance (of observed lateral spread displacement). The columns in the training data matrix represent the various predictor variables (refer to Table 1) and the corresponding predicted variable $[\log(D_H)]$, whereas the columns in the testing data matrix represent the various predictor variables only.
3. The data must be scaled to prevent columns of greater numeric ranges dominating over columns of smaller numeric ranges (the e1071 package does this by default).

4. In the SVR implementation, there are three parameters to be optimized (Gaussian radial basis function parameter $\gamma$, magnitude of penalty term $C$, and the width of the error margin $\varepsilon$). In order to optimize these parameters, the data are randomly divided into $K$ ($K=5$, number of folds for cross validation) subsets of approximately equal number of rows.
5. The three parameters are optimized by a grid search using $K$-fold cross validation. The grid search is done in two steps. Initially a coarser grid is applied with an exponentially growing sequence of $C=2^{-5},2^{-3},\ldots,2^{15}$, $\gamma=2^{-15},2^{-13},\ldots,2^{3}$ and $\varepsilon=2^{-10},2^{-8},\ldots,2^{-1}$. The coarser grid search is followed by a finer grid search in the region (value of $C$, $\gamma$, and $\varepsilon$) that produced the highest $E$ for the cross validation.
6. The finer grid search is repeated until the change in $E$ is not significant.
7. Finally the optimal values of $C$, $\gamma$, and $\varepsilon$ are used to train the SVR model and quantify the predictive capability using cross validation.

The data used for the SVR model (for each of the five splits of free face cases) and the code used for SVR algorithm in R can be obtained from the URL http://ase.tufts.edu/cee/geohazards/peopleOommen.asp.

## References


Bartlett, S. F., and Youd, T. L. (1992). *Empirical analysis of horizontal ground displacement generated by liquefaction-induced lateral spread*, National Center for Earthquake Engineering Research, Buffalo, N.Y.

Bartlett, S. F., and Youd, T. L. (1995). "Empirical prediction of liquefaction-induced lateral spread." *J. Geotech. Engrg.*, 121(4), 316–329.

Baziar, M. H., and Ghorbani, A. (2005). "Evaluation of lateral spreading using artificial neural networks." *Soil Dyn. Earthquake Eng.*, 25(1), 1–9.

Breiman, L., and Spector, P. (1992). "Submodel selection and evaluation in regression, the x-random case." *Int. Statist. Rev.*, 60(3), 291–319.

Cetin, K. O., et al. (2004). "Standard penetration test-based probabilistic and deterministic assessment of seismic soil liquefaction potential." *J. Geotech. Geoenviron. Eng.*, 130(12), 1314–1340.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A. (2007). *e1071 Misc functions*, Dept. of Statistics, TU Wien, Wien, Austria.

Foody, G. M., and Mathur, A. (2004). "Toward intelligent training of supervised image classification: Directing training data acquisition for SVM classification." *Remote Sens. Environ.*, 93, 107–117.

Franklin, A. G., and Chang, F. K. (1977). "Permanent displacements of earth embankments by Newmark sliding block analysis." *Misc. Paper No. s-71-17*, Soil and Pavements Lab, U.S. Army Eng. Waterway Experiment Station, Vicksburg, Miss.

Goh, A. T. C., and Goh, S. H. (2007). "Support vector machines: Their use in geotechnical engineering as illustrated using seismic liquefaction data." *Comput. Geotech.*, 34, 410–421.

Gu, W. H., Morgenstern, N. R., and Robertson, P. K. (1994). "Postearthquake deformation analysis of wildlife site." *J. Geotech. Geoenviron. Eng.*, 120(2), 274–289.

Gunn, S. (1998). "Support vector machines for classification and regression." *Technical Rep.*, Univ. of Southampton.

Hamada, M., Yasuda, S., Isoyama, R., and Emoto, K. (1986). *Study on liquefaction induced permanent ground displacements*, Association for the Development of Earthquake Prediction in Japan, Tokyo.

Hamada, M., Towhata, I., Yasuda, S., and Isoyama, R. (1987). "Study on permanent ground displacements induced by seismic liquefaction." *Comput. Geotech.*, 4, 197–220.

Hashash, Y. M. A., Marulanda, C., Ghaboussi, J., and Jung, S. (2003). "Systematic update of a deep excavation model using field performance data." *Comput. Geotech.*, 30(6), 477–488.

Hastie, T. R. T., and Friedman, J., eds. (2003). *The elements of statistical learning: Data mining, inference and prediction*, Springer, New York.

Javadi, A. A., Rezania, M., and Nezhad, M. M. (2006). "Evaluation of liquefaction induced lateral displacements using genetic programming." *Comput. Geotech.*, 33(4–5), 222–233.

Kecman, V. (2000). *Learning and soft computing: Support vector machines, neural networks and fuzzy logic models*, MIT Press, Cambridge, Mass.

Keerthi, S. S., and Lin, C. J. (2003). "Asymptotic behaviors of support vector machines with Gaussian kernel." *Neural Comput.*, 15(7), 1667–1689.

Kramer, S. L. (1996). *Geotechnical earthquake engineering*, Prentice-Hall, Upper Saddle River, N.J.

Legates, D. R., and McCabe, G. J. (1999). "Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation." *Water Resour. Res.*, 35(1), 233–241.

Makdisi, F. I., and Seed, H. B. (1978). "Simplified procedure for estimating dam and embankment earthquake induced deformations." *J. Geotech. Engrg. Div.*, 104(7), 849–867.

Meyer, D. (2006). *Support vector machines: The interface to libsvm in package e1071*, Technical Univ., Wien, Austria.

Misra, D., Oommen, T., Agarwal, A., Mishra, S. K., and Thompson, A. M. (2009). "Application and analysis of support vector machine based simulation for runoff and sediment yield." *Biosyst. Eng.*, 103(4), 527–535.

Moss, R. E. S., Seed, R. B., Kayen, R. E., Stewart, J. P., Kiureghian, A. D., and Cetin, K. O. (2006). "CPT-based probabilistic and deterministic assessment of in situ seismic soil liquefaction potential." *J. Geotech. Geoenviron. Eng.*, 132(8), 1032–1051.

Nash, J. E., and Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models. Part I—A discussion of principles." *J. Hydrol.*, 10(3), 282–290.

Newmark, N. M. (1965). "Effects of earthquake on dams and embankments." *Geotechnique*, 15(2), 139–160.

Oommen, T., Misra, D., Twarakavi, N. K., Prakash, A., Sahoo, B. C., and Bandopadhyay, S. (2008). "An objective analysis of support vector machine based classification for remote sensing." *Mathematical Geosciences*, 40(4), 409–424.

Pal, M. (2006). "Support vector machines-based modeling of seismic liquefaction potential." *Int. J. Numer. Analyt. Meth. Geomech.*, 30(10), 983–996.

Sahoo, B. C., Oommen, T., Misra, D., and Newby, G. (2007). "Using the one-dimensional S-transform as a discrimination tool in classification of hyperspectral images." *Can. J. Remote Sens.*, 33(6), 551–560.

Scholkopf, B., and Smola, A. (2002). *Learning with kernels*, MIT Press, Cambridge, Mass.

Team, R. D. C. (2007). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Tokida, K., Matsumoto, H., Azuma, T., and Towhata, I., eds. (1993). "Simplified procedure to estimate lateral ground flow by soil liquefaction." *Proc., 5th Int. Conf. on Soil Dynamics and Earthquake Engineering, VI*, Elsevier, New York, 381–396.

Towhata, I., et al. (1992). "Prediction of permanent displacement of liquefied ground by means of energy principle." *Soils Found.*, 32(3), 97–116.

Vapnik, V. (1995). *The nature of statistical learning theory*, Springer, New York.

Wang, J., and Rahman, M. S. (1999). "A neural network model for liquefaction-induced horizontal ground displacement." *Soil Dyn. Earthquake Eng.*, 18(8), 555–568.

Yasuda, S., Nagase, H., Kiku, H., and Uchida, Y. (1992). "The mechanism and simplified procedure for analysis of permanent ground displacement due to liquefaction." *Soils Found.*, 32(1), 149–160.

Youd, T. L., and Perkins, D. M. (1987). "Mapping of liquefaction severity index." *J. Geotech. Geoenviron. Eng.*, 113(11), 1374–1391.

Youd, T. L., Hansen, C. M., and Bartlett, S. F. (2002). "Revised multilinear regression equations for prediction of lateral spread displacement." *J. Geotech. Geoenviron. Eng.*, 128(12), 1007–1017.

Zhang, G., Robertson, P. K., and Brachman, R. W. I. (2004). "Estimating liquefaction-induced lateral displacements using the standard penetration test or cone penetration test." *J. Geotech. Geoenviron. Eng.*, 130(8), 861–871.

Zhang, J., and Zhao, J. X. (2005). "Empirical models for estimating liquefaction-induced lateral spread displacement." *Soil Dyn. Earthquake Eng.*, 25(6), 439–450.