Wet aggregate stability modeling based on support vector machine in multiuse soils

Ruizhi Zhai¹, Jianping Wang²,³, Deshun Yin¹ and Ziheng Shangguan⁴

Abstract
Accurate assessment of wet aggregate stability is critical in evaluating soil quality. However, a few general models are used to assess it. In this work, we use the support vector machine to evaluate wet aggregate stability and compare it with a benchmark model based on artificial neural networks. One hundred thirty-four soil samples from various land uses, such as crops, grasslands, and bare land are adopted to verify the effectiveness of the proposed method and confirm the valid input parameters. We select 107 samples for calibrating the prediction model and the rest for evaluation. Experiments show that organic carbon is the main control parameter of wet aggregate stability, although the most influential factors for different land use are various. Comparing the determination coefficient and the root mean square error, it proves that the support vector machine method is superior to the artificial neural network method. In addition, the relative importance analysis shows that contents of organic carbon, silt, and clay are the primary input parameters. Finally, the impact of land use and management types is evaluated.

Keywords
Soil analysis, support vector machine, artificial neural network, wet aggregate stability, organic carbon

Date received: 25 November 2021; accepted: 4 May 2022
Handling Editor: Yanjiao Chen

Introduction
Soil aggregate stability (SAS) is a critical factor for soil resistance to deformation by external force.¹ Lower SAS means the soil is more likely to break down into finer particles,² leading to soil disintegration, deformation, and failure, thus to soil erosion and engineering structure damage.³,⁴ Therefore, accurate assessment of SAS is of great significance for preventing water conservancy disasters.

Until now, SAS is usually evaluated by the wet sieving method and expressed as wet aggregate stability (WAS).⁵,⁶ However, the standard procedures and apparatuses are unestablished for measuring SAS. Moreover, the SAS measurement is time-consuming and costly. The pretransfer function (PTF) is often used to estimate the SAS based on linear regression. J Cañasveras et al.⁷ evaluate the ability of diffuse reflectance spectra (DRS) to predict the SAS indices by introducing partial least-squares (PLS) regression. The results indicate that DRS can estimate SAS quickly and categorize soil zones according to surface sealing and susceptibility to water erosion. C Gomez et al.⁸ explore a two-steps approach as an alternative to the normalized international measurements for estimating SAS indexes. First, the elementary properties are

¹College of Mechanics and Materials, Hohai University, Nanjing, China
²School of Information Engineering, Henan University of Science and Technology, Luoyang, P. R. China
³Postdoctoral Research Base, Henan Institute of Science and Technology, Xinxiang, P. R. China
⁴School of Public Administration, Hohai University, Nanjing, China

Corresponding author:
Deshun Yin, College of Mechanics and Materials, Hohai University, Nanjing 211100, China.
Email: yindeshun72@163.com
estimated by soil spectra. Then, the PTF is proposed to predict SAS by multilinear regression (MLR) models. The result shows that visible-near infrared spectroscopy may be used directly to estimate SAS indexes with accuracy comparable to the multiple linear models in the PTF approach. M Annabi et al.\textsuperscript{9} apply the regression-kriging method to estimate SAS based on existing geological information as ancillary data. The result shows that the estimation of regression kriging is the same to that of PTF. X Qiao et al.\textsuperscript{10} identify the SAS distribution and explore the possibility of assessing aggregate characters based on hyperspectral technology and the PLS method. It is found that soil spectra significantly respond to SAS, and most of the soil aggregate characters achieve good prediction using spectral technology. Jones et al.\textsuperscript{11} investigate the use of the regression-kriging approach and digital soil mapping techniques to assess the variation in the values of the slaking index in a landscape with diverse agricultural and natural land uses. The result shows that the model has highly accurate predictions produced with leave-one-out cross-validation, giving a Lin concordance correlation coefficient of 0.85 and a root mean square error (RMSE) of 1.1. Similar validation metrics are observed in an independent set consisting of 50 samples. E Afriyie et al.\textsuperscript{12} estimate SAS using mid-infrared spectroscopy. In this work, PLS is employed to build calibration models for slaking index, using a calibration set accounting for 70% of the samples, which are validated by using the rest of the samples. However, the main drawback of the linear regression model is that it usually yields lower prediction accuracy due to the inability to capture the nonlinear and complex association between SAS and soil parameters.\textsuperscript{13}

Machine learning (ML) techniques have proven more effective in soil engineering in recent years. J Rivera and Bonilla\textsuperscript{13} build a PTF for estimating SAS using artificial neural network (ANN) and generalized linear model (GLM) techniques from typically measured soil properties. The experimental results indicate that PTF performs better than GLM. A Besalatpour et al.\textsuperscript{14} estimate SAS from readily available properties. They compare the mean weight diameter (MWD) estimation capabilities of ANN, GLM, MLR, and the adaptive neuro-fuzzy inference system (ANFIS). The correlation coefficient values for MLR, GLM, ANN, and ANFIS are 0.24, 0.35, 0.84, and 0.73, respectively. The result shows that ANN and ANFIS models can predict SAS, whereas linear regression approaches do not perform well. A Besalatpour et al.\textsuperscript{15} explore the applicability of support vector machine (SVM) and ANN to estimate MWD. A parallel genetic algorithm (PGA) is employed to feature selection and clay, aspect parameters, and normalized difference vegetation index (NDVI) are accounted as the redundant features. The result shows that the ANN model achieves higher accuracy in predicting geometric mean diameter (GMD) than MLR and SVM. A Usta et al.\textsuperscript{16} investigate the soil properties that are routinely measured in semi-arid ecosystems and the predictability of the WAS stability using ANN and MLR. The results show that it has better performance in predicting soil properties than MLR. The primary defect of the ANN model is that it is easy to fall into the local optimum, and the convergence rate is slow. In addition, the optimal solution cannot be obtained. M Zeraatpisheh et al.\textsuperscript{17} attempt to predict the SAS indices MWD, GMD, WAS using digital soil mapping and ML models based on the environmental covariates. Results demonstrate that the random forest (RF) model outperformed MWD, GMD, and WAS. P Bhattacharya et al.\textsuperscript{18} present MLR, ANN, and SVM to predict the MWD of agricultural soils using sand, silt, clay, bulk density (BD), and organic carbon (OC) as input variables. The experimental results show that the prediction capability of SVM is better than that of the MLR and ANN models with the same number of input parameters and data points. Y Bouslihim et al.\textsuperscript{19} propose ML approaches to predict SAS. In the study, they compared the ability of RF and MLR to predict MWD as an SAS index. The results achieved are acceptable for predicting SAS and similar for both models. Nevertheless, the above studies confine to land use or a small area and have the disadvantage of uneasily measured variables such as normalized difference vegetation index (NDVI) and remote-sensing attributes. Therefore, ML models for SAS estimation still need to be explored and improved.

We build a reliable model for predicting WAS based on SVM in this work. To the best of our knowledge, it is an original work that SVM is applied to the WAS prediction. The main contribution of this work list as follows: first, we establish the SVM model; second, we build the WAS prediction method based on the SVM model; and third, the proposed method is compared to existing models regarding performance. Finally, we reveal the relevant parameters that are decisive for SAS under this method.

The rest of this article is organized as follows: section “WAS prediction based on SVM” introduces the WAS predicting process based on SVM. Section “Results” proposes the experimental results. Finally, section “Conclusion” concludes the study.

**WAS prediction based on SVM**

**Data preprocessing and selection**

In this work, 134 sample data\textsuperscript{20} are chosen to develop the WAS model, which has already eliminated the influence caused by differences in various measurement procedures. The percentage content of sand, silt, and clay in the soil sample is obtained by the hydrometer technique.\textsuperscript{21} The OC is determined by the wet oxidation
method.\textsuperscript{22} The particle density (DP) is measured by the pycnometer.\textsuperscript{23} WAS is measured by the following procedure analogous to Nimmo and Perkins.\textsuperscript{24} After air-dried, pre-wetted, wetted, shok, sieved, re-dried, and weighed, the wet aggregate stability is calculated by equation (1)

$$\text{WAS\%} = \frac{w_{ds} - w_{da}}{w_{dry} - w_{sand}} \times 100\%$$  \hspace{1cm} (1)

where $w_{ds}$ is the mass of soil dispersed in solution, $w_{da}$ is the mass of dispersant, $w_{sand}$ is the mass of sand particles, and $w_{dry}$ is the mass of the initial air-dried aggregates (1–2 mm).

Variables in the data set have different dimensions and orders of magnitude, which leads to significant differences in the levels of various indicators. Suppose original variables are used directly for analysis. In this case, the variable with a higher value will be strengthened whereas that with a lower one will be weakened. Data standardization is needed to convert the raw data set to a standard format to guarantee the model reliability. The zero-mean normalization is considered to make the data conform to the standard normal distribution as in equation (2)

$$\bar{x}_j = \frac{x_j - \bar{x}}{\sigma}$$  \hspace{1cm} (2)

where $\bar{x}_j$ denotes the normalized value of $x_j$, $\bar{x}$ is the mean of indicator $x$, and $\sigma$ is the variance of $x$. First, 80\% of the data set is split randomly to calibrate the model, and the rest 20\% for testing. To improve the reliability and stability, the cross-validation is applied to separate the calibration data into 10 subsets. Nine subsets are used to train and tune the model, while the remaining part is used for validation.

### WAS model based on SVM

The SVM is a non-parametric learning method,\textsuperscript{25} which is widely recognized to solve small sample, nonlinear, and high-dimensional problems. For a WAS data set: $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, $x_i$ is the soil factors affecting WAS, and $y_i$ is the WAS value. The WAS prediction based on SVM is to establish a function between soil factors and WAS,\textsuperscript{26} as shown in equation (3)

$$y_i = w^T x_i + b$$  \hspace{1cm} (3)

where $w$ is the weight vector, and $b$ is the bias.

The parameter determination of $w$ and $b$ can be transformed into the following optimization problem,\textsuperscript{27} as shown in equation (4)

$$\min \frac{1}{2}||w||_2^2 + C \sum_{i=1}^{m} (s_i^+ + s_i^-)$$  \hspace{1cm} \text{s.t.} \hspace{0.5cm} \begin{align*}
h(x_i) - w \phi(x_i) - b & \leq \delta + s_i^- \vspace{0.5cm} \\
h(x_i) - w \phi(x_i) - b & \geq \delta + s_i^+ \vspace{0.5cm} \\
s_i^+, s_i^- & \geq 0 \end{align*}$$  \hspace{0.5cm} (4)

where $C$ is the regularization coefficient, $\delta$ is the prediction error, and $s_i^+$ and $s_i^-$ are the slack variables to handle the noise.

Followed by, the Lagrange function is used to handle this optimization problem, as shown in equation (5)

$$\ell(w, b, \lambda^i, \lambda^u, \xi^i, \xi^u, \eta^i, \eta^u) = \frac{1}{2}||w||_2^2 + C \sum_{i=1}^{m} (\xi_i^+ + \xi_i^-) + \sum_{i=1}^{m} \lambda_i^i(-\delta - \xi_i^+ - b - h(x_i) + w \phi(x_i))$$  \hspace{0.5cm} (5)

$$+ \sum_{i=1}^{m} \lambda_i^u(-\delta - \xi_i^- + h(x_i) - b - w \phi(x_i)) - \sum_{i=1}^{m} \eta_i^i \xi_i^+ - \sum_{i=1}^{m} \eta_i^i \xi_i^-$$

where $\lambda_i^i, \lambda_i^u, \xi_i^i, \xi_i^u, \eta_i^i, \eta_i^u \geq 0$ are the Lagrange coefficients.

According to the strong duality theorem,\textsuperscript{28} the optimization problem is shown as equation (6)

$$\max \lambda^i, \lambda^u, \xi^i, \xi^u \min \ell(w, b, \lambda^i, \lambda^u, \xi^i, \xi^u, \eta^i, \eta^u)$$  \hspace{0.5cm} (6)

The min $\ell$ can be obtained by getting the partial derivatives with respect to the parameters $w, b, \xi^i, \xi^u$, as shown in equation (7)

$$\begin{align*}
\frac{\partial \ell}{\partial w} &= 0 \Rightarrow w = \sum_{i=1}^{m} (\lambda_i^u - \lambda_i^i) x_i \\
\frac{\partial \ell}{\partial b} &= 0 \Rightarrow \sum_{i=1}^{m} (\lambda_i^u - \lambda_i^i) = 0 \\
\frac{\partial \ell}{\partial \lambda_i^u} &= 0 \Rightarrow C - \lambda_i^u - \eta_i^u = 0 \\
\frac{\partial \ell}{\partial \lambda_i^i} &= 0 \Rightarrow C - \lambda_i^i - \eta_i^i = 0
\end{align*}$$  \hspace{0.5cm} (7)

Substituting equation (7) into equation (5), the optimal problem could be rewritten, as shown in equation (8)

$$\min \frac{1}{2} \sum_{i=1}^{m} \lambda_i \lambda_i + \sum_{i=1}^{m} \lambda_i \lambda_i x_i^2 + \delta \sum_{i=1}^{m} \lambda_i$$  \hspace{1cm} \text{s.t.} \hspace{0.5cm} \sum_{i=1}^{m} \lambda_i = 0, \lambda_i \in [0, C]$$  \hspace{0.5cm} (8)

where $\lambda_i = \lambda_i^u - \lambda_i^i$. 

\title{Zhai et al.}

3
For nonlinear multi-dimensional data, the SVM maps the original data to a higher-dimensional space by kernel function $K(x_i, x) = \langle \phi(x_i), \phi(x) \rangle$. In this way, the separable hyperplane is identified by transforming the nonlinear problem into the linear one, as shown in equation (9)

$$y = \sum_{i=1}^{m} \lambda_i K(x_i, x) + b^*$$  \hspace{1cm} (9)

**Quadratic programming problem solution based on the sequential minimum optimization**

The sequential minimum optimization (SMO) method is applied to handle the quadratic programming problem in equation (6) to determine $w$ and $b$. To facilitate the solution, make the following transformation

$$\begin{align*}
\hat{\lambda}_i &= \lambda_i^k + d_i, \\
\lambda_i^k &= \lambda_i^k + d_i, \\
d_i &= -d_j
\end{align*}$$  \hspace{1cm} (10)

Then, substituting equation (10) into equation (8), the optimal problem with two variables is reduced, as shown in equation (11)

$$\min_{\lambda_i, \lambda_j} \frac{1}{2} \sum_{i,j} (K_{ij} - 2K_{ij}^k + K_{ij}^k + \delta)d_j^2 + |\nabla F(\lambda_i^k) - \nabla F(\lambda_j^k)|d_j$$

$$s.t. L \leq d_j \leq H$$  \hspace{1cm} (11)

where

$$\begin{align*}
\nabla F(\lambda_i^k) &= \begin{bmatrix} K_{ii} & K_{ij} \\
K_{ji} & K_{jj} \end{bmatrix} \begin{bmatrix} \lambda_i^k \\
\lambda_j^k \end{bmatrix} + \begin{bmatrix} y_i \\
y_j \end{bmatrix} + \delta \begin{bmatrix} d_i \\
- d_j \end{bmatrix}
\end{align*}$$

$$\begin{align*}
L &= \max(\lambda_i - C, - C - \hat{\lambda}_i) \\
H &= \min(\hat{\lambda}_i + C, C - \lambda_i)
\end{align*}$$

and $k$ represents the $k$th iteration.

Suppose that the optimal parameters $\lambda_i^k, i = 1, 2, \ldots, m$ and $b^*$ are obtained, the regression function is given by equation (12)

$$y = \sum_{i=1}^{m} \lambda_i K(x_i, x) + b^*$$  \hspace{1cm} (12)

Commonly, linear, polynomial, and Gaussian functions are used as kernel functions. The linear kernel function has the advantages of simplicity and strong interpretability, but it can only be used to solve linearly separable problems. Polynomial kernel functions can be used to solve nonlinear problems, but many hyperparameters are generally used in the case of small power. The Gaussian kernel could map the raw data into the high dimensionality space and only need to tune two hyperparameters. Thus, in this study, the Gaussian function is used as kernel function, as shown in equation (13)

$$K(x_i, x) = \exp\left(-\gamma ||x_i - x||^2\right)$$  \hspace{1cm} (13)

where $\gamma$ means the Gaussian coefficient. The prediction performance is achieved by tuning the values of $\gamma$ and $C$ by the cross-validation process at the calibration stage. The pseudocode of the WAS prediction algorithm is expressed in Table 1.

### Results

**Exploratory data analysis**

The descriptive statistical results of the WAS are visible in Table 2. As shown in Table 2, the mean WAS value for the entire data set is 64.4%. For other soil parameters, the clay, sand, silt, and OC have mean values of 16.94, 27.07, 55.99, and 0.87. The mean value of DP is 2.48 g/cm³. The highest mean value of WAS is grass soil (61.80%). For the rest of the soil properties, the cultivated soil shows the highest amount of clay and silt, while the highest level of sand is observed on the bare land. OC is different in various land uses, and the increase of the mean values in the order bare < crop < grass. The change of DP is insignificant in all land uses.

### Table 1. The WAS prediction algorithm.

| INPUT: Choose training set randomly (80% data set) |
|---|
| Procedure: |
| 1. Data preprocessing |
| 2. Selection of the kernel function and give the range of hyperparameters $C$ and $\gamma$ |
| 3. For $C = C_{\min}$ : $C_{\max} : \Delta C$; |
| 4. For $\gamma = \gamma_{\min}$ : $\gamma_{\max} : \Delta \gamma$; |
| 5. Initialization variables: $d$, $b_{low}$ and $b_{up}$, $\lambda^0 = 0$, set $k = 0$ |
| 6. while (false) do |
| 7. $\lambda_i, \lambda_j$ ← select parameters to optimize |
| 8. Solve the equation (11) |
| 9. Update $\hat{\lambda}_i, \hat{\lambda}_j ← \hat{\lambda}_i + 1, \hat{\lambda}_j + 1$ |
| 10. $k ← k + 1$ |
| 11. End while, Obtained $\hat{\lambda}, b$ |
| 12. Establish regression function (12) |
| 13. Calculate loss value |
| 14. IF $\text{error}(C, \gamma) < \text{error}(C_{new}, \gamma_{new})$ |
| 15. ELSE IF $(C, \gamma) ← (C_{new}, \gamma_{new})$ |
| 16. ENDFOR |
| 17. ENDFOR |
| 18. Obtain optimal $C$ and $\gamma$ corresponding to min{error} |
| 19. Verified model performance in test data set (20% data set) |
| OUTPUT: target value $y$ |

### Table 2. Descriptive statistical results of the WAS. |

| Soil property | Mean value |
|---|---|
| WAS | 64.4% |
| Clay | 16.94 g/cm³ |
| Sand | 27.07 g/cm³ |
| Silt | 55.99 g/cm³ |
| OC | 0.87 g/cm³ |

Note: The descriptive statistical results are visible in Table 2.
Pearson’s correlation coefficients among the soil properties in the total data set and in the different land-use patterns are visible in Figure 1. In the various land-use types, the OC is positively correlated with the WAS, while a negative correlation is presented between the DP and the WAS, and all the OC correlations are more significant than 0.4. A negative relationship is observed between clay and WAS (except for grass). The silt is associated with the WAS positively in all data set, bare and grass land uses, while it is uncorrelated with the DP and the WAS.

### Table 2. Descriptive statistics for WAS in various land uses.

| Land use | N  | Clay   | Silt   | Sand   | DP     | OC     | WAS   |
|----------|----|--------|--------|--------|--------|--------|-------|
|          | %  | %      | %      | g/cm   | %      | %      | %     |
| Bare     | 56 | Minimum 6.68 | 6.98   | 36.2   | 2.24   | 0.37   | 33.73 |
|          | Maximum 35.22 | 42.03   | 79.54  | 2.62   | 1.10   | 95.92 |
|          | Mean 16.74    | 24.79   | 58.57  | 2.50   | 0.77   | 64.00 |
| Crop     | 47 | Minimum 7.36 | 13.44  | 34.2   | 2.26   | 0.1    | 25.14 |
|          | Maximum 25.71 | 42.94   | 73.68  | 2.65   | 1.95   | 91.86 |
|          | Mean 17.41    | 28.96   | 53.63  | 2.48   | 0.87   | 61.80 |
| Grass    | 31 | Minimum 9.14 | 13.60  | 36.29  | 2.21   | 0.49   | 43.16 |
|          | Maximum 27.97 | 48.01   | 74.15  | 2.55   | 2.05   | 93.65 |
|          | Mean 16.59    | 28.31   | 55.1   | 2.44   | 1.04   | 69.04 |
| All soil | 134| Minimum 6.68 | 6.98   | 34.2   | 2.22   | 0.1    | 25.14 |
|          | Maximum 35.22 | 48.01   | 79.54  | 2.65   | 2.05   | 95.92 |
|          | Mean 16.94    | 27.07   | 55.99  | 2.48   | 0.87   | 64.40 |

DP: particle density; OC: organic carbon; WAS: wet aggregate stability.

**Figure 1.** The correlation coefficients among soil properties and WAS for various land uses: (a) all data sets, (b) crop data set, (c) bare data set, (d) grass data set.
WAS in the crop soil. The sand is negatively correlated with WAS in grassland and has weak positive correlations in cropland, but there are no correlations in all data sets and bare land.

The WAS prediction based on SVM

A couple of standard statistical indices are applied, including coefficient of determination ($R^2$) and RMSE, to evaluate the performance. In this work, the parameters $C$ and $\gamma$ are changed, respectively, from 5 to 25 and 0.01 to 0.50 by an increment of 1 and 0.01. Then, the optimal hyperparameters are determined by the Grid-search approach. Subsequently, the validation is applied to obtain the best parameters for WAS prediction by considering $R^2$. Figure 2 displays the result of cross-validation in a three-dimensional (3D) space. As indicated in Figure 2, the minimum value of the cross-validation RMSE is 0.6536, and the corresponding parameters of $C$ and $\gamma$ are 5 and 0.5. Moreover, the optimal result of cross-validation $R^2$ appears to check the hyperparameters’ effect. It is shown that SVM model has the optimal parameters $C = 5$ and $\gamma = 0.5$.

Subsequently, SVM model is developed by the optimal hyperparameters. The WAS prediction is compared to the actual value to evaluate the performance of the SVM model, as shown in Figure 3. As indicated, the SVM-based model performs well in the training phase ($R^2 = 0.776$, RMSE = 7.50) and the testing phase ($R^2 = 0.763$, RMSE = 6.81). It is obvious that the SVM model has proper generalization performance in estimating WAS.

The WAS prediction based on ANN

In this work, the ANN algorithm is recommended as a reference to evaluate the performance of SVM for WAS prediction. The “relu” function is used as an activation function and an “adam” solver for optimizing weight and bias to generate the ANN model. Then, the ANN model is obtained by adjusting the hidden layer and the number of neurons in each layer. The prediction results in the training and testing phases are illustrated in Figure 4. The ANN has value of $R^2 = 0.664$, RMSE = 8.16 in the training stage and $R^2 = 0.665$, RMSE = 7.09 in the testing stage. The results indicate that SVM has the edge over ANN in estimating WAS for multiple land-use types.

Comparison with other aggregate stability methods on training and testing data

To evaluate the performance of the proposed method, the SVM model is compared with the existing methods proposed in Usta et al., Zeraatpisheh et al., and Ye et al. First, the training data set is predicted. The $R^2$ and RMSE values of four methods are calculated. The results are shown in Table 3.

As shown in Table 3, when comparing to Usta et al., Zeraatpisheh et al., and Ye et al., the SVM model provides higher accurate estimation, which reduces the RMSE by 5.36, 7.32, and 5.16, and increases the $R^2$ by 0.246, 0.406, and 0.226, respectively. Then, the testing data set is predicted by four methods, and the values of $R^2$ and RMSE is listed in Table 4.

For the testing data set, the SVM increases the $R^2$ value by 0.283, 0.483, and 0.273, respectively, and the RMSE value decreases 6.84, 9.01, and 6.57, respectively. The comparison illustrates that it has higher performance on the WAS estimation based on SVM.

Analysis of influencing factors

The relative importance (RI) of features is assessed using the SVM model with the sensitivity analysis. The RI is shown in equation (14)
RI% = \frac{R^2_{\text{Ommitt}}}{R^2_{\text{All}}} \times \% \quad (14)

where $R^2_{\text{Ommitt}}$ is the determination coefficient by omitting a variable, and $R^2_{\text{All}}$ is the determination coefficient of all variables.

As shown in Figure 5, OC is the most influential variable, accounting for 33.17% of the variations in the WAS. As the second and third predominant factors, silt and clay provide 21.47% and 19.08% contribution for WAS, which shows the role of micro-particles in soil aggregates, followed by DP (16.46%). By contrast, sand (9.81%) is not critical for WAS.

Significant properties (OC, silt, and clay) are extracted from Pearson’s correlation and the RI analysis. In this work, the WAS is statistically positively and inversely correlated with OC and DP, respectively, which is in line with other experts.\textsuperscript{33,34} Although A Usta et al.\textsuperscript{16} and C Gomez et al.\textsuperscript{8} observe that WAS increased with clay content, unusual results are offered in the crop, bare, and entire data categories in this research. These results indicate that OC more strongly controls the soil aggregate than the clay and that the WAS is negatively affected by the clay content.\textsuperscript{34,35} For the whole data, the low range of clay content could result in the different behavior of the soils; moreover, the different behavior in crop and bare areas might be due to the change in the mineralogy of the clay fraction. In a relatively dry soil environment, the carbonate component of clay harmfully influences soil stability.\textsuperscript{36}

Despite a low range of OC content in soils, the result shows that OC is the most significant governing property to estimate WAS. OC is the main cementation factor in forming stable aggregates, which forms organo-mineral assemblages with other substances.\textsuperscript{37} The application of OC has beneficial impacts on improving soil structure, soil fertility, and reducing surface crusts, which is conducive to plant growth.\textsuperscript{38,39} Furthermore, the application of organic carbon is easier to achieve than altering the texture of the soil in large areas.

In addition to soil properties, the remarkable effect of land use and management in WAS can be observed in Table 2. As evident in the table, the WAS differs in various land uses and decreases in the order of grass > bare > crop.\textsuperscript{39} There are abundant plant roots in grasslands, a higher abundance of earthworms, and more robust microorganism activity, promoting soil stability.\textsuperscript{40} By contrast, the tillage destructs the soil...
structure and reduces biological activity such as earthworm abundance and microbial population.\cite{41,42} Consequently, the average WAS value of the crop soil is lower than that of bare soil. This finding is in line with the observations of S An et al.\cite{43} In general, WAS needs to be improved by increasing vegetation and reducing human disturbance.

**Conclusion**

This work uses SVM to develop the estimation model for WAS in multiuse soils. The existing models are applied to the WAS data set and compared with the proposed method. The result shows that SVM has the highest performance in predicting WAS with readily measured parameters. Furthermore, WAS is affected by varying parameters and management models. Although the importance of soil parameters is different in various land uses, OC is the most significant factor for predicting WAS. Owing to the broader range of soil components and land uses, the modeling of WAS could be applied to evaluate soil quality in the area short of soil aggregate determination. Our work could be feasible and effective in employing SVM-based model to determine WAS for multiuse soils. In the future, we will establish a larger-scale data set to verify the method comprehensively.

**Table 3.** Comparing the performance with existing methods on training data.

| Methods                      | RMSE | $R^2$ |
|------------------------------|------|-------|
| The proposed method          | 7.5  | 0.776 |
| Usta et al.\cite{16}          | 12.86| 0.53  |
| Zeraatpisheh et al.\cite{17}  | 14.82| 0.37  |
| Ye et al.\cite{32}           | 12.66| 0.55  |

RMSE: root mean square error.

**Table 4.** Comparing the performance with existing methods on testing data.

| Methods                      | RMSE | $R^2$ |
|------------------------------|------|-------|
| The proposed method          | 6.81 | 0.763 |
| Usta et al.\cite{16}          | 13.65| 0.48  |
| Zeraatpisheh et al.\cite{17}  | 15.82| 0.28  |
| Ye et al.\cite{32}           | 13.38| 0.49  |

RMSE: root mean square error.
Acknowledgements

The authors thank the referees for their constructive comments, the Editor-in-Chief for helpful suggestions, and the reviewers of the article who helped in improving this article significantly.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the National Natural Science Fund (no. 11872173), in part by the Science and Technology Major Project of Xinxiang City under grant no. 21ZD003, in part by the Key Scientific and Technological Project of Henan Province under grant nos 222102310087, and 2021 02210388, in part by the Young Scholar Training Program of Colleges and Universities in Henan Province under grant nos 20A520002 and 21A520001.

ORCID iDs

Ruizhi Zhai https://orcid.org/0000-0002-8883-9454
Jianping Wang https://orcid.org/0000-0002-8159-1054

References

1. Amézqueta E. Soil aggregate stability: a review. J Sustain Agric 1999; 14(2–3): 83–151.
2. Yan F, Shi Z, Li Z, et al. Estimating interrill soil erosion from aggregate stability of Ultisols in subtropical China. Soil Till Res 2008; 100(1–2): 34–41.
3. Xiao H, Liu G, Liu P, et al. Developing equations to explore relationships between aggregate stability and erodibility in Ultisols of subtropical China. Catena 2017; 157: 279–285.
4. Canton Y, Sole-Benet A, Asensio C, et al. Aggregate stability in range sandy loam soils relationships with runoff and erosion. Catena 2009; 77: 192–199.
5. Almajmaie A, Hardie M, Acuna T, et al. Evaluation of methods for determining soil aggregate stability. Soil Till Res 2017; 167: 39–45.
6. Saygm S, Cornélis W, Erpul G, et al. Comparison of different aggregate stability approaches for loamy sand soils. Appl Soil Ecol 2012; 54: 1–6.
7. Cañasveras J, Barron V, Campillo M, et al. Estimation of aggregate stability indices in Mediterranean soils by diffuse reflectance spectroscopy. Geoderma 2010; 158(1–2): 78–84.
8. Gomez C, Bissonnais Y and Annabi M. Laboratory Vis-NIR spectroscopy as an alternative method for estimating the soil aggregate stability indexes of Mediterranean soils. Geoderma 2013; 209: 86–97.
9. Annabi M, Raclot D, Bahri H, et al. Spatial variability of soil aggregate stability at the scale of an agricultural region in Tunisia. Catena 2017; 153: 157–167.
10. Qiao X, Wang C, Feng M, et al. Hyperspectral response and quantitative estimation on soil aggregate characters. Catena 2021; 202: 45–56.
11. Jones E, Filippi P, Wittig R, et al. Mapping soil slaking index and assessing the impact of management in a mixed agricultural landscape. Soil 2021; 7(1): 33–46.
12. Afriyie E, Verdoordt A and Mouazen A. Estimation of aggregate stability of some soils in the loam belt of Belgium using mid-infrared spectroscopy. Sci Total Environ 2020; 744: 1–14.
13. Rivera J and Bonilla C. Predicting soil aggregate stability using readily available soil properties and machine learning techniques. Catena 2020; 187: 1–9.
14. Besalatpour A, Ayoubi S, Hajabbasi MA, et al. Estimating wet soil aggregate stability from easily available properties in a highly mountainous watershed. Catena 2013; 11: 72–79.
15. Besalatpour A, Ayoubi S, Hajabbasi A, et al. Feature selection using parallel genetic algorithm for the prediction of geometric mean diameter of soil aggregates by machine learning methods. Arid Land Res Manag 2014; 28(4): 383–394.
16. Usta A, Yilmaz M and Kocamanoglu Y. Estimation of wet soil aggregate stability by some soil properties in a semi-arid ecosystem. Fresen Environ Bull 2018; 27(12A): 9026–9032.
17. Zeraatpisheh M, Ayoubi S, Mirbagheri Z, et al. Spatial prediction of soil aggregate stability and soil organic carbon in aggregate fractions using machine learning algorithms and environmental variables. Geoderma Region 2021; 27: e00440
18. Bhattacharya P, Maity P, Ray M, et al. Prediction of mean weight diameter of soil using machine learning approaches. Agronom J 2021; 113(2): 1303–1316.
19. Bouslihim Y, Rochdi A and Paaza N. Machine learning approaches for the prediction of soil aggregate stability. Heliyon 2021; 7(3): e06480.
20. Rahmati M, Weihermuller L, Vanderborght J, et al. Development and analysis of the Soil Water Infiltration Global Database. Earth Syst Sci Data 2018; 10(3): 1237–1263.
21. Gee GW and Or D. Particle size analysis. In: Dane JH and Topp GC (eds) Methods of soil analysis: part 4—physical methods (Soil Science Society of America Book Series 5). Madison, WI: Soil Science Society of America, 2002, pp.255–293.

22. Nelson DA and Sommers L. Total carbon, organic carbon, and organic matter. In: Page AL (ed.) Methods of soil analysis: part 2—chemical and microbiological properties (Agronomy Series No. 9). Madison, WI: Soil Science Society of America, 1983, pp.539–579.

23. Flint AL and Flint LE. Particle density. In: Dane JH and Topp GC (eds) Methods of soil analysis: part 4—physical methods (Soil Science Society of America Book Series 5). Madison, WI: Soil Science Society of America, 2002, pp.229–240.

24. Nimmo JR and Perkins KS. Aggregate stability and size distribution. In: Dane JH and Topp GC (eds) Methods of soil analysis: part 4—physical methods (Soil Science Society of America Book Series 5). Madison, WI: Soil Science Society of America, 2002, pp.317–328.

25. Vapnik V. The nature of statistical learning theory. Berlin: Springer, 1999.

26. Smola A and Schölkopf B. A tutorial on support vector regression. Stat Comput 2004; 14(3): 199–222

27. Barbero A and Dorronsoro J. Cycle-breaking acceleration for support vector regression. Neurocomputing 2011; 74(16): 2649–2656.

28. Bertsekas D, Nedic A and Ozdaglar A. Convex analysis and optimization, vol. 1. Belmont, MA: Athena Scientific, 2003.

29. Cortes C and Vapnik V. Support-vector networks. Mach Learn 1995; 20(3): 273–297

30. Platt J. Fast training of support vector machines using sequential minimal optimization. In: Schölkopf B and Burges CJC (eds) Advances in Kernel methods. Cambridge, MA: MIT Press, 1998, pp.185–208.

31. Tan M, Song X, Yang X, et al. Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: a comparative study. J Nat Gas Sci Eng 2015; 26: 792–802.

32. Ye L, Tan W, Fang L, et al. Spatial analysis of soil aggregate stability in a small catchment of the Loess Plateau, China: H. Spatial prediction. Soil Till Res 2019; 192: 1–11.

33. Martinez E, Fuentes J, Silva P, et al. Soil physical properties and wheat root growth as affected by no-tillage and conventional tillage systems in a Mediterranean environment of Chile. Soil Till Res 2008; 99(2): 232–244.

34. Pituello C, Dal N, Francioso O, et al. Effects of biochar on the dynamics of aggregate stability in clay and sandy loam soils. Eur J Soil Sci 2018; 69(5): 827–842.

35. Almajmaie A, Hardie M, Doyle R, et al. Influence of soil properties on the aggregate stability of cultivated sandy clay loams. J Soil Sediment 2016; 17(3): 800–809.

36. Dimoyiannis D. Wet aggregate stability as affected by excess carbonate and other soil properties. Land Degrad Dev 2012; 23(5): 450–455.

37. Tejada M, Garcia C, Gonzalez J, et al. Organic amendment based on fresh and composted beet vinasse. Soil Sci Soc Am J 2006; 70(3): 900–908.

38. Karami A, Homaei M, Afzalinia S, et al. Organic resource management: impacts on soil aggregate stability and other soil physico-chemical properties. Agric Ecosyst Environ 2012; 148: 22–28.

39. Zhang S, Wang R, Yang X, et al. Soil aggregation and aggregating agents as affected by long term contrasting management of an Anthrosol. Sci Rep 2016; 6: 1–11.

40. Kavdir Y, Ozcan H, Ekinici H, et al. The influence of clay content, organic carbon and land use types on soil aggregate stability and tensile strength. Turk J Agric For 2014; 28(3): 155–162.

41. Fiorini A, Boselli R, Maris S, et al. Soil type and cropping system as drivers of soil quality indicators response to no-till: a 7-year field study. Appl Soil Ecol 2020; 155: 1–9.

42. Ayoubi S, Karchegani P, Mosaddeghi M, et al. Soil aggregation and organic carbon as affected by topography and land use change in Western Iran. Soil Till Res 2012; 121: 18–26.

43. An S, Mentler A, Mayer H, et al. Soil aggregation, aggregate stability, organic carbon and nitrogen in different soil aggregate fractions under forest and shrub vegetation on the Loess Plateau, China. Catena 2010; 81(3): 226–233.