Improving the undulation estimation accuracy by Genetic Algorithm based Least Squares Support Vector Machine

Chia-Hsin Chen (1), Lao-Sheng Lin (1)

[1] Department of Land Economics, National Chengchi University
No. 64, Sec.2, Zhi-Nan Rd, Wenshan District, Taipei City 11605, Taiwan(R. O. C)
Email: 106257032@nccu.edu.tw; lslin@nccu.edu.tw

KEY WORDS: Undulation model, Least Squares Support Vector Machine (LSSVM), Genetic Algorithm (GA), Orthometric height, Ellipsoidal height

ABSTRACT: Traditionally, the orthometric height H used in engineering application can be derived by leveling, which requires high cost of labor and time. On the other hand, the ellipsoidal height h derived by Global Positioning System (GPS) has the advantage of lower cost. The method of GPS Levelling can be applied to obtain the orthometric height from GPS-derived data. And for the transformation between ellipsoidal heights h and orthometric heights H, undulation N with sufficient accuracy is the main study goal.
There exist a number of methods for approximating the undulation model. The polynomial method is the most widely used method to fit the geoidal undulation. However, the polynomial fitting method has its limitation when determined the undulation model in large areas with complex terrain. In order to improve the undulation estimation accuracy, the Genetic Algorithm (GA) is first used to search and optimize the parameters of LSSVM (i.e., LSSVM(GA)), and then use LSSVM(GA) to establish the undulation model.
In this paper, 283 benchmark points distributed throughout the central part of Taiwan region with its orthometric height, ellipsoidal height and plane coordinates were used as test data. According to the test results, the accuracies of undulation estimation are improved about 42.83% (reduced from 0.0523m to 0.0299m) after using genetic algorithm based least squares support machine. The proposed method, LSSVM(GA), and test results will be presented in this paper.

1. INTRODUCTION

Orthometric heights are referred to the geoid. They can be used in engineering application with their physical meaning. Traditionally, orthometric heights can be obtained by spirit leveling, which are quite arduous and time-consuming (Kavzoglu and Saka, 2003). In comparison to leveling, Global Positioning System (GPS) provides more practical, rapid, precise observation and can obtain three-dimensional coordinate simultaneously anywhere on the earth (Gullu et al., 2011). However, GPS-derived ellipsoidal height is merely geometric value. To be able to use ellipsoidal heights in most engineering and surveying projects, their transformation to orthometric heights can be conducted according to the following equation (Featherstone et al., 2000; Lin, 2014; Doganalp and Selvi., 2015):

$$H = h - N \tag{1}$$

where H is the orthometric height, the distance of a point on the earth from the geoid along curved plumb line; h is the ellipsoidal height, the distance of a point on the earth from the surface of the reference ellipsoid along the normal; N denotes the geoid undulation, the difference between WGS84 ellipsoidal height and the orthometric height with respect to the geoid (Gullu et al., 2011).
According to Eq. (1), orthometric heights can be obtained in combination of ellipsoidal heights and undulation value, which is called GPS leveling method (Mårtensson, 2002; Gullu et al., 2011). Thus, the establishment of undulation model is known to be the crucial part of the GPS leveling method.
In the aspect of determined the undulation model, there are two main approaches: the gravimetric method and the geometric method (Gullu et al., 2011). For the gravimetric method, it requires uniform distribution, high precision gravity information and terrain data, which is difficult to achieve in the actual engineering application (Liu et al., 2014). As for the geometric method, many methods like polynomial and Kriging method are common ways to be used. Nevertheless, the shape of the geoid is very complex and the task of approximating the geoid surface by a relatively simple mathematical expression is hardly easy (Stopar et al., 2006). In recent years, there are lots of algorithm such as Artificial Neural Network, Support Vector Machine, Particle Swarm Optimization are proposed to establish the undulation model with the advancement of computer science.
Least Squares Support Vector Machine (LSSVM) were proposed by Suykens and Vandewalle in 1999. It not only has the capability of solving the problems of small sample size, nonlinearity, high dimension and local minimum, but it also requires few parameters and solves the problem fast in comparison with SVM. However, many papers have shown that the parameters selection of LSSVM is still the problem to be solved. Therefore, in this study, the Genetic Algorithm (GA), which has the ability to obtain the globally optimal solution, will be used to optimize the parameters of LSSVM for the purpose of improving the accuracy of the undulation.

2. METHODOLOGY

2.1 The undulation model developed by geometric methods

The orthometric height H is the distance of a point on the earth from the geoid along curved plumb line. In addition, the ellipsoidal height h means the distance of a point on the earth from the surface of the reference ellipsoid along the normal. To be able to use ellipsoidal heights in most engineering and surveying projects, a GPS-derived ellipsoidal height is converted to an orthometric height using a knowledge of the undulation. As shown in Fig.1 and expressed by Eq. (2), the undulation value is the difference between the ellipsoidal height and the orthometric height. (Featherstone et al., 2000; Lin, 2014; Doganalp and Selvi, 2015):

$$N = h - H \tag{2}$$

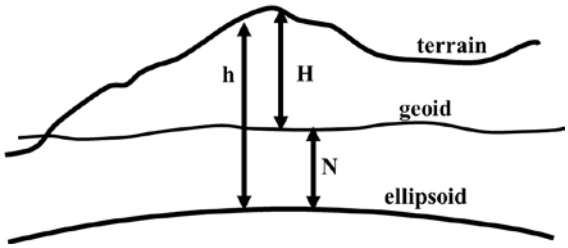


Figure 1. The relationship between orthometric height (H), ellipsoidal height (h) and undulation (N)

According to Eq. (2) , the undulations can be derived by subtracting the orthometric height from the ellipsoidal height of a point (shown as Eq. (3)). Therefore, the undulation data derived by Eq. (3) can be used as training data to establish the undulation model.

$$N(X_i, Y_i) = h_i - H_i, \;\; i = 1,2, \dots, n \tag{3}$$

where (X_i, Y_i) represents the plane coordinate of each point; h_i is the ellipsoidal height of each point; H_i is the orthometric height of each point; N_i is the undulation value of each point.

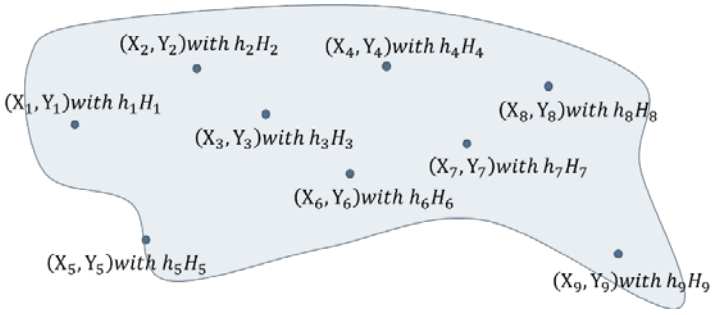


Figure 2. The distribution of points used to train the undulation model.

2.2 Least Squares Support Vector Machine

Least Squares Support Vector Machine (LSSVM) is the improvement of standard Support Vector Machine (SVM). SVM was proposed by Vapnik and was firstly used in classification and non-linear function estimation (Zhang et al., 2009). Nevertheless, the constraint of SVM consists an inequality, which causes complex computation. To solve this problem, Suykens and Vandewalle (1999) constructed LSSVM by substituting the insensitive loss function with the least squares quadratic loss function (Zhang et al., 2009; Kao et al., 2014). With the quadratic loss function, the optimization problem reduces to finding the solution of a set of linear equations (Samui and Kothari, 2011).
To train the geoidal undulation model by LSSVM, the regression function can be expressed as:

$$f(x_i) = w \cdot x_i + b \tag{4}$$

where x_i is the input value, that is, the plane coordinate of each training point (X ,Y); $f(x_i)$ is the output value, that is, the undulation corresponding to the training point; w is the normal to the hyperplane; b is the bias term (Suykens and Vandewalle, 1999).
The optimization problem is given:

$$\text{target function: } J = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} e_i^2 \rightarrow \min \tag{5}$$

$$\text{constraint condition: } w \cdot x_i + b + e_i = y_i \tag{6}$$

where C is the penalty function used to determine trade-off between minimizing the training errors and minimizing the model complexity; e_i is the error between the actual and predicted output at ith sample point.
Lagrange multiplier $\alpha_1, \alpha_2, \ldots\ldots, \alpha_n$ are introduced to change the function J into a quadratic equation:

$$J = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i(w \cdot x_i + b + e_i - y_i) \tag{7}$$

To determine the optimal solution for function J, J are derivative by w, b, e_i and α_i respectively and setting all derivatives equal to zero:

$$\begin{cases} \frac{\partial J}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{n} \alpha_i x_i \\ \frac{\partial J}{\partial b} = 0 \rightarrow \sum_{i=1}^{n} \alpha_i = 0 \\ \frac{\partial J}{\partial e_i} = 0 \rightarrow \alpha_i = 2Ce_i \\ \frac{\partial J}{\partial \alpha_i} = 0 \rightarrow (w \cdot x_i + b + e_i - y_i) = 0 \end{cases} \tag{8}$$

Eq. (8) can be expressed as the following matrix:

$$\begin{bmatrix} I & 0 & 0 & -x^T \\ 0 & 0 & 0 & -u^T \\ 0 & 0 & 2C \cdot I & -I \\ x & u & I & 0 \end{bmatrix} \begin{bmatrix} \omega \\ b \\ e_i \\ \alpha_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ y_i \end{bmatrix} \tag{9}$$

After elimination of e_i and ω, the following linear equation can be obtained:

$$\begin{bmatrix} 0 & u^T \\ u & x_i x_i^T + (2C)^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha_i \end{bmatrix} = \begin{bmatrix} 0 \\ y_i \end{bmatrix} \tag{10}$$

where $x_i = [x_1, x_2 \ldots x_n]^T$; $\alpha_i = [\alpha_1, \alpha_2 \ldots \alpha_n]^T$; $y_i = [y_1, y_2 \ldots y_n]^T$, represents the undulation of each point; $e_i = [e_1, e_2 \ldots e_n]^T$; $u = [1,1, \ldots, 1]^T$; I stands for the unit matrix.
After b and α_i are solved in Eq. (10), a regression function of LSSVM is formed:

$$f(x) = \sum_{i=1}^{n} (\alpha_i x_i) \cdot x + b \tag{11}$$

If the input data are nonlinear, LSSVM maps the training samples from the input space into a higher-dimensional feature space via a mapping function φ to form a linear problem (Huang and Wang, 2006; Kao et al., 2014; Suykens and Vandewalle, 1999). Then the following equation is formed:

$$f(x) = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \cdot \varphi(x) + b \tag{12}$$

Due to the complex computation in higher dimensional space, the kernel trick is introduced to simplify the calculation process. Any function $K(x, x_i)$ satisfying Mercer's condition can be used as the kernel function. By selecting an appropriate kernel function, the nonlinear relation between points' plane coordinates and its corresponding undulation value based on LSSVM is established (Jung et al., 2015).
Due to the kernel function $K(x, x_i)$ equals the inner product of mapping function φ, which is $K(x, x_i) = \varphi(x_i) \cdot \varphi(x)$, Eq. (12) can be expressed as follows:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b \tag{13}$$

As Mustaffa and others mentioned in 2014, an inappropriate selection of kernel function and the parameters of the kernel function may cause the LSSVM prediction model vulnerable to over fitting or under fitting. Therefore, the selection of kernel function and the parameters of the kernel function is known to be a key factor in determining the performance of the support vector machine. Commonly used kernel function are listed as following table:

Table 1. Kernel functions commonly used when appyling LSSVM

Linear kernel (LIN)	$K(x, x_i) = x \cdot x_i$ where $x = (x_1, x_2, \dots, x_n)$ are the training samples	(14)
Polynomial kernel (POLY)	$K(x, x_i) = (x \cdot x_i + t)^d$ where d is the order of polynomial; t represents the intercept of polynomial.	(15)
Radial basis function (RBF)	$K(x, x_i) = \exp(-\frac{\|x-x_i\|^2}{\sigma^2})$ where σ^2 is the bandwidth of the kernel function, which determines the generalization performance and prediction accuracy of RBF (Zhang et al., 2009).	(16)

2.3 Using Genetic Algorithm based LSSVM to establish the undulation model

The parameters of LSSVM play a crucial role in the performance of LSSVM. Nevertheless, inappropriate selection of parameters of LSSVM may lead to over-fitting or under-fitting, and further affects the performance (Jung et al., 2015). So far, there are no guidelines available for parameters selection (Jung et al., 2015). Additionally, it is time-consuming, blind and difficult to select parameters by ways like cross validation or trial and error(Liu et al., 2014; Jung et al., 2015). Therefore, in this study, genetic algorithm (GA) is proposed to optimized the parameters of LSSVM. GA was introduced by Holland in 1975. It is a searching method with the concept of Darwin's theory of natural evolution that allows global optimization (Cai et al., 2015; Jahromi and Ameli, 2018). In this study, the detailed description of steps of gene encoding, fitness function and evolutionary system for GA-based LSSVM parameters optimization were as follows (Yang et al., 2010):

(1) Fitness function evaluation

In this study, gene, referred to the parameters of LSSVM (C and σ^2) , are encoded to string of real number, which formed a chromosome. At the same time, an initial population of chromosomes with two parameters in their allowable ranges is randomly generated. Then, the LSSVM regression function is constructed with the given set of parameters (C and σ^2) using the training data set. Then the performance of the set of parameters (C and σ^2) ,which can be known as the fitness of an individual, is evaluated by the root mean squares error (RMSE) of reference points and check points.

(2) Selection and genetic evolution

To select the appropriate parents for reproduction, the chromosomes with large fitness (better individuals) are selected with higher probability in the stage of selection. Then new offspring will formed with given crossover rate and mutation rate in the stage of genetic evolution.

(3) Create new generation

After the genetic operation, the fitness of new offspring would be calculated again. Then the best N results of offspring or parents would form new generation.

(4) GA iteration

Repeat step (1) to step (3) until the end condition is satisfied or the number of iteration is equal to the presetting maximum.

(5) Termination condition

When the termination criteria are satisfied, the process ends. Then the LSSVM parameters (C and σ^2) with the best fitness are obtained as a result of the algorithm.

(6) Establishment of undulation model

Input the plane coordinates and corresponding undulation of reference points and check points as their input and output data, and establish the undulation model with LSSVM.

2.4 Statistical Analysis Procedures

In order to evaluate the performance of the proposed algorithm, the difference of known undulation N_i^{known} of check points and its corresponding estimated undulation $N_i^{estimated}$ are calculated as Eq.(17) expressed. Besides, Root Mean Squares Error (RMSE) (shown as Eq.(18))will be used in this study to evaluate the performance of the proposed algorithm.

$$\Delta N_i = N_i^{known} - N_i^{estimated} \quad (17)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \Delta N_i \times \Delta N_i}{n}} \quad (18)$$

where n is the number of the check points; N_i^{known} is the known undulation of the check points; $N_i^{estimated}$ is the estimated undulation obtained by either LSSVM or other proposed interpolation methods.

3. STUDY AREA AND TEST DATA

The study area in this paper is located in middle of Taiwan. The test data, which included the GPS and leveling data of the 283 benchmarks of the middle of Taiwan, was collected between 2000 and 2003 by the Satellite Survey Center, Department of Land Administration, Ministry of Interior, Taiwan. The test area size is about 6,321 km^2. The GPS data were collected by the static GPS surveying method with the accuracy of $\pm$36mm, and the leveling data were obtained by the first-order geodetic leveling method with the accuracy of $\pm$8.8mm. (Lin, 2007)

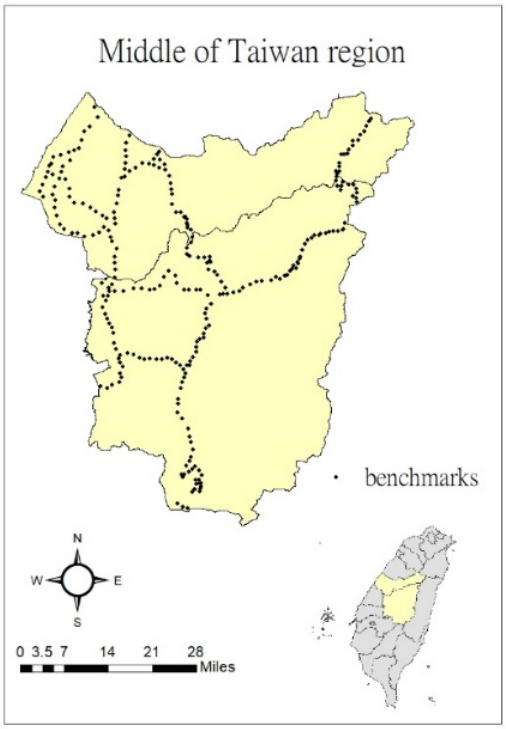


Figure 3. Point distribution map of the 283 benchmarks of the middle of Taiwan.

4. DATA ANALYSIS

4.1 Using LSSVM to establish the undulation model

In order to see how different reference points and check points ratio and different kernel function will affect the performance of LSSVM, 283 test data distributed in the middle of Taiwan are first separated with 1:1、1:2、1:3、2:1、3:1 respectively. Then different kernel functions (POLY, LIN and RBF kernel function) are applied with different reference points and check points ratio to train and evaluate LSSVM model (that is, the undulation model). According to the test results shown in Fig. 4, the best undulation estimation accuracy occurred when applying 1:1 reference points and check points ratio and using RBF kernel function.

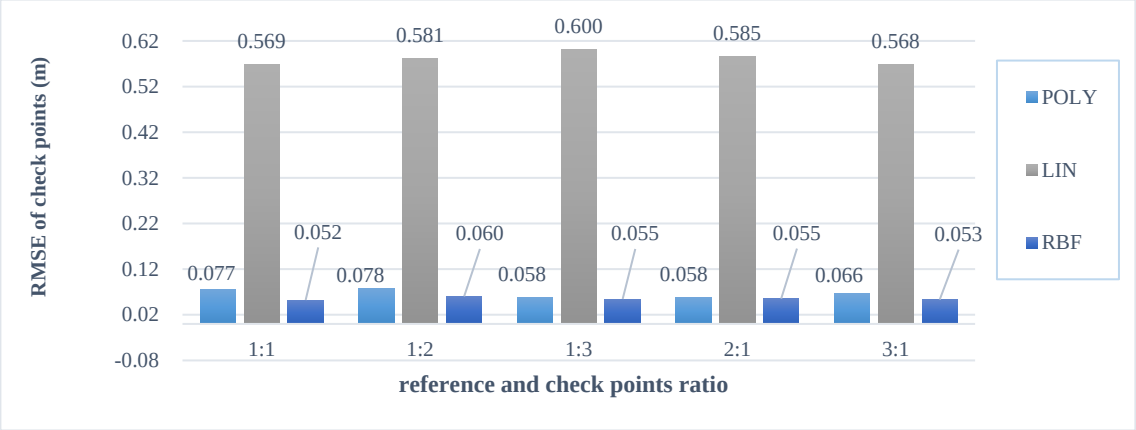


Figure 4. The size of check points' RMSE when applying POLY, LIN and RBF kernel function to train LSSVM model in different reference points and check points ratio.

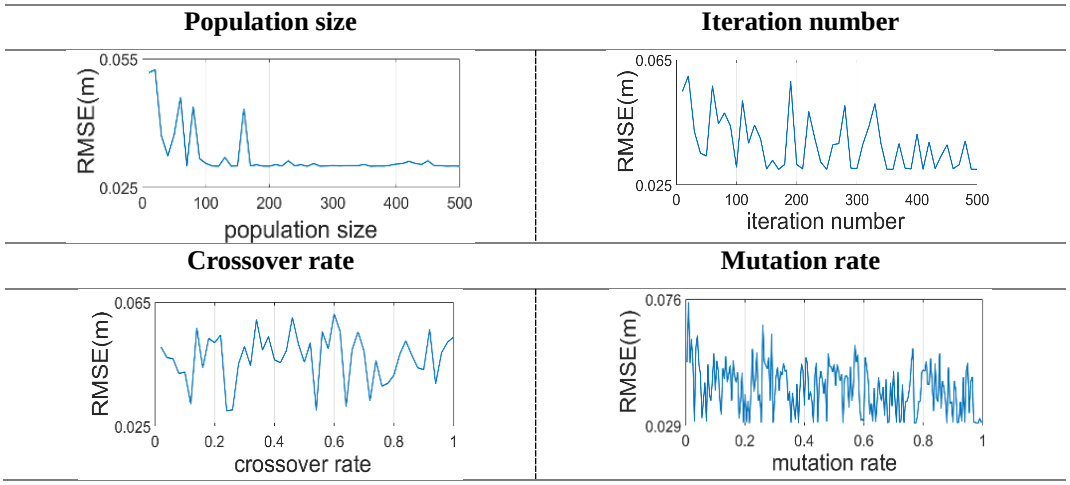
4.2 Using different GA parameters to optimize LSSVM

To see how the different GA parameters set will affect the search of LSSVM parameters and further the performance of the undulation model train by GA-based LSSVM. The main parameters of GA like population size, iteration number, crossover rate and mutation rate are set different value as table 2 shown. With the initial value of population size 10, iteration number 20, crossover rate 0.6 and mutation rate 0.001 (Yang, 1998), the test results are shown as table 3. According to table 3, we know that: (1) The larger the population size, the more accurate the undulation model. (2) There are no significant relationship between undulation estimation accuracy and its corresponding iteration number, crossover rate and mutation rate setting.

Table 2. GA parameters setting table

GA parameters	Population size	Iteration number	Crossover rate	Mutation rate
Initial/terminal value	**10/500**	**10/500**	**0.02/1**	**0.005/1**
Interval	**10**	**10**	**0.02**	**0.005**
Number of GA parameters set	**50 groups**	**50 groups**	**50 groups**	**200 groups**

Table 3. The size of check points' RMSE derived by LSSVM(GA) when using different GA parameters combinations.



To further see the relationship between LSSVM(GA)-derived undulation estimation accuracy (i.e. RMSE of check points) and its corresponding LSSVM parameters optimized by GA, all test results are first rearranged according to the size of check points' RMSE from the smallest one to the largest one. Afterwards, the corresponding LSSVM parameters are shown as Fig. 5. From Fig. 5, it can be seen that: (1) The performance of LSSVM(GA) can cause approximately 4.5 cm difference in accuracy when applying different GA parameters setting. (2) The smaller the size

of C and σ^2, the better the LSSVM performance. (3) When C is smaller than 4000 and $\sigma^2$is smaller than 0.05 simultaneously, the undulation estimation accuracy derived by LSSVM is 0.0299m (the best).

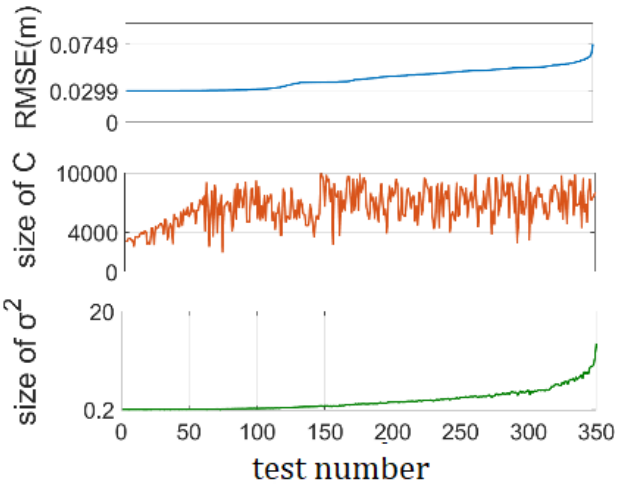


Figure 5. The relationship between RMSE and corresponding LSSVM parameters optimized by GA.

4.3 Comparison between LSSVM and LSSVM(GA)

Table 4 summarizes the performance of LSSVM before and after optimized by GA."LSSVM(RBF)" represents the undulation regression model establish by LSSVM; "LSSVM(RBF+GA)" stands for the undulation regression model establish by GA-based LSSVM. From Table 4, it can be seen that: (1) After optimizing by GA (LSSVM(RBF+GA)), LSSVM(RBF+GA) had improved 42.83% (reduced from 0.0523m to 0.0299m) in accuracy. (2) LSSVM(RBF+GA) is more efficient according to the execution time (reduced from 2.014 s to 1.946 s). (3) From the results of F test, LSSVM(RBF+GA) had significantly improved when comparing to LSSVM(RBF).

Table 4. The comparison between LSSVM(RBF) and LSSVM(RBF+GA).

Estimation method	**execution time (s)**	**RMSE (m)**	**Imrpovement in RMSE (%)**	F-test **(α=5%)**
LSSVM(RBF)	2.014	0.0523	-	Reject H_0
LSSVM(RBF+GA)	1.946	0.0299	42.83	-

5. CONCLUSIONS

In this study, LSSVM is proposed to establish the geoidal undulation model. Additionally, GA is employed to optimize the parameter of LSSVM for the purpose of improving the accuracy of undulation model. With the accurate undulation model derived by GA-based LSSVM, the orthometric height can be obtained from GPS-derived ellipsoidal height instead of conventional spirit leveling. In this study, 283 benchmarks distributed in the middle of Taiwan are used as test data. According to the test result, they show that:(1) The best estimated accuracy 0.0523m is obtained when applying the RBF kernel function and using 1:1 as reference point and check point ratio. (2) According to the relationship between LSSVM parameters (C, σ^2) optimized by GA and their corresponding performance of LSSVM(GA) (i.e. RMSE of check points), it can be seen that when C is smaller than 4000 and $\sigma^2$is smaller than 0.05 simultaneously, using LSSVM(GA) to establish the undulation model can derive the estimation accuracy about 0.0299m. (3) After optimized by GA (LSSVM(GA)), the accuracy of undulation is improved about 42.83% (reduced from 0.0523m to 0.0299m) when comparing to the one which is not optimized. (4) According to the F test, the accuracy of undulation estimation is significantly improved after using GA-based LSSVM.

6. REFERENCE

1. Cai, Z., W. Xu, Y. Meng, C. Shi, & R. Wang, 2016. Prediction of landslide displacement based on GA-LSSVM with multiple factors. Bulletin of engineering geology and the environment, 75 (2), pp. 637-646.
2. Doganalp, S. and Selvi, H. Z., 2015. Local geoid determination in strip area projects by using polynomials, least-

squares collocation and radial basis functions. Measurement, 73, pp. 429-438.
3. Featherstone, W., 2000. Refinement of gravimetric geoid using GPS and leveling data, J. Surv. Eng., 126 (2), pp. 27-56.
4. Gullu, M., Yilmaz, M., and Yilmaz, I., 2011. Application of back propagation artificial neural network for modelling local GPS/levelling geoid undulations: A comparative study. In FIG Working Week, pp. 18-22.
5. Jung, H. C., J. S. Kim, & H. Heo, 2015. Prediction of building energy consumption using an improved real coded genetic algorithm based least squares support vector machine approach, Energy and Buildings, 90, pp. 76-84.
6. Jahromi, M. E. N., and Ameli, M. T., 2018. Measurement-based modelling of composite load using genetic algorithm. Electric Power Systems Research, 158, pp. 82-91.
7. Kavzoglu, T. and Saka, M. H., 2005. Modelling local GPS/levelling geoid undulations using artificial neural networks. Journal of Geodesy, 78 (9), pp. 520-527.
8. Kao, S. P., C. N. Chen, H. C. Huang, & Y. T. Shen, 2014. Using a least squares support vector machine to estimate a local geometric geoid model, Boletim de Ciências Geodésicas, 20 (2), pp. 427-443.
9. Lin, L.S., 2007. Application of a Back-Propagation Artificial Neural Network to Regional Grid-Based Geoid Model Generation Using GPS and Leveling Data, Journal of Surveying Engineering, 133(2), pp. 81-89.
10. Lin, L.S., 2014. Orthometric Height Improvement in Tainan City using RTK GPS and Local Geoid Corrector Surface Models, Journal of Surveying Engineering, 140(1), pp. 35-43.
11. Liu, L. L., T. X. Zhang, M. Zhou, W. Wang, & L. K. Huang, 2014. Research of GPS elevation conversion based on least square support vector machine and BP neural network, Applied Mechanics and Materials, 501, pp. 2166-2171.
12. Mårtensson, S. G., 2002. Height determination by GPS: Accuracy with respect to different geoid models in Sweden. In XXII FIG International Congress, April 19-26 2002, Washington, DC, USA, pp. 106-113.
13. Mustaffa, Z., Yusof, Y. and Kamaruddin, S. S., 2014. Gasoline price forecasting: an application of LSSVM with improved ABC. Procedia-Social and Behavioral Sciences, 129, pp. 601-609.
14. Stopar, B., T. Ambrožič, M. Kuhar & G. Turk, 2006. GPS-derived geoid using artificial neural network and least squares collocation, Survey Review, 38 (300), pp. 513-524.
15. Suykens, J.A.K. & J. Vandewalle,1999. Least Squares Support Vector Machine Classifiers, Neural Processing Letters, 9 (3), pp. 293-300.
16. Samui, P. and Kothari D. P., 2011. Utilization of a least square support vector machine (LSSVM) for slope stability analysis. Scientia Iranica, 18(1) , pp. 53-58
17. Yang, Z., X. S. Gu, X. Y. Liang, and L. C. Ling, 2010, Genetic algorithm-least squares support vector regression based predicting and optimizing model on carbon fiber composite integrated conductivity. Materials & Design, 31 (3), pp. 1042-1049.
18. Zhang, W., Li, C., and Zhong, B., 2009, LSSVM parameters optimizing and non-linear system prediction based on cross validation. In 2009 Fifth International Conference on Natural Computation. 1, pp. 531-535
19. Jahromi, M. E. N., and Ameli, M. T., 2018. Measurement-based modelling of composite load using genetic algorithm. Electric Power Systems Research, 158, pp. 82-91.