Assessing the effect of various training and testing set ratios to model the satellite derived bathymetry

R S Dewi*, Y Lumban-Gaol, A N Safi’I, A Rizaldy, ASyetiawan and A Rahadiati

Geospatial Information Agency (Badan Informasi Geospasial)
e-mail: ratna.sari@big.go.id

Abstract: Depth determination in shallow water area is critical to model for instance, a detailed shoreline position, a change in beach topography, and the potency of beach erosion. Multispectral images can provide a complete map of areas that are difficult to map by conventional hydrographic surveys due to their logistics difficulties and limited spatial coverage. The aim of this study is to evaluate the effect of various training and testing set ratios to model bathymetric data from remotely sensed imagery. This research applied three methods to derive shallow water bathymetric data tested on two subsets located along Tanjung Kelayang coastal areas. The methods combined echo-sounding measurements and the reflectance of blue, green, red and near infrared of Sentinel 2A image with 10 m spatial resolution. In the experiment, the echo-sounding data set was split into training and testing set in three different ratios to see the effect of these various training and testing ratios to the accuracy of all algorithms. From the results, we can see that all models perform well in estimating bathymetric data for the shallow water depth, however, the accuracies were slightly changing by the variation of the training and testing data included in the model. In general, all methods provide a comparable performance for shallow water depth with RMSE less than 1 m and can be used effectively for deriving accurate and updated medium resolution bathymetric maps.

1. Introduction

Generally, the bathymetric data is measured by using echo sounders (Multi/Single Beam Echo Sounders) or LiDAR (Light Detection and Ranging) installed on airborne. Airborne LiDAR bathymetric sensor uses a green wavelength that can estimate depths of up to 70 m [1] and produces a reliable bathymetric data in certain conditions [2]. Both echo sounding and LiDAR methods produce an accurate depth. However, they are high in operational cost and time consuming as well. Furthermore, echo sounders measurement has drawbacks; for safety reason, it cannot be used in shallow water area.

Remote sensing technology can be an alternative tools to map the bathymetry of shallow water regions by pairing remote sensing data with ground-based measurements [3]. The use of remote sensing data is excellence in terms of cost saving, time, and the availability of time series data. Studies related to the Satellite Derived Bathymetry (SDB) emphasize the potential use of optical satellite remote sensing sensors for bathymetric mapping purposes. In this case, analytical, semi-analytical and empirical approaches may be implemented for the estimation of bathymetry of up to 30 m depth [4]. Depth extraction using the SDB method depends on the ability to extract information about the attenuation of light in the water column which is a function of the wavelength. Satellite sensors receiving spectral radiances in shallow water areas are generally consisting of four components: atmospheric scattering, sea surface reflection, in-water scattering, and bottom reflection [5]. Bottom reflectance can be transformed into depth values after removing the atmospheric scattering, surface reflection, and in-water scattering components [6].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
In clear water conditions, Satellite Derived Bathymetry is promising for filling the data gap in shallow water area. In this study, training data is needed to develop SDB model. For this case, single beam echo sounder data was used as training data as well as testing data. The current research aims to assess the effect of various training and testing set ratios to model the Satellite Derived Bathymetry. Sentinel 2A images were used and three algorithms were applied to test their capabilities in extracting bathymetry information.

2. Methodology

2.1. Study Area
This research focuses on deriving bathymetry for the coastal area of Tanjung Kelayang, Indonesia covering more than 81 km$^2$ (Figure 1). The central point of the area is at geographical coordinates 2° 34' 1.2" S and 107° 38' 9.6" E. Tanjung Kelayang is one of the strategic tourism areas in Bangka Belitung Province and a place for coral reef cultivation [7] as it has a relatively clear water with sandy beaches. The prevailing surface current of this area during dry season (June to August) comes from southeast to northwest while during wet season (December to February) comes from north to southeast [8]. The area has a diurnal tide with one high and low tides and the average tidal range is 2.4 m [9]. Shallow waters have an important role both economically and ecologically. Thus, detail bathymetry over the shallow water area is important for management and protection of the area especially around Tanjung Kelayang.

Figure 1. Study area in Tanjung Kelayang with two subsets S1 and S2 in yellow rectangles and RGB 432 of Sentinel 2 is used as background (red pixels represent vegetation, grey pixels are built up, and blue and dark blue pixels are water). White-dashed rectangle shows the position of deep water pixels and green points represent the location of SBES measurements.
2.2. Supporting data and pre-processing

2.2.1. Satellite imagery.
In this research, Sentinel 2A imagery and echo-sounding measurement were used to model the bathymetry of the shallow water area. Sentinel 2A is a high-resolution imagery with a sun synchronous polar orbit for the global and sustained monitoring of Earth land and coastal areas. The satellite imagery is available freely from USGS website [10]. For this study, only four bands consisting of visible and near infrared bands were used with 10 m spatial resolution and the image was recorded in 31 March 2018. Detailed information relating to the spatial and spectral resolutions of full bands of Sentinel 2A can be seen in Table 1.

| Band  | Central Wavelength (μm) | Color range             | Resolution (m) |
|-------|------------------------|-------------------------|----------------|
| Band 1| 0.443                  | Coastal aerosol         | 60             |
| Band 2| 0.490                  | Blue                    | 10             |
| Band 3| 0.560                  | Green                   | 10             |
| Band 4| 0.665                  | Red                     | 10             |
| Band 5| 0.705                  | Vegetation Red Edge     | 20             |
| Band 6| 0.740                  | Vegetation Red Edge     | 20             |
| Band 7| 0.783                  | Vegetation Red Edge     | 20             |
| Band 8| 0.842                  | Near Infrared           | 10             |
| Band 8A| 0.865               | Vegetation Red Edge     | 20             |
| Band 9| 0.945                  | Water vapour            | 60             |
| Band 10| 1.375                | SWIR-Cirrus             | 60             |
| Band 11| 1.610               | SWIR                    | 20             |
| Band 12| 2.190                | SWIR                    | 20             |

2.2.2. Bathymetric measurement
Single Beam Echo Sounder (SBES) is used to build and validate the models (Figure 1). The SBES was collected in August 2018. Tide correction has been applied so that the data has zero tide influence. For each subset, we set different ratios of training and testing data, i.e., 75%:25%, 50%:50%, and 25%:75%, in order to evaluate the accuracy of SDB resulted from the model. These data were selected randomly and used to develop empirical SDB model and to assess the accuracy of the derived SDB estimates.

2.3. Method

2.3.1. Pre-processing of images
Sentinel 2A adopted for this study is Level-1C product which is provided with top of atmosphere (ToA) reflectance. The product has been applied standard radiometric and geometric corrections including orthorectification and spatial registration on a global reference system with sub-pixel accuracy [11]. From the previous study, it was mentioned that an atmospheric correction process reduced the accuracy of SDB estimates [12]. Thus, for this study, only the ToA correction was applied to the image to generate SDB results. However, future study will be required in determining if atmospheric correction will beneficial for SDB. Afterwards, for applying SDB algorithm, two subsets were created denoted as S1 and S2 as seen in Figure 1. For this research, we tested two different datasets; the first dataset consists
of three bands in red, green and blue canals, while the second dataset consists of red, green, blue and near infra-red (NIR) canals.

2.3.2. Water correction method
Kanno [5] and Vinayaraj, Raghavan [13] assumed that the spectral radiance \( L_s \) observed by a satellite sensor in shallow water is consisting of four components: atmospheric scattering \( L_a \), surface reflection \( L_r \), in-water scattering \( L_w \), and bottom reflection \( L_b \) (see Figure 2). Thus, the observed spectral radiance in shallow water \( L_s \) can be expressed by a function of wavelength as:

\[
L_s = L_a + L_r + L_w + L_b
\]  

In this study, water correction method was performed following Lyzenga [14] and Gholamalifard, Esmaili Sari [6]. The method ignores the variation of sea-surface and atmospheric scattering over the water area; it is assumed to be homogenous. Furthermore, the method assumed that in deep water, the observed spectral radiance does not include bottom reflectance. The transformed radiance value \( L_s \) was estimated from the spectral properties of the deep water area which has low spectral value represented by dark blue pixels (in white-dashed rectangle) using RGB 432 in Figure 1. We assumed that the reflectance value in that area must be due to scattering. As in Gholamalifard, Esmaili Sari [6], we calculated the average value of pixels in the deep water area. The measured radiances are transformed according to the following equation:

\[
X_i = \log(L_s)_i - \text{mean}(L_d)_i
\]  

Where \( (L_s)_i \) is the measured radiance in shallow water for band \( i \) and \( (L_d)_i \) is the deep water radiance. The resulting values which are the transformed radiance \( (X_i) \) are linear function of the water depth and showing a linear relation between spectral value in shallow water and deep water. By using this equation, we actually removed the effect of sea-surface and atmospheric scattering.

2.3.3. Depth estimation.
The principle of SDB from passive multispectral image is based on a simple assumption that when a depth increase, the water leaving radiance decreases until reaches a point where bottom reflectance are no longer exist or detectable referred to as deep water [12]. The following section describes three methods applied in this study.

(1) Multiple Linear Regressions. Several studies used multiple linear regressions (MLR) in estimating depth information using multispectral bands over shallow water area [15-17]. Hengel and Spitzer [18] mentioned that this method works by assuming that the bottom reflectance and water composition are constant within all part of the image and the multispectral bands of the image are influenced by the
bottom reflectance. By using this model, echo sounding measurement data is considered as the dependent variable, and transformed radiance $X_i$ is considered as the independent variable. The dependent variable (SBES point) was used to determine the regression coefficient and estimates the bathymetry. The water depth can be estimated by means of linear regression analysis of $X_i$ with known water depth data for a selected of pixels using the following equation [18, 19]:

$$W_d = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_i$$  (3)

where $X_1$, $X_2$, ..., $X_i$ are the transformed radiance derived from Equation 2 for each band, $\beta_0$ is representing the $y$-intercept, while $\beta_1$, $\beta_2$, ..., $\beta_i$ are slope for each band. These $\beta$-coefficients are obtained from the multiple linear regressions with echo-sounding points. For this research, we set standard parameter when applying the algorithm.

(2) Random Forest. Random Forest (RF) is one of the non-linear methods that are suitable to develop regression model that relates satellite imagery to water depth data. RF is formed by creating trees by using the training data, for instance water depth, and predictor variables, e.g., spectral value of satellite image. RF depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The representation of the tree classification found in RF depends on the correlation between the trees. Readers are advised to refer to previous studies for detailed description regarding RF for classifications [20] and RF for depth estimations [21, 22]. For this study, we set various values for ntree parameter. The ntree values were set from 100 to 1000 in the steps of 100. This parameter represents number of trees to grow for the decision tree estimates. However, we set a standard mtry when applying RF. This mtry represents number of variables sampled randomly as candidates at each split.

(3) Support Vector Regression. Initially, Support Vector Machine (SVM) was used for classification and statistical learning purposes [23]. It was developed for binary classification that search for an optimal separating hyperplane between two classes by maximizing the margin between the closest points of the class. These points are lying on the boundaries that are called support vector, and the middle of the margin is the optimal separating plane [24].

Nowadays, this machine learning algorithm has been extended to solve regression problem in particular for bathymetry calculation [25-27]. In regression-based SVM training data is used by the learning machine to define the problem and learn the input-output relation. Readers are advised to refer to Mohamed, Negm [27] and Wang, Liu [26] for a detailed description of the algorithm in deriving depth information. For this experiment, radial basis kernel was used and we set various gamma values and a standard cost parameter of SVM. We set gamma values equal to 0.1, 0.3, 0.5, 0.7, 0.9 and 1.0, whereas cost value was set to 1. The cost is a general penalizing parameter and gamma is the radial basis function-specific kernel parameter [24].

3. Results

3.1. Multiple linear regressions

Table 2 and Figure 3 present the results of accuracy assessment of the MLR by applying three different ratios with two different datasets. From the table, we can see that adding the fourth band (NIR) improved the accuracy of the SDB model. For example, by setting training at 75%, and using three bands of Sentinel, we obtained RMSE values 1.013 and 1.011 for S1 and S2, respectively; whereas, when using four bands, we obtained 0.99 and 0.99 for S1 and S2, respectively. However, setting different ratios to the training and testing data has little influence to the RMSE results (less than 1% for all combinations) and setting 75%:25% of training and testing ratio tends to obtain a higher RMSE value.
Table 2. The comparison of RMSE values resulted when applying MLR using two datasets with various training and testing ratios. B234 and b2348 represent composites 234 and 2348 datasets, respectively (2=blue, 3=green, 4=red, and 8=NIR)

| Training - testing ratio | S1     | S2     | S1     | S2     |
|--------------------------|--------|--------|--------|--------|
|                          | b234   | b2348  | b234   | b2348  |
| 75%:25 %                 | 1.013  | 0.990  | 1.011  | 0.996  |
| 50%:50 %                 | 1.009  | 0.986  | 1.011  | 0.994  |
| 25%:75 %                 | 1.008  | 0.985  | 1.012  | 0.995  |

Figure 3. RMSE values for various training and testing ratios when applying MLR for deriving depth information (see notations in Table 2 for image bands used)

3.2. Random Forest.
We tested various $n_{tree}$ for parameterization of RF. From the results, we can see that variation in $n_{tree}$ values has a little effect to the estimation results (see Figures 4 and Table 3). However, setting higher $n_{tree}$ required a longer processing time.

From Figure 4, we can see that variation in the setting of training and testing ratios influenced the SDB algorithm. For both band combinations, the lowest RMSE were achieved by setting 75%:25% of training and testing ratios, respectively. However, when we set the ratio into 25%:75%, the RMSE values are double the values of the 75%:25% ratio (see Table 3). In this case, reducing the number of training data would give a worse result. Applying 75%:25% for training and testing set ratios only obtained a slightly better result than setting 50%:50% set ratio, but a longer time was needed. Given this fact, ratio 50%:50% for training and testing data may probably be an option when applying RF for depth estimation.
3.3. Support Vector Machine

To estimate depth information by using SVM, various gamma values were tested. From the results in Figure 5, it can be seen that adding NIR band provided a lower RMSE values (<20%). Furthermore, for all subsets, variation in gamma values for i.e., increasing gamma values resulted in a low RMSE. However, an increase of gamma required a longer processing time. From Figures 5 and Table 4, little differences in RMSE values (less than 3%) are obvious when applying various training and testing set ratios. However, ratio 75:25% provided a slightly better result.
Figure 5. RMSE for various gamma values applied to two different band combinations by using three training and testing ratios when performing SVM for deriving SDB at two different subsets (see notations in Table 2 for image bands used)

Table 4. The comparison of RMSE values when applying SVM in subset S1 using two band combinations with various training and testing ratios (see notations in Table 2 for image bands used)

| gamma | b2348 | b234 |
|-------|-------|-------|
|       | 75%:25% | 50%:50% | 25%:75% | 75%:25% | 50%:50% | 25%:75% |
| 0.1   | 0.60 | 0.60 | 0.61 | 0.67 | 0.67 | 0.67 |
| 0.3   | 0.56 | 0.56 | 0.57 | 0.65 | 0.65 | 0.65 |
| 0.5   | 0.54 | 0.55 | 0.55 | 0.64 | 0.64 | 0.64 |
| 0.7   | 0.53 | 0.53 | 0.54 | 0.63 | 0.63 | 0.63 |
| 0.9   | 0.52 | 0.52 | 0.53 | 0.63 | 0.63 | 0.63 |
| 1.0   | 0.51 | 0.51 | 0.53 | 0.63 | 0.63 | 0.63 |

Figure 6 shows the comparison of bathymetry maps of an area in Tanjung Kelayang (see black dashed rectangle in the insert map of Figure 6). From the figure we can see that MLR (Figure 6c) has a large disagreement with the SDB reference image (Figure 6d), in particular at 0-1 m depth (violet colour) which were also indicated by the highest RMSE value 0.99 as compared to other methods.

Depths extracted over the crossprofile line (blue line, A to B in Figure 6) in Figure 7 show that SDB from RF are fit well with the 2018 sounding measurement which are also indicated by the lowest RMSE 0.12 (Figure 6a). However, SVM and MLR curves are also close to the sounding data (green and red curves in Figure 7). The largest deviation of those curves is located near 109 m point of the cross profile line (see yellow dashed-rectangle in Figure 7) at 6-8 m depth.
Figure 6. Comparison of bathymetry map of Tanjung Kelayang (0-13m). Grey colour represents land area. Sea depth is reported as positive value. Bathymetry maps were derived from (a) RF; (b) SVM; (c) MLR; and (d) in situ measurements. Light blue lines show the cross profiles from A to B.

Figure 7. Cross profile showing comparison between SDB (from SVM, MLR and RF) and 2018 in situ measurement (SBES). The cross-profile locations are indicated as blue lines in Figure 6. The profile begins at point A, ending at point B.
4. Discussions
In our experiments, adding NIR band provides a better result. With only three bands in the visible spectrum, multispectral satellites are unable to explain many unknown parameters in the model. Furthermore, the number of bands used in the model is related to the capability of the algorithm to discriminate different bottom types and water masses, thus more bands may produce a more accurate result [28].

Our results demonstrated the outperformance of RF compared to SVM and MLR. However, the RF was influenced by the variation on the training-testing set ratio. Furthermore, it requires a large number of training data in order to have a good result. From the cross profile in Figure 7, RF curve overfitted the in situ measurement curve. This might be due to a high correlation between training and testing data. In this study, the training and testing set ratio was selected randomly, so it was possible that the training and testing data were from the same distribution and location. Therefore, future study will be required in determining that training and testing data are separated.

With error less than 1 m, SVM and MLR are promising for SDB estimation. Moreover, both algorithms were not largely influenced by the variation in the training and testing set ratio indicated by a good result could be achieved by using relatively small number of training data.

Some sources of error in this study could be introduced by the bathymetry data sets from echosounding measurements collected using ships sounding. In shallow water ship sounding may be unreliable. Other source of error could be due to water quality. The methods assume that the water quality condition to be homogenous over the study area. In fact, any differences of water quality among training points will cause error in the algorithm.

5. Conclusions
This research provides a comparative analysis of the three bathymetry models by incorporating variation of training and testing set ratios. RF model obtained an RMSE of 0.12 providing the best model. Whereas, MLR and SVM were also promising and the methods obtained RMSE of 0.99 and 0.62 m, respectively. These results were obtained by using two datasets with three and four bands. Adding NIR band to the data set produced higher accuracy. Further study is needed to account for error in the model and the correlation between training and testing data.

6. Acknowledgements
The authors express their gratitude to Center for Marine and Coastal Mapping of Badan Informasi Geospasial for providing the echosounding data.

7. References
[1] Gao, J., Bathymetric mapping by means of remote sensing: methods, accuracy and limitations. Progress in Physical Geography, 2009. 33(1): p. 103-116.
[2] Pe’eri, S., et al., Satellite Remote Sensing as a Reconnaissance Tool for Assessing Nautical Chart Adequacy and Completeness. Mar. Geod., 2014. 37(3): p. 293-314.
[3] Winterbottom, S.J. and D.J. Gilvear, Quantification of channel bed morphology in gravel-bed rivers using airborne multispectral imagery and aerial photography. Regulated Rivers: Research & Management, 1997. 13, : p. 489-499.
[4] Evagorou, E., et al., Bathymetric maps from multi-temporal analysis of Sentinel-2 data: the case study of Limassol, Cyprus. Adv. Geosci., 2019. 45: p. 397–407.
[5] Kanno, A., Shallow Water Bathymetry from Multispectral Satellite Images: Extensions of Lyzenga’s Method for Improving Accuracy. Coastal Engineering Journal, 2011. 53(4): p. 431–450.
[6] Gholamalifard, M., et al., Bathymetric Modeling from Satellite Imagery via Single Band Algorithm (SBA) and Principal Components Analysis (PCA) in Southern Caspian Sea. Int J Remote Sens, 2013. 7(4): p. 877-886.
[7] Akbar, A., Analisis potensi dan daya dukung ekologi kawasan wisata pantai Tanjung Kelayang dan Tanjung Tinggi untuk wisata pantai, wisata selam dan snorkeling di Kecamatan Sijuk, Kabupaten Belitung, in Sekolah Pasca Sarjana Institut Pertanian Bogor. 2013, IPB: Bogor.
[8] PURISKEL. Tanjung Kelayang - Bangka Belitung. Riset Wisata Bahari 2019 [cited 2019 14 July 2019]; Available from: http://pusriskel.litbang.kkp.go.id/index.php/en/data/riset-wisata-bahari/category/486-tanjung-kelayang-bangka-belitung.
[9] BIG. Online Tidal Prediction. 2019 [cited 2017 1 August 2017]; Available from: Available online: http://tides.big.go.id/.
[10] USGS. EarthExplorer. 2019 [cited 2019 1 April 2019]; Available from: http://earthexplorer.usgs.gov/.
[11] SUHET, Sentinel-2 User Handbook. 2015, ESA: EU. p. 64.
[12] Chénier, R., M.-A. Faucher, and R. Ahola, Satellite-Derived Bathymetry for Improving Canadian Hydrographic Service Charts. ISPRS Int. J. Geo-Inf., 2018. 7(306): p. 15.
[13] Vinayaraj, P., V. Raghvan, and S. Masumoto, Satellite-Derived Bathymetry using Adaptive Geographically Weighted Regression Model. Mar. Geod., 2016. 39(6): p. 458–478.
[14] Lyzenga, D.R., Remote sensing of bottom reflectance and water attenuation parameters in shallow water using aircraft and Landsat data. Int J Remote Sens, 1981. 2(1): p. 71-82.
[15] Clark, R.K., T.H. Fay, and C.L. Walker, Bathymetry calculations with Landsat 4 TM imagery under a generalized ratio assumption. Applied Optics, 1987. 26(19): p. 4036-4038.
[16] Hamilton, M.K., et al., Estimating Chlorophyll Content and Bathymetry of Lake Tahoe Using AVIRIS Data. REMOTE SENS. ENVIRON., 1993. 44: p. 217-230.
[17] Kanno, A. and Y. Tanaka, Modified Lyzenga’s Method for Estimating Generalized Coefficients of Satellite-Based Predictor of Shallow Water Depth. IEEE Geosci. Remote S, 2012. 9(4): p. 715-719.
[18] Hengel, W.V. and D. Spitzer, Multi-temporal water depth mapping by means of Landsat TM. Int J Remote Sens, 1991. 4(4): p. 703-712.
[19] Poliyapram, V., Development of Algorithms for Near-shore Satellite Derived Bathymetry Using Multispectral Remote Sensing Images, in Department of Mathematics, Osaka City University. 2017, Osaka City University: Osaka.
[20] Breiman, L., Random Forests. Machine Learning, 2001. 45: p. 5–32.
[21] Manessa, M.D.M., et al., Satellite-Derived Bathymetry using Random Forest Algorithm and WorldView-2 Imagery. 2016.
[22] Sagawa, T., et al., Satellite Derived Bathymetry Using Machine Learning and Multi-Temporal Satellite Images. Remote Sens., 2019. 11(1155).
[23] Vapnik, V.N., Statistical Learning Theory. 1998, Canada: JOHN WILEY & SONS, INC.
[24] Meyer, D., Support Vector Machines: The Interface to libsvm in package e1071. 2019: Austria. p. 8.
[25] Misra, A., et al., Shallow water bathymetry mapping using Support Vector Machine (SVM) technique and multispectral imagery. Int J Remote Sens, 2018. 39(13): p. 4431-4450.
[26] Wang, L., et al., Bathymetry retrieval from optical images with spatially distributed support vector machines. GIScience & Remote Sensing, 2019. 56(3): p. 323-337.
[27] Mohamed, H., et al., Assessment of proposed approaches for bathymetry calculations using multispectral satellite images in shallow coastal/lake areas: a comparison of five models. Arab J Geosci, 2017. 10(42): p. 41-42.
[28] Bramante, J.F., D.K. Raju, and T.M. Sin, Multispectral derivation of bathymetry in Singapore’s shallow, turbid waters. International Journal of Remote Sensing, 2013. 34(6): p. 2070–2088.