Tide modelling using partial least squares regression

ONUWA OKWUASHI (PhD) [Corresponding Author]

Department of Geoinformatics and Surveying,

University of Uyo, Uyo, Nigeria

Email: onuwaokwuashi@gmail.com

CHRISTOPHER NDEHEDEHE (PhD)

Australian Rivers Institute, Griffith University, Nathan,

Queensland 4111, Australia

e-mail: c.ndehedehe@griffith.edu.au

HOSANNA ATTAI (MSc)

Department of Geoinformatics and Surveying,

University of Uyo, Uyo, Nigeria

Email: hosannaattai@gmail.com

Abstract

This research explores the novel use of the Partial Least Squares Regression (PLSR) as an alternative model to the conventional Least Squares (LS) model for modelling tide levels. The modelling is based on twenty tidal constituents: M2, S2, N2, K1, O1, MO3, MK3, MN4, M4, SN4, MS4, 2MN6, M6, 2MS6, S4, SK3, 2MK5, 2SM6, 3MK7, and M8. The 1st, 2nd, and 3rd PLSR components are selected from 40 PLSR components for the modelling based on the computed variances, Yloadings, and Yscores. The PLSR results are compared to those of the LS. The normality of the model residuals are evaluated by the Jarque-Bera statistical test. The computed probabilities of the normality test for 1st, 2nd, and 3rd PLSR components and LS are $p=0.0611$, $p=0.0656$, $p=0.916$, and $p=0.0517$ respectively, which all indicate $p>0.05$, and imply that the residuals are normally distributed. The nature of tide criterion is verified by computing the Tidal Form Factor F . The computed Tidal Form Factor for the 1st, 2nd, and 3rd PLSR components and LS are $F=0.1794$, $F=0.1696$, $F=0.1599$, and $F=0.1848$ respectively. All the models satisfy the semi-diurnal criterion of $0 \leq F \leq 0.25$ at the 95% confidence level, since the observed tide is characteristically semi-diurnal. The computed Coefficient of Determination for the 1st, 2nd, and 3rd PLSR components and LS are $r^2=0.9134$, $r^2=0.9825$, $r^2=0.9933$, and $r^2=0.7861$ respectively. These results prove that the PLSR model outperformed the conventional LS model, and therefore a viable alternative to the conventional LS model.

Keywords: Tide; partial least squares regression; least squares; modelling

1. Introduction

Tide usually occurs along the coast when there is a rise and fall of sea levels mainly induced by the gravitational forces of the moon, sun and rotation of the earth (Chang & Lin, 2006; de Jong *et al.*, 2010; Cai *et al.*, 2018; Guan *et al.* 2019). The diurnal tide, semi-diurnal tide and mixed tide are the three main categories of tide. Accurate tide forecast is vital for navigation, installation of platforms, petroleum exploration and other coastal activities (Hong *et al.*, 2020).

Least squares (LS) regression is the conventional model for forecasting tide levels (Lee & Jeng, 2002). Okenwa (1978) forecasted the height of tide by applying LS polynomial approximation of the observed time series and by calculating the coefficients of the polynomials used to determine the range ratio and time difference of designated points. Shu (2003) forecasted tidal heights by applying an enhanced system of linear equations of the constraint LS solutions and adopting the Goertzel's recurrence formula of the model. But in recent times several contemporary models have been used for tide forecasting as alternative models to the conventional LS technique. Some of these contemporary models that have been applied to tide forecast are notably Artificial Neural Network (ANN), Kalman Filter (KF), and Support Vector Machine Regression (SVMR). The presentation of the ANN, KF and SVMR tide applications are as follows. Lee (2004) applied the Back-Propagation Neural Network (BPNN) to short-term tide prediction at the Taichung Harbor in Taiwan. The BPNN was compared with the harmonic method. The BPNN was found to be more efficient than the harmonic method. Okwuashi and Olayinka (2017) used the KF to forecast the height of tide based on seven main tidal constituents M2, S2, N2, K2, K1, O1 and P1. The KF was found to outperform the conventional LS technique. Okwuashi and Ndehedehe (2017) investigated the application of the SVMR for forecasting the height of tide based on tidal constituents M2, S2, N2, K2, K1, O1 and P1. The SVMR results were found to be better than those of the conventional LS. It was also noted that unlike the LS, the SVMR has the advantage of not being constrained by the statistical normality criterion of its computed residuals.

These contemporary mathematical models presented have proven to be more robust than the conventional LS, but in the quest for sustainable tide forecasting, there is need for continuous investigation of contemporary models. Hence this research explores the novel use of the Partial Least Squares Regression (PLSR) for forecasting tide levels. The PLSR has been applied in other fields as follows. Cheng and Sun (2017) applied PLSR to near-infrared (NIR) and hyperspectral imaging (HSI) spectral data for rapid and chemical-free modelling and predicting chemical properties of fish muscle. The result of the experiment showed that PLSR is a veritable alternative to the traditional chemical analytical methods such as Kjeldahl, Soxhlet, and chromatography methods for detecting chemical information of fish muscle. Liu *et al.* (2018) used the PLSR to predict reservoir permeability. The parameters used for the modelling were extracted from the Mercury Injection Capillary Pressure (MICP) curves. A comparison of the result of the PLSR with that of the Ordinary Least Squares (OLS) based on the MICP curves proved that the PLSR permeability model was superior to that of the OLS. Straus and Skogestad (2019) developed a new incremental sampling method for the generation of surrogate models based on PLSR as a termination criterion. Compared to existing incremental and adaptive methods, the proposed technique allowed the sampling algorithm to stop without having to fit a surrogate model at each iteration step. Applying the proposed method to a motivating pipe model and two case studies showed that this method allowed a small number of sampling points that corresponded to a regular grid with less than two points in each independent variable. These three PLSR applications showed the robust predictive capability of the PLSR over conventional methods. This research will investigate the novel use of the PLSR for tide forecasting, and will compare its result with the result of the conventional LS model.

2. Partial least squares regression algorithm

PLSR is a statistical technique that bears similarity with the Principal Component Regression (PCR). Instead of seeking hyperplanes of maximum variance between the response and predictor variables, the PLSR finds a linear regression model by projecting the predictor variables and the response variable to a new space (Haenlein & Kaplan, 2004).

The goal of the PLSR is to decompose both the design matrix X and response matrix Y reminiscent of Principal Component Analysis (PCA) (Chen *et al.*, 2018). The PLSR is based on the properties of the Nonlinear Iterative Partial Least Squares (NIPALS). Let us formulate a regression between the scores of the X block and Y block. We assume that the PLSR model consists of outer relations (X block and Y block) and an inner relation (containing both blocks). Let the outer relation for the X block be expressed as,

$$X = TP' + R^* = \sum t_h p'_h + R^*, \quad (1)$$

and the outer relation for Y block as,

$$Y = UQ' + S^* = \sum U_h q'_h + S^*, \quad (2)$$

where R^* and S^* are constants. Let,

$$\hat{u}_h = b_h t_h \quad (3)$$

where,

$$b_h = u'_h t_h / t'_h t_h \quad (4)$$

b_h denotes the regression coefficients. In order to ensure that X and Y have a weak relationship between each other, their principal components are computed separately.

For the X block, step 1: $t_{start} = x_j$, step 2: $p' = t'X/t't$ ($= u'X/u'u$), step 3: $p'_{new} = p'_{old} / \|p'_{old}\|$, step 4: $t = X_p / p'p$, step 5: compare t in step 2 and step 4 if they are equal stop, else go to step 2.

For the Y block, step 1: $u_{start} = y_j$, step 2: $q' = u'Y/u'u$ ($= t'Y/t't$), step 3: $q'_{new} = q'_{old} / \|q'_{old}\|$, step 4: $u = Yq/q'q$, step 5: compare u in step 2 and 4, and if they are equal stop, else go to step 2.

The inner relation can be improved by exchanging the scores. In order to make the two blocks exchange information we let t and u change place in step 2. Therefore the algorithm can be revised as, step 1: $u_{start} = y_j$, step 2: $p' = u'X/u'u$ ($w' = u'X/u'u$), step 3: $p'_{new} = p'_{old} / \|p'_{old}\|$ ($w'_{new} = w'_{old} / \|w'_{old}\|$), step 4: $t = X_p / p'p$ ($t = Xw/w'w$), step 5: $q' = t'Y/t't$, step 6: $q'_{new} = q'_{old} / \|q'_{old}\|$, step 7: $u = Yq/q'q$, step 8: compare the t in step 4 with the one in the preceding iteration step. Assuming block Y has only one variable, steps 5 – 8 can be omitted by equating $q = 1$. The algorithm achieves convergence quickly in order to rotate the components for X and Y blocks (Geladi & Kowalski, 1986).

3. Harmonic analysis

The height of tide at time t can be computed by expressing it as the sum of harmonic terms,

$$h(t) = s_0 + \sum_{i=1}^{\infty} H_i \cos(w_i t + \alpha_i). \quad (5)$$

Where $h(t)$ denotes the height of tide at time t , s_0 is the height of mean water level above the datum used, H_i is the amplitude of the constituents, w is the constituents' speed, t is time, and α is the phase lag of the constituents. H_i and α_i are known as the harmonic constants (Doodson & Warburg, 1941; Dronkers, 1972).

Since the phase lag of the constituent α_i depends on the varying mean longitude of the moon's perigee and sun's perigee with periods of approximately 8.61 and 2100 years respectively, as well as the ecliptic longitude of the moon's ascending node with a period of 18.61 tropical years, f_i , v_i and u_i are introduced into equation 5 to account for these effects as,

$$h(t) = s_0 + \sum_{i=1}^n f_i H_i \cos(w_i t + (v_i + u_i) - \alpha_i) \quad (6)$$

Where n is the number of harmonic constituents, v_i is the phase angle at time zero, f_i is the nodal factor, and u_i is the nodal angle (Cheng & Gartner, 1985). By introducing the trigonometric identity, equation 6 can be expanded as,

$$h(t) = s_0 + \sum_{i=1}^n f_i H_i \cos(w_i t + (v_i + u_i)) \cos \alpha_i + \sum_{i=1}^n f_i H_i \sin(w_i t + (v_i + u_i)) \sin \alpha_i \quad (7)$$

Let,

$$a_i = H_i \cos \alpha_i, \quad (8)$$

and

$$b_i = H_i \sin \alpha_i. \quad (9)$$

a_i and b_i are called the constituents' constants. Dividing equation 9 by equation 8 yielded,

$$\alpha_i = \tan^{-1}\left(\frac{b_i}{a_i}\right), \quad (10)$$

By introducing equations 8 and 9 into equation 7 yielded,

$$h(t) = s_0 + \sum_{i=1}^n [a_i f_i \cos(w_i t + (v_i + u_i)) + b_i f_i \sin(w_i t + (v_i + u_i))] \quad (11)$$

The design matrix can be created from equation 11 in the following form,

$$X_{n \times m} = \begin{bmatrix} 1 & f_1 \cos(w_1 t_1 + (v_1 + u_1)) & f_1 \sin(w_1 t_1 + (v_1 + u_1)) & f_2 \cos(w_2 t_1 + (v_2 + u_2)) & f_2 \sin(w_2 t_1 + (v_2 + u_2)) \dots \\ 1 & f_1 \cos(w_1 t_2 + (v_1 + u_1)) & f_1 \sin(w_1 t_2 + (v_1 + u_1)) & f_2 \cos(w_2 t_2 + (v_2 + u_2)) & f_2 \sin(w_2 t_2 + (v_2 + u_2)) \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & f_1 \cos(w_1 t_n + (v_1 + u_1)) & f_1 \sin(w_1 t_n + (v_1 + u_1)) & f_2 \cos(w_2 t_n + (v_2 + u_2)) & f_2 \sin(w_2 t_n + (v_2 + u_2)) \dots \end{bmatrix} \quad (12)$$

And the vector Y can be obtained as,

$$Y_{n \times 1} = [h(t_1), h(t_2), h(t_3), \dots, h(t_n)]^T \quad (13)$$

While the unknown parameters are,

$$\hat{\beta} = [s_0, a_1, b_1, a_2, b_2, \dots, a_m, b_m]^T \quad (14)$$

To obtain the weight P , each observation is considered as having been made independently with the same amount of reliability. Hence the error in observations ($\mathcal{G}_{\hat{Y}}$) can be assumed to be equal to the resolution of the tide gauge used, therefore,

$$\sum_{\hat{Y}} = \text{diag} \left[\mathcal{G}_{\hat{Y}_1}^2, \mathcal{G}_{\hat{Y}_2}^2, \mathcal{G}_{\hat{Y}_3}^2, \dots, \mathcal{G}_{\hat{Y}_n}^2 \right]. \quad (15)$$

n and m denote the number of observations and unknown parameters respectively. Therefore the corresponding weight can be expressed as,

$$P = \sum_{\hat{Y}}^{-1} = \text{diag} \left[\frac{1}{\mathcal{G}_{\hat{Y}_1}^2}, \frac{1}{\mathcal{G}_{\hat{Y}_2}^2}, \frac{1}{\mathcal{G}_{\hat{Y}_3}^2}, \dots, \frac{1}{\mathcal{G}_{\hat{Y}_n}^2} \right] \quad (16)$$

4. Modelling

The data used for the modelling were based on tide gauge observations at Lagos, Nigeria located at latitude: 6.423300°N and longitude: 3.411700°E (WOCE Sea Level, 2006). The data were obtained from consecutive hourly tidal observations carried out for a one year period from 1st January 00:00 hrs to 31st December 23:00 hrs. One thousand consecutive hourly tidal observations that form the composite data were extracted for the modelling (Figure 1). The one thousand tidal observations were split into two (Figure 2).

Figure 1: Composite dataset

Figure 2: Training and test datasets

The first 900 datasets (900 hours) were used for training while the second 100 datasets (100 hours) were used for testing (or prediction) based on the hold-out method of cross-validation (Okwuashi & Ndehedehe, 2015; Lee *et al.*, 2018). Twenty tidal constituents M2, S2, N2, K1, O1, MO3, MK3, MN4, M4, SN4, MS4, 2MN6, M6, 2MS6, S4, SK3, 2MK5, 2SM6, 3MK7, and M8 were selected having met the criterion for inclusion for the tidal prediction based on the findings by Foreman (1996) on tidal constituents and the length (in hours) required for their inclusion for tidal prediction expressed in Table 1. The selected twenty tidal constituents required not more 900 hours for their inclusion. The constituents asterisked in the table required beyond 900 hours for their inclusion.

Table 1: Tidal constituents and length (in hours) required for their inclusion for prediction. *Required more than 900 hours for constituent inclusion [adapted from Foreman (1996)]

S/N	Shallow tidal constituent	Record length (in hours) required for constituent inclusion	Main tidal constituent	Record length (in hours) required for constituent inclusion
1	SO1*	4383	S2	355
2	MKS2*	4383	M2	13
3	MSN2*	4383	O1	328
4	MO3	656	K1	24
5	SO3*	4383	N2	662
6	MK3	656	K2*	4383
7	SK3	355		
8	MN4	662		
9	M4	25		
10	SN4	764		
11	MS4	355		
12	MK4*	4383		
13	S4	355		
14	SK4*	4383		
15	2MK5	24		
16	2SK5	178		
17	2MN6	662		
18	M6	26		
19	2MS6	355		
20	2MK6*	4383		
21	2SM6	355		
22	MSK6*	4383		
23	3MK7*	24		
24	M8	26		

The values of the constituents' speed (°/hr), constituents' astronomical arguments (degree), and constituents' nodal factor are expressed in table 2. The accuracy evaluation of the PLSR and LS models was based on the Root Mean Square Error (RMSE) and the Coefficient of Determination known as r^2 . The RMSE is computed as,

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2} . \quad (17)$$

Where y_j is the observed tide, $\hat{y}_j$ is the predicted tide, while n represents the number of observations (DiStefano *et al.*, 2018). Then r^2 was used to evaluate the correlation or how close the predicted tide fits the observed tide; r^2 is computed as,

$$r^2 = 1 - \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{\sum_{j=1}^n (y_j - \bar{y}_j)^2} , \quad (18)$$

where y_j is the observed tide, $\hat{y}_j$ is the predicted tide, $\bar{y}_j$ is the mean of observed tide, and n denotes the number of observations (Hu & Miao, 2018).

Table 2: Constituents' speed, constituents' astronomical arguments, and constituents' nodal factor

S/N	Tidal constituent	Speed (°/hr)	V+U (degree)	f
1	M2	28.984104	1.2854605886	1.0063923852
2	S2	30.000000	360.0000000000	1.0000000000
3	N2	28.439730	281.6303114864	1.0063923852
4	K1	15.041069	228.0175372109	0.9953671368
5	O1	13.943036	93.4467282115	0.9920402910
6	MO3	42.9271400	94.7321888001	0.9983817946
7	MK3	44.0251729	229.3029977995	1.0017299069
8	MN4	57.4238337	282.9157720751	1.0128256329
9	M4	57.9682084	2.5709211773	1.0128256329
10	SN4	58.4397300	281.6303114864	1.0063923852
11	MS4	58.9841042	1.2854605886	1.0063923852
12	2MN6	88.4079380	284.2012326637	1.0193000044
13	M6	86.9523127	3.8563817659	1.0193000044
14	2MS6	87.9682080	2.5709211773	1.0128256329
15	S4	60.0000000	359.9430262177	1.000000798
16	SK3	45.0410686	280.6504715533	1.000022106
17	2MK5	74.0251728	335.8143655541	0.9880002314
18	2SM6	88.9841042	182.3890143401	1.0005902126
19	3MK7	101.9933600	197.0421939936	0.9876986591
20	M8	115.9364169	214.3360589725	1.0509418653

Figure 3: Computed variances for 1st to 40th components

Using the PLSR, the predictor variables X were regressed on the response variable Y . From Figure 3, $x_{40 \times 40}$ PLSR components were obtained, since there are 40 variables. Each tidal constituent yielded two variables. Since there were twenty tidal constituents, we had a total of forty variables which yielded forty PLSR components. Each column contained the PLSR component of one variable. It can be discerned that the variances decreased from the 1st to the 3rd component and it remained the same from the 3rd to the 40th component. Each component represents each regressor, it implied that the regressors with higher variances were better than the regressors with lower variances. From Figure 4, the higher the r^2 value the more accurate the number of components used for the modelling. The RMSE results were the opposite of the r^2 results.

Figure 4: RMSEs and r^2 results for 1st to 40th PLSR components

The computed RMSE decreased from the 1st component to the 3rd component and remained same from the 3rd component till the 40th component, while the computed r^2 increased from the 1st component to the 3rd component and remained same from the 3rd component till the 40th component. It can be discerned that the computed RMSE and r^2 did not change after the 3rd component. Hence needless using the 4th component to the 40th component for the prediction. Therefore only the 1st component, 2nd component, and the 3rd component were used for the prediction. The predicted constituent constants, phase lags, and amplitudes for the 1st component, 2nd component, and the 3rd component were presented in Tables 3 and 4.

Table 3: Computed constituent constants for 1st to 3rd PLSR components

Constituents		PLSR components		
		1st	2 nd	3rd
	s_0	1.8076	1.8076	1.8077
M2	a_1, b_1	-0.2883, -0.2251	-0.3119, -0.2409	-0.3120, -0.2404
S2	a_2, b_2	0.1234, 0.0752	0.1105, 0.1288	0.1106, 0.1293
N2	a_3, b_3	0.1058, 0.0965	0.1028, 0.0288	0.1050, 0.0270
K1	a_4, b_4	0.0492, -0.0560	0.0578, -0.0637	0.0586, -0.0633
O1	a_5, b_5	-0.0047, 0.0211	-0.0070, 0.0213	-0.0057, 0.0213
MO3	a_6, b_6	-0.0022, 0.0011	-0.0010, 0.0008	-0.0030, 0.0009
MK3	a_7, b_7	-0.0035, -0.0048	-0.0007, -0.0048	-0.0036, -0.0048
MN4	a_8, b_8	0.0097, -0.0021	0.0097, -0.0036	0.0089, -0.0028
M4	a_9, b_9	-0.0178, -0.0096	-0.0200, -0.0084	-0.0205, -0.0098
SN4	a_{10}, b_{10}	0.0019, 0.0008	0.0020, 0.0031	0.0007, 0.0032
MS4	a_{11}, b_{11}	0.0086, 0.0083	0.0093, 0.0120	0.0104, 0.0114
2MN6	a_{12}, b_{12}	0.0004, -0.0010	0.0011, -0.0002	0.0009, -0.0004
M6	a_{13}, b_{13}	0.0021, 0.0035	0.0029, 0.0034	0.0022, 0.0031
2MS6	a_{14}, b_{14}	-0.0008, -0.0032	-0.0013, -0.0030	-0.0016, -0.0030
S4	a_{15}, b_{15}	0.0003, -0.0024	0.0009, -0.0023	0.0012, -0.0024
SK3	a_{16}, b_{16}	-0.0017, -0.0001	-0.0014, -0.0030	-0.0012, -0.0033
2MK5	a_{17}, b_{17}	0.0010, 0.0004	0.0005, 0.0005	0.0010, 0.0005
2SM6	a_{18}, b_{18}	0.0004, -0.0007	0.0006, -0.0004	0.0002, -0.0004
3MK7	a_{19}, b_{19}	-0.0011, 0.0006	-0.0005, 0.0004	-0.0005, 0.0005
M8	a_{20}, b_{20}	0.0002, -0.0012	0.0010, -0.0018	0.0007, -0.0013

Table 4: PLSR computed phase lag and amplitude for the 1st to the 3rd component

Constituents	Phase lag (in degrees)			Amplitude (m)		
	1st	2nd	3rd	1st	2nd	3rd
M2	217.9821	217.6812	217.6147	0.3658	0.3999	0.3999
S2	31.3581	49.3731	49.4571	0.1699	0.2392	0.2773
N2	42.3679	15.6505	14.4208	0.1432	0.1068	0.1084
K1	311.3016	312.2199	312.7920	0.0745	0.0860	0.0863
O1	102.5576	108.1925	104.9816	0.0216	0.0224	0.0220
MO3	153.4349	141.3402	163.3008	0.0025	0.0013	0.0031
MK3	233.9017	261.7029	233.1301	0.0059	0.0049	0.0060
MN4	347.7843	339.6384	342.5361	0.0099	0.0103	0.0093
M4	208.3391	202.7824	205.5501	0.0202	0.0217	0.0227
SN4	22.8337	57.1715	77.6609	0.0021	0.0037	0.0033
MS4	43.9830	52.2243	47.6264	0.0120	0.0152	0.0154
2MN6	291.8014	349.6952	336.0375	0.0011	0.0011	0.0010
M6	59.0362	49.5378	54.6375	0.0041	0.0045	0.0038
2MS6	255.9638	246.5713	241.9275	0.0033	0.0033	0.0034
S4	277.1250	291.3706	296.5651	0.0024	0.0025	0.0027
SK3	183.3665	244.9831	250.0169	0.0017	0.0033	0.0035
2MK5	21.8014	45.0000	26.5651	0.0011	0.0007	0.0011
2SM6	299.7449	326.3099	296.5651	0.0008	0.0007	0.0004
3MK7	151.3895	141.3402	135.0000	0.0013	0.0006	0.0007
M8	279.4623	299.0546	298.3008	0.0012	0.0021	0.0015

The classification of the predicted tide can be determined by computing the tidal form factor F as,

$$F = \frac{K_1 + O_1}{M_2 + S_2}. \quad (19)$$

The tide is then said to be, semidiurnal if $0 \leq F \leq 0.25$, mixed if $0.25 < F \leq 3.00$, and diurnal if $F > 3.00$ (Dietrich, 1963; Foreman, 1996; Pugh, 2004). From Table 4, the computed amplitudes for

PLSR for the 1st component were 0.0745, 0.0216, 0.3658 and 0.1445 for K_1 , O_1 , M_2 , S_2 and respectively. Therefore F was computed as, $F = \frac{0.0745 + 0.0216}{0.3658 + 0.1699} = 0.1794$. The computed F fell between 0 and 0.25. The computed amplitudes for PLSR for the 2nd component were 0.0860, 0.0224, 0.3941 and 0.1697 for K_1 , O_1 , M_2 , S_2 and respectively. Therefore F was computed as, $F = \frac{0.0860 + 0.0224}{0.3999 + 0.2392} = 0.1696$. The computed F fell between 0 and 0.25. The computed amplitudes for PLSR for the 3rd component were 0.0863, 0.0220, 0.3939 and 0.1701 for K_1 , O_1 , M_2 , S_2 and respectively. Therefore F was computed as, $F = \frac{0.0863 + 0.0220}{0.3999 + 0.2773} = 0.1599$. The computed F fell between 0 and 0.25.

Figure 5: Computed Yloadings for the 1st to the 36th component

Figure 6: Computed Yscores for 1st to 40th principal component

From Figure 5, the computed Yloadings showed that the 1st, 2nd, and 3rd components were sufficient for the prediction, since the computed Yloadings from the 3rd to the 40th component were virtually the same. Also from Figure 6 aside the 1st, 2nd and 3rd components, the computed Yscores from the 4th to the 40th component were practically the same. This implied that the 1st, 2nd and 3rd components were reliably adequate for the prediction.

Figure 7 depicted the graphs of the observed tide and PLSR predicted tide, while Figure 8 depicted the graph of the correlation between the observed tide and PLSR predicted tide. The computed r^2 for the PLSR 1st, 2nd and 3rd components were 0.9134, 0.9825, 0.9933 respectively. The next experiment was based on the LS. LS minimises,

$$\sum_{i=1}^n Y_i - f(X_i, \beta). \quad (20)$$

Where β is the vector of the unknown parameters, X denotes the design matrix of the predictor variables, Y represents the vector of the response variable, and n expresses the number of observations. The solution of the vector of adjusted unknown parameters $\hat{\beta}$ can be written as,

$$\hat{\beta} = (X^T P X)^{-1} X^T P Y. \quad (21)$$

Where P represents the matrix of weight of observations. The LS results for the computed constituents' constants, amplitude, and phase lag were presented in Table 5. From Table 5, the computed amplitudes for LS were 0.0829, 0.0211, 0.3997 and 0.1632 for K_1 , O_1 , M_2 and S_2 respectively. Using equation 19 F was

computed as $F = \frac{0.0829 + 0.0211}{0.3939 + 0.1632} = 0.1848$. The computed F fell between 0 and 0.25. The graph of the LS predicted tide was shown in Figure 7, while Figure 8 depicted the graph of the correlation between the observed tide and LS predicted tide. The computed r^2 for the LS was 0.7861.

Figure 7: Graph of observed tide and PLSR and LS predicted tide

Figure 8: Correlation between the observed tide and, PLSR and LS predicted tide

Table 5: LS results for computed constituent constants (*significant at $\pm t > 1.96$ or $p < 0.05$ at the 95% confidence level).

Tidal constituent	Constituent constant	Coefficient	Phase lag (degrees)	Amplitude	t stat	p value
	s_0	1.8075			824.1118	0.0000*
M2	a_1, b_1	-0.3162, -0.2445	217.7173	0.3997	-99.4909, -76.9601	0.0000*, 0.0000*
S2	a_2, b_2	0.1060, 0.1241	49.4878	0.1632	33.9202, 39.6799	0.0000*, 0.0000*
N2	a_3, b_3	0.0979, 0.0260	14.8573	0.1013	31.0512, 8.2375	0.0000*, 0.0000*
K1	a_4, b_4	0.0558, -0.0613	312.2982	0.0829	17.8227, -19.6424	0.0000*, 0.0000*
O1	a_5, b_5	-0.0071, 0.0198	109.6277	0.0211	-2.2599, 6.3125	0.0241*, 0.0000*
MO3	a_6, b_6	-0.0008, 0.0003	157.2877	0.0009	-0.2577, 0.1077	0.7967, 0.9143
MK3	a_7, b_7	-0.0004, -0.0048	265.6816	0.0048	-0.1148, -1.5214	0.9086, 0.1285
MN4	a_8, b_8	0.0086, -0.0030	340.5881	0.0092	2.7433, -0.9668	0.0062*, 0.3339
M4	a_9, b_9	-0.0204, -0.0093	204.5326	0.0224	-6.4199, -2.9311	0.0000*, 0.0035*
SN4	a_{10}, b_{10}	0.0011, 0.0034	72.1850	0.0036	0.3446, 1.0719	0.7305, 0.2841
MS4	a_{11}, b_{11}	0.0100, 0.0117	49.2231	0.0154	3.1464, 3.6478	0.0017*, 0.0003*
2MN6	a_{12}, b_{12}	0.0010, -0.0003	340.5488	0.0010	0.3083, -0.1089	0.7579, 0.9133
M6	a_{13}, b_{13}	0.0027, 0.0032	49.8145	0.0042	0.8705, 1.0313	0.3843, 0.3027
2MS6	a_{14}, b_{14}	-0.0011, -0.0030	249.3347	0.0032	-0.3679, -0.9756	0.7131, 0.3296
S4	a_{15}, b_{15}	0.0011, -0.0024	294.6156	0.0026	0.3506, -0.7650	0.7260, 0.4445
SK3	a_{16}, b_{16}	-0.0011, -0.0032	250.7125	0.0034	-0.3555, -1.0156	0.7223, 0.3101
2MK5	a_{17}, b_{17}	0.0004, 0.0005	49.2667	0.0006	0.1341, 0.1557	0.8933, 0.8763
2SM6	a_{18}, b_{18}	0.0003, -0.0004	305.5548	0.0005	0.0974, -0.1363	0.9224, 0.8916
3MK7	a_{19}, b_{19}	-0.0004, 0.0003	142.1550	0.0006	-0.1421, 0.1104	0.8871, 0.9121
M8	a_{20}, b_{20}	0.0008, -0.0017	294.5319	0.0019	0.2626, -0.5751	0.7929, 0.5654

The graphs of the computed residuals as well as the histogram plots of the residuals for PLSR and LS were presented in Figures 9 and 10. The residuals were computed by subtracting the predicted tide from the observed tide. The histogram plot was used to visualise the normality of the computed residuals. The Jarque-Bera test statistic was used to test the normality of the computed residuals. The Jarque-Bera test statistic is computed as,

$$JB = \frac{n}{6} \left[s^2 + \frac{(k-3)^2}{4} \right]. \quad (22)$$

Where JB denotes Jarque-Bera statistic, n is the sample size, s is the sample skewness, and k is the sample kurtosis (Jarque & Bera, 1980; Abdellatif *et al.*, 2018). The hypotheses for the normality test can be stated as H_0 : the residuals are normally distributed at the 95% confidence level, H_1 : the residuals are not normally distributed at the 95% confidence level (reject H_0 if p-value < 0.05). The Jarque-Bera test results for PCR and LS were presented in Table 6. We accept H_0 (that the residuals of the PLSR and LS were normally distributed) since the p values of the two models were greater than 0.05. From the bar plot in Figure 11, the computed RMSEs for PLSR 1st, 2nd, 3rd components and LS were 0.1731, 0.0834, 0.0655, and 0.2488 respectively. Also from Figure 11 the box-plot showed that, for the PLSR 1st component, the computed median, minimum, and maximum residuals were 0.0624, -0.2244, and 0.3855 respectively; the PLSR 2nd component, the computed median, minimum, and maximum residuals were 0.0533, -0.0576, 0.1862 respectively; and the PLSR 3rd component, the computed median, minimum, and maximum residuals were 0.0596, -0.0299, 0.1602 respectively. For the LS the computed median, minimum, and maximum residuals were 0.0404, -0.2367, and 0.3608 respectively.

Figure 9: Computed residuals for PLSR and LS

Figure 10: Histogram plot of the PLSR and LS residuals

Table 6: The Jarque-Bera test for PLSR and LS residuals (*significant at $p < 0.05$)

Model	JB statistic	Critical value	P value
PLSR (1 st comp.)	4.8482	5.4314	0.0611
PLSR (2 nd comp.)	4.6710	5.4314	0.0656
PLSR (3 rd comp.)	3.8608	5.4314	0.0916
LS	5.3241	5.4314	0.0517

Figure 11: Box plot of the computed residuals and bar plot of the computed RMSE for the PLSR and LS

5. Discussion and conclusion

The 1st, 2nd, and 3rd components yielded the highest variances therefore they had the greatest impact in the model (Figure 3). The higher the variance the more significant the regressor (Ndehedehe *et al.*, 2018a; Ndehedehe *et al.*, 2018b; Ndehedehe *et al.*, 2018c). It can be discerned that the computed variances from the 4th to the 40th components were the same. This implied that the 1st, 2nd, and 3rd components were sufficient for the prediction, hence the 4th to the 40th were not used for the prediction. The computed r^2 and RMSE in Figure 4 corroborated the result in Figure 3, because changes only occurred in the 1st, 2nd, and 3rd components that implied that they were the most potent components. The computed tidal form factor F for the 1st, 2nd, and 3rd PLSR components as well the LS fell between $0 \leq F \leq 0.25$ indicating a semi-diurnal tide. This implied that the results of the experiment can be trusted since the coastal water of the observed tide is naturally semi-diurnal. From Figure 5, the 1st, 2nd, and 3rd components yielded the highest Yloadings. This is another indication that the 1st, 2nd, and 3rd components had the greatest influence in the experiment. From Figure 6, but for the 1st and 2nd components, the 3rd to the 40th component of the computed Yscores were virtually the same. This implied that the experiment was mostly influenced by the 1st, 2nd, and 3rd components, hence the selection of best PLSR components was truncated at the 3rd component.

Another technique for validating the models is by testing the normality of the computed residuals using the Jarque-Bera statistical normality test. The result of the normality test indicated that all the models' residuals yielded $p > 0.05$ which showed that the residuals were normally distributed at the 95% confidence level. From Figure 11 the box plot showed that the 3rd component had the smallest box, which implied that it was the most accurate model. The computed RMSE and r^2 indicated that the 3rd component was the most accurate while the LS was the least accurate. Generally the computed RMSE and r^2 proved that the PLSR performed far better than the conventional LS model. The PLSR has the advantage of reducing the number of regressors by reducing the number of components, then discarding the redundant components and eventually selecting the most potent components. This helps in mitigating the effect of multicollinearity (Ndehedehe *et al.*, 2016a; Ndehedehe *et al.*, 2016b; Krepper *et al.*, 2018). The result of this experiment showed that the PLSR is a veritable alternative model to the conventional LS model for tide level prediction.

References

- Abdellatif, D., El Moutaouakil, K., & Satori, K. (2018). clustering and jarque-bera normality test to face recognition. *Procedia Computer Science*, 127, 246-255.
- Cai, S., Liu, L., & Wang, G. (2018). Short-term tidal level prediction using normal time-frequency transform. *Ocean Engineering*, 156, 489-499.

- Chang, H. K., & Lin, L. C. (2006). Multi-point tidal prediction using artificial neural network with tide generating forces. *Coastal Engineering*, 53(10), 857-864.
- Cheng, J. H., & Sun, D. W. (2017). Partial least squares regression (PLSR) applied to NIR and HSI spectral data modeling to predict chemical properties of fish muscle. *Food Engineering Reviews*, 9(1), 36-49.
- Chen, H., Sun, Y., Gao, J., Hu, Y., & Yin, B. (2018). Solving partial least squares regression via manifold optimization approaches. *IEEE Transactions on Neural Networks and Learning Systems*, (99), 1-13.
- Cheng, R. T., & Gartner, J. W. (1985). Harmonic analysis of tides and tidal currents in South San Francisco Bay, California. *Estuarine, Coastal and Shelf Science*, 21(1), 57-74.
- de Jong, C. D., Lachapelle, G., Skone, S., & Elema, I. A. (2010). *Hydrography*. Delft: Delft University Press.
- Dietrich, G. (1963). *General oceanography*. New York, NY: Interscience Publishers.
- DiStefano, C., Liu, J., Jiang, N., & Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness?. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3), 453-466.
- Doodson, A. T., & Warburg, H. D. (1941). *Admiralty manual of tides*. Admiralty, UK: Hydrographic Department.
- Dronkers, J. J. (1972). *Tidal theory and computations*. The Hague: Hydraulic Department of Delta Works.
- Foreman, M. G. G. (1996). Manual for tidal heights analysis and prediction. Pacific Marine Sciences Report. *Institute for Ocean Sciences*.
- Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185, 1-17.
- Guan, M., Li, Q., Zhu, J., Wang, C., Zhou, L., Huang, C., & Ding, K. (2019). A method of establishing an instantaneous water level model for tide correction. *Ocean Engineering*, 171, 324-331.
- Haenlein, M., & Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. *Understanding Statistics*, 3(4), 283-297.
- Hong, B., Liu, Z., Shen, J., Wu, H., Gong, W., Xu, H., & Wang, D. (2020). Potential physical impacts of sea-level rise on the Pearl River Estuary, China. *Journal of Marine Systems*, 201, 103245.
- Hu, W., & Miao, H. (2018). Sub-pixel displacement algorithm in temporal sequence digital image correlation based on correlation coefficient weighted fitting. *Optics and Lasers in Engineering*, 110, 410-414.
- Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econometric Letters*, 6, 255-259.
- Krepper, G., Romeo, F., de Sousa Fernandes, D. D., Diniz, P. H. G. D., de Araújo, M. C. U., Di Nezio, M. S., Pistonesi, M. F., & Centurión, M. E. (2018). Determination of fat content in chicken hamburgers using NIR spectroscopy and the Successive Projections Algorithm for interval selection in PLS regression (iSPA-PLS). *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 189, 300-306.
- Lee, T. (2004). Back-propagation neural network for long-term tidal predictions. *Ocean Engineering*, 31 (2), 225-238.
- Lee, T. L., & Jeng, D. S. (2002). Application of artificial neural networks in tide-forecasting. *Ocean Engineering*, 29(9), 1003-1022.
- Lee, L. C., Liong, C. Y., & Jemain, A. A. (2018). Validity of the best practice in splitting data for hold-out validation strategy as performed on the ink strokes in the context of forensic science. *Microchemical Journal*, 139, 125-133.
- Liu, M., Xie, R., Wu, S., Zhu, R., Mao, Z., & Wang, C. (2018). Permeability prediction from mercury injection capillary pressure curves by partial least squares regression method in tight sandstone reservoirs. *Journal of Petroleum Science and Engineering*, 169, 135-145.
- Ndehedehe, C. E., Agutu, N., & Okwuashi, O. (2018b). Is terrestrial water storage a useful indicator in assessing the impacts of climate variability on crop yield in semi-arid ecosystems? *Ecological Indicators*, 88 (2018) 51-62.
- Ndehedehe, C. E., Agutu, N. O., Okwuashi, O., & Ferreira, V. G. (2016b). Spatio-temporal variability of droughts and terrestrial water storage over Lake Chad Basin using independent component analysis. *Journal of Hydrology*, 540, 106-128.
- Ndehedehe, C. E., Awange, J. L., Agutu, N., & Okwuashi, O. (2018a). Changes in hydro-meteorological conditions over tropical West Africa (1980-2015) and links to global climate. *Global and Planetary Change* 162 (2018) 321-341.
- Ndehedehe, C. E., Awange, J. L., Corner, R. J., Kuhn, M., & Okwuashi, O. (2016a). On the potentials of multiple climate variables in assessing the spatio-temporal characteristics of hydrological droughts over the Volta Basin. *Science of the Total Environment*, 557-558, 819-837.
- Ndehedehe, C. E., Okwuashi, O., Ferreira, V. G., & Agutu, N. O. (2018c). Exploring evapotranspiration dynamics over Sub-Sahara Africa (2000-2014). *Environmental Monitoring And Assessment*, 190(7), 400.

- Okenwa, E. G. (1978). *Automated tidal reduction of soundings*. Technical report of the Department of Surveying Engineering, University of New Brunswick, Canada.
- Okwuashi, O., & Ndehedehe, C. (2015). Digital terrain model height estimation using support vector machine regression. *South African Journal of Science*, 111(9/10), 1-5.
- Okwuashi, O., & Ndehedehe, C. (2017). Tide modelling using support vector machine regression. *Journal of Spatial Science*, 62(1), 29-46.
- Okwuashi, O., & Olayinka, D. N. (2017). Tide modelling using the Kalman filter. *Journal of Spatial Science*, 62(2), 353-365.
- Pugh, D. T. (2004). *Changing sea levels: effects of tides, weather and climate*. Cambridge, UK: Cambridge University Press.
- Shu, J. (2003). Prediction and analysis of tides and tidal currents. *International Hydrographic Review*, 4 (2), 24–29.
- Straus, J., & Skogestad, S. (2019). A new termination criterion for sampling for surrogate model generation using partial least squares regression. *Computers & Chemical Engineering*, 121, 75-85.
- WOCE Sea Level, WSL (2006). *Hourly sea level from coastal tide gauge station Lagos-C in 1993 (Research quality database)*. Sea Level Center, University of Hawaii, doi:<http://dx.doi.org/10.1594/PANGAEA.435683>.

AUTHORS' RESPONSE TO REVIEWERS COMMENTS

- (1) **Reviewers' comment:** The authors fail to provide an objective criteria for deciding which tidal constituents to include and exclude in their analysis.

Authors' response: The basis for the selection of the tidal constituents has now been included. Based on the analysis by Forman (1996), the appropriate tidal constituents have now been included in the tidal prediction.

- (2) **Reviewers' comment:** Page 4: Line 50 – Page 5: Line 28: The lists of constituent speeds, nodal factors, and astronomical arguments would be more readable if it was provided in a table. Equation (11) I believe there is a typographical error in the coefficient of sine; I believe it should be b_i instead of a_i . Please ensure this error does not propagate through the remainder of the derivation.

Authors' response: a_i has been corrected to b_i in Equation 11.