Journal of Applied Structural Equation Modeling

eISSN: 2590-4221

Journal of Applied Structural Equation Modeling: 5(1),1-8, Jan 2021
DOI: 10.47263/JASEM.5(1)01

RELATIVE IMPORTANCE OF WATER QUALITY PARAMETERS ON FISH GROWTH USING PLS REGRESSION

Job Ombiro Omweno[a*], Albert Getabu[a], Paul Sagwe Orina[b]
Simion Kipkemboi Omasaki[a], Wilfred Obwoge Zablon[a]

[a]*Department of Aquatic and Fishery Sciences, Kisii University, Kisii, Kenya.*
[b]*Kenya Marine & Fisheries Research Institute (KMFRI), Kegati Aquaculture Research Centre, Kisii, Kenya.*

[*]*omwenojob@gmail.com*

ABSTRACT

Partial least squares (PLS) is a multivariate dimension reduction technique which is not based on ordinary regression assumptions. The use of PLS regression in life sciences is still a novel concept despite having many scientific applications. This paper analyses the relative importance of physicochemical parameters on the growth of *Oreochromis jipe* using *Oreochromis niloticus* as a control. Modelling and the graphical display of the regression coefficients were performed using a suite of open access R-software packages. The modelling hypotheses were assessed using experimental data collected from 270 fingerlings cultured for the period of 84 days. The findings revealed that significant linear correlation exists between water quality variables and the mean body weight of both *O. jipe* and *O. niloticus* fish species. The study provides baseline information to assess the growth of *O. jipe* under aquaculture conditions; therefore, we recommend a further study to be conducted on several other predictor variables that can be measured under controlled aquaculture conditions.

Keywords: *Water quality; PLS regression;* multi-co-linearity *; Oreochromis jipe ;* Mean prediction error

INTRODUCTION

Multivariate analysis methods such as the Partial Least Squares Regression (PLSR) have a wide range of scientific applications such as spectroscopy to compare the effect of several variables. This method has been highly developed to handle multi-co-linearity and a large number of explanatory variables which do not exhibit a direct relationship with the response variable (Frank & Friedman, 1993). Although there may be many manifest variables, only few latent variables account for the underlying relationships. Graphical displays may show heteroscedascities

although diagnostic tests may indicate that residuals are indeed homogenous, leading to incorrect conclusions (Gelman & Unwin, 2013). This challenge has been partly solved through the common practice of log - transforming data prior to regression analysis which increases the tendency for normal and homoscedastic distribution of residuals to satisfy some of the regression assumptions (Alan et al., 2020). In fisheries, log transformation is purposefully performed to eliminate curvilinear relationships when determining the fish condition (Spare & Venema, 1998). However, linearization only increases the percentage variance in the response variable explained by the independent variable (Pinheiro et al., 2020), which yields a highly significant F-ratio value and coefficient of determination (R^2) which are frequently misinterpreted to mean a well-fitting regression model. Nevertheless, it is possible that none of the variables can be truly independent as both variables covary in response to unknown extraneous variables. Fish growth is as a result of interactions among several variables in the culture environment which are likely not independent of each other. For instance, temperature influences many physicochemical variables such as DO and salinity directly and pH indirectly (Nehemia et al., 2012). Although the PLS method may not be useful for singling out variables that have a negligible effect on the response variable, it has the advantage of incorporating several controllable variables which affect dozens of outputs or responses hence has found industrial and scientific applications.

LITERATURE REVIEW

The correlation between water quality variables and fish growth parameters is controlled by several extraneous variables within a complex aquaculture system. Several multivariate techniques have been developed to decompose a whole-complex system into separate components which can be studied simultaneously (Ward, 2009). This is advantageous to provide crucial information for management and the general understanding of fisheries by predicting the adaptability of a given species to the culture environment (Juan-Jordá et al., 2015; Okomoda et al., 2018). The use of projection based inverse regression methods such as PCR and PLSR has recently gained popularity because these methods allow post-evaluation of the regression results using bias and mean prediction errors, which is not possible with multiple linear regression (MLR) because they lack elaborate external validation procedures. Also, MLR models are highly sensitive to multi-co-linearity between independent variables; hence generate unstable prediction coefficients resulting from unreliable prediction errors. Nevertheless, these models can be modified or up-scaled to allow generalizations of future possibilities due to overdependence on conventional data which may be lacking at times (Ayata et al., 2013; Colléter et al., 2013). Several renowned authors (e.g. Abdel-Raheam et al., 2017; De Graaf et al., 2005; Conceicäo, 1997) have studied fish growth and condition using models. This is especially for commercially cultured species such as *O. niloticus* which contribute to food and nutrition security among the vulnerable rural populations in the sub-saharan region. The relative importance of physico-chemical parameters can be assessed by correlation using linear regression models (Jopp et al., 2011). Thus, the robust projection PLS regression used in this study will be more useful in handling outliers attributed to shooters than ordinary MLR which generate more efficient estimators using small datasets (Rodionova & Pomerantsev, 2020).

MATERIALS AND METHODS

Data was collected during the experimental growth studies of *O. jipe* and *O. niloticus* conducted at KMFRI, located in Kisii County, Kenya. Total length (TL) and body weight measurements from 270 stocked fingerlings were taken biweekly using the methods described by Omweno et al. (2020). Water quality variables: temperature, pH, dissolved oxygen (DO), salinity, conductivity and total dissolved solids were recorded using a YSI multi-parameter meter (H9829 model). The

number of culture days (age) was counted from the stocking day. All data was organized into two-dimensional data frames containing parameters and their corresponding variables which were used for building the models. Prior assessment of the variance inflating Factors (VIF) showed no co-linearity among the independent variables although there were heteroscedascities observed in graphical displays. The initial step involved selection of the variable components, (X_1 to X_{10}), which were regressed against the response variable(y). The number of components selected was pegged on the assumption that each additional component significantly increased the variance explained by the model. Models were generated by submitting the datasets to modeling arguments coded in the 64-bit R software 3.6.3 (R-Core Team, 2020). The software provides a platform for modeling using different packages such as semPlot (Epskamp, 2019) and lavaan (Rosseel, 2012). These packages translate syntax of input commands specified by functions into high quality graphics which indicate different variables and relationships using nodes and edges. The model was run as an assignment to the object created in R workspace. The single term representing the matrix of regressors on the right hand side was separated from the response variables by a tilde (~) operator whereas the plus (+) operator was used to show the additive effect of regressors. The the I( ) operator is used to protect the additive regressors to nullify their interpretation as separate regressors (Horton and Kleinman, 2015). The layout and layoutSplit arguments were used to control the type and the placement of the nodes. The Model information was extracted using summary ( ) and coef( ) functions. The significance of parameter estimates in the model and the overall model performance were assessed using the model p-values and loadings. Observations with a large leverage (distance between the individual observation and its projection on the variable component retained in the model) were regarded as potential outliers. In this study however, influential outliers were not eliminated from the model because they originated from shooters which exhibited faster growth than the rest of mixed-sex fingerlings and were less than 5% of the total observations. Model validation was used to optimize the number of latent variables (LVs) used to build a well-fitting model. The mean prediction errors such as root mean square error in external prediction (RMSEP) were used to estimate how well the model predicts external datasets. The calculated prediction error was used as a measure of prediction accuracy, and a resampling of 4000 iterations was used to generate partial regression estimates for the hypothesized relationships through bootstrapping for inference purposes. Cross validation (CV) was performed to determine the optimal number of dimensions to be used in building the model using package pls (Mevik et al., 2020). The CV was performed using the Leave-Out One (LOO) approach which selects variable components repeatedly until all components have been left out and computes CV and Adj. CV estimates. Validation used both training and test data sets. However, there was no external validation due to lack of sufficient data.

RESULTS

The regression output indicates five variables which were used in predicting the fish mean weight. The raw associations among the variables determining growth before any modeling was performed are presented in the correlation plot shown in figure 1. Water temperature showed negative correlation with DO and the culture period in both species whereas pH showed negative correlation with DO in *O. jipe*.

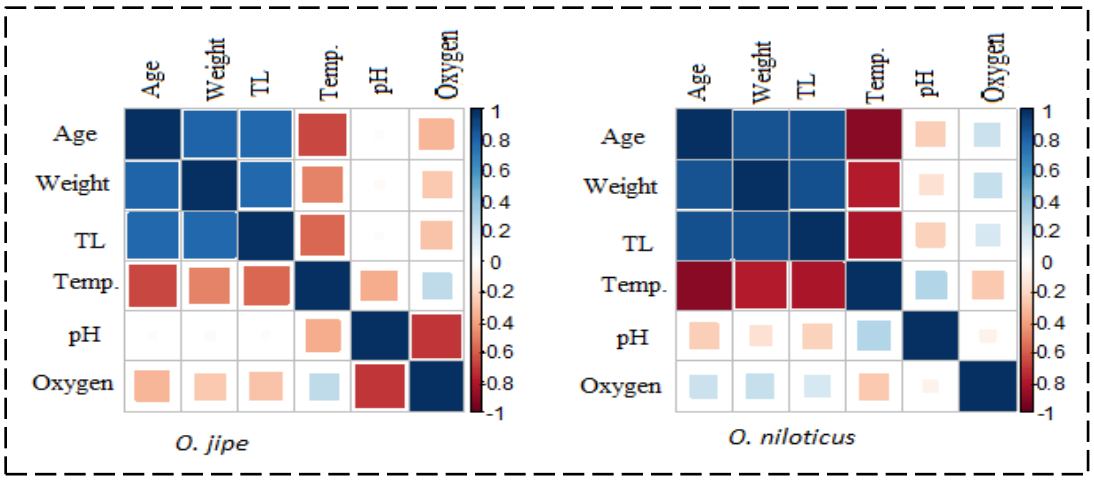


Figure 1: Correlation plots showing the raw associations between water quality and fish growth parameters in an aquaculture system

Table 1 and Table 2 show the regression coefficients, Z - values, standard error and significance of the model parameter estimates. Results indicated that in both *O. jipe* and *O. niloticus*, dissolved oxygen and pH had the greatest influence on mean weight consequently producing the most important relationships which exhibited the highest variance explained by the models (highest R^2 value). In *O. jipe* pls model, dissolved oxygen had a strong negative relationship with fish total length, but the relationship was not significant (β = -0.138, $p < 0.000$). Similarly, in *O. niloticus*, dissolved oxygen was negatively related to pH (β = -0.064) and total length (β = -0.201) but the relationships were also not significant ($p > 0.05$).

Table 1: Coefficients of partial least squares regression dimensions in the *Oreochromis jipe* model

Regression	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
Weight ~ TL	0.925	0.073	12.731	0.000	1.26	0.16
Weight ~ Age	0.103	0.007	15.484	0.000	2.96	0.37
Weight ~ Temp.	0.865	0.125	6.916	0.000	0.87	0.11
Weight ~ DO	2.487	0.519	4.794	0.000	2.71	0.34
Weight ~ pH	2.463	0.473	5.205	0.000	1.26	0.16

Table 2: Coefficients of partial least squares regression dimensions in the *Oreochromis niloticus* model

Regression	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
Weight ~ TL	2.167	0.143	15.119	0.000	0.76	0.14
Weight ~ Age	0.181	0.019	9.531	0.000	0.52	0.09
Weight ~ Temp.	0.689	0.414	1.663	0.002	2.17	0.41
Weight ~ DO	2.521	0.645	3.909	0.000	1.76	0.33
Weight ~ pH	2.040	0.644	3.169	0.096	0.76	0.14

The study also normalized the output of partial regression coefficients so their absolute sum is 100 and sorted the result before it was displayed. The results below indicate that all the assessed independent variables sorted on both scales (head and tail) were positive predictors of fish bodyweight.

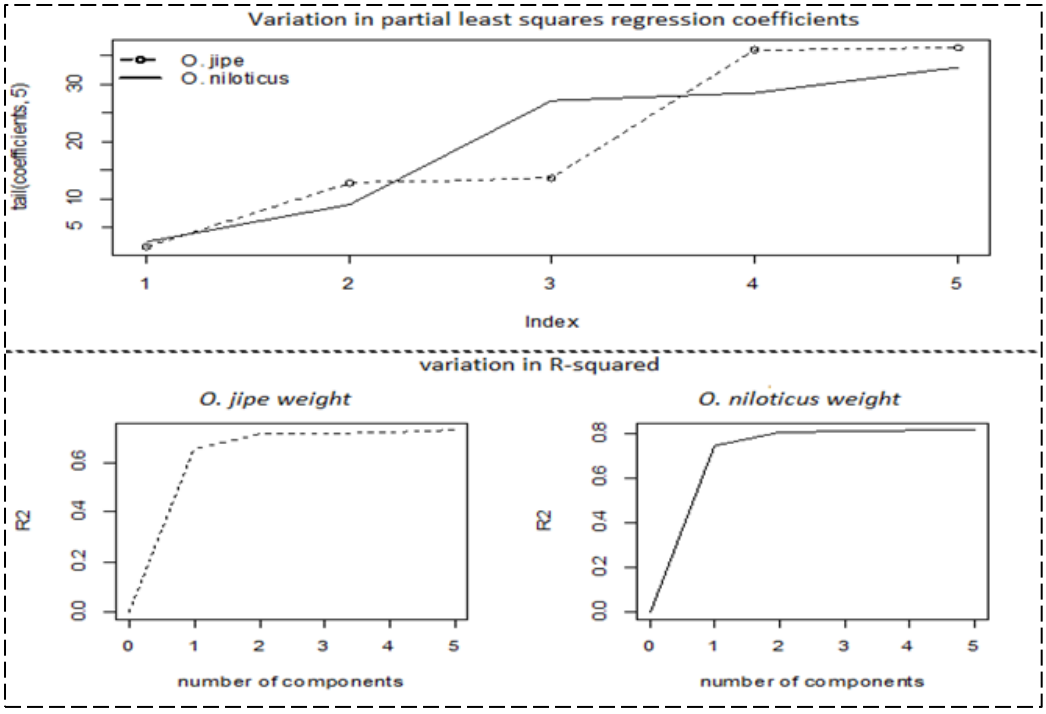
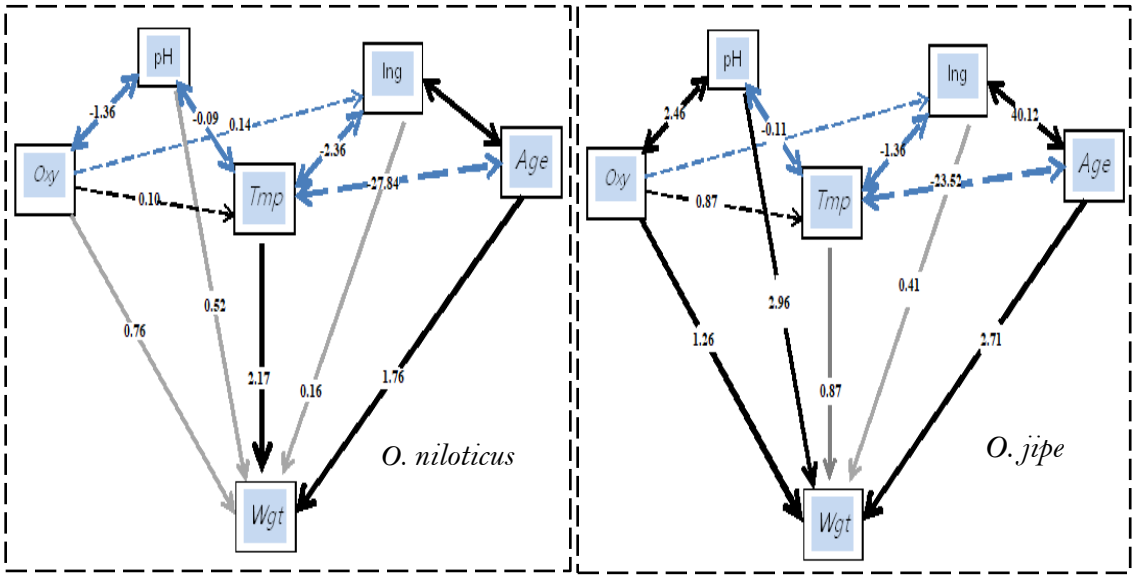


Figure 2: Variations of R-squared and partial regression coefficients in *O. jipe* and *O. niloticus* species (1 – Culture period, 2 – Temperature, 3 – DO, 4 – pH and 5 – TL).

The specific influence of each variable component on fish growth is directional in nature; therefore, the study quantifies the relative influence of each independent variable on the mean response variable. Figure 3 indicates negative, positive, negative, strong and weak relationships between water quality variables and fish body weight in *O. jipe* and *O. niloticus* distinguished by the colour and width of the arrows.



Noted: Thick bold arrows indicate strong positive correlations while thin-faded and broken arrows weak and negative correlations between modeled variables.

Figure 3: Relationships between water quality variables and fish growth parameters in aquaculture.

The overall PLS model R^2 showed that all the independent variables explained 71.2% and 84.5% percent of the response variable in *O. jipe* and *O. niloticus* respectively as shown in table 3. Validation datasets had low mean prediction errors and high R^2 compared to prediction datasets in both MLR and PLS regressions. However, the MLR yielded higher mean prediction errors compared to the PLS regression.

Table 3: Comparison of the model performance statistics between calibration and validation datasets in *O. jipe* and *O. niloticus*

Treatment	Performance statistics	Calibration		Validation	
		MLR	PLS	MLR	PLS
Oreochromis jipe	RMSEP	1.2239	0.4492	0.2365	0.1286
	SEP	1.0936	0.3487	0.2136	0.1010
	Bias	0.6931	0.1000	0.4968	-0.1020
	Adj. R^2	0.6413	0.7120	0.6495	0.6691
Oreochromis niloticus	RMSEP	0.5834	0.3984	0.4103	0.3021
	SEP	0.5245	0.4607	0.3973	0.2794
	Bias	0.3156	0.0000	0.1020	-0.8089
	Adj. R^2	0.8180	0.8450	o.9171	0.9379

DISCUSSION

Regression models show correlation between variables and the parameter estimate coefficients (Fox, 2016). Different statistical software use matrix algebra to conveniently compute parameter estimate coefficients (Faraway, 2005; Rawlings et al., 1998). The most common parameter for estimating model accuracy is the coefficient of multiple determination (R^2). However, a large number of variables having few observations is likely to generate a multiple regression model that fits the observed data set perfectly (has a large R^2 value) but will fail to give accurate prediction of different data because of over-fitting. Overfitting occurs where there are many manifest variables in the model but only a few underlying (or latent) variables explain most of the variation in the response. Thus, the PLS techniques works to extract these latent variables accounting for the variance in the response variable. All the five predictors together explained 71.2% and 84.5% percent of the variance in the mean body weight of *O. jipe* and *O. niloticus* respectively. A larger R^2 indicates that the observed data in a population will most likely assuming the model with similar coefficient values. This is supported by low RMSEP and SEP values which showed that the models yielded stable partial coefficients indicating high accuracy in making predictions outside the datasets used for model calibration. However, this could be a misleading generalization suggesting that the outdoor wooden ponds provide a more stable environment for the culture of *O. jipe* species whose growth was investigated. This could be due to unobserved fluctuations in physico-chemical parameters because the data modeled was manually collected over regular time intervals, hence did not cover multiple fluctuations within the aquaculture system. In other related studies, the water quality data is usually collected using automatic sensor detectors over infinitesimal time intervals which capture a wide range of multiple fluctuations (Platikanov, 2016). The CV and the bias-corrected Adj. CV were used to train the estimates and highly significant models were selected from the nested models with a different the number of components for comparison. The final predictive models illustrated in figure 3 were linear combination of five independent variable combinations was based on conceptualized relationships. The directional arrows indicate that variables exhibited both

negative and positive correlations with the response variable. Specifically, the culture period with β estimates of 0.103 and 0.181 in *O. jipe* and *O. niloticus* had very strong positive correlation with fish total weight. This is confirmed by the standardized coefficients in table 1 and table 2. Each slope coefficient (beta) represents the average change in the response variable attributed to unit change in the predictor variable when all other predictor variables held constant. Moreover, they provide a measure of the average contribution of the predictor variables towards the variance in the response variable. Besides, PLS is particularly suited to project predicted and observed variables to a new dimensional space unlike the Principal Component Analysis (PCA), which ignores the response variable. The technique maps predictors into a small set of components and regresses them against the response variable but does not use hyperplanes to maximize the variance explained by the independent variables. The P-value < 0.000 shows highly significant partial coefficients. For instance, the influence of pH on the growth of *O. niloticus* was rather insignificant ($p > 0.05$) compared to other variables. This may be because the pH was within the optimum range of 6.5 - 8.5 recommended for Tilapia culture (Nehemiah et al., 2012). Apart from model p-values, the relative influence of each predictor estimate on the response variable was assessed using the Z-statistics. Many methods can be used to assess the relative influence of physico-chemical parameters on fish growth but the PLS regression enables the researcher to select variable components that maximally explain the response variable. It is important to note that that the first three LVs (Culture period, pH and dissolved oxygen) account for 90% of the total variance in the response variable and form the most important factors affecting the aquaculture potential of *O. jipe*. This is confirmed by cross-validation in which PLS with five factors achieved the mean squared error which is not significantly different than the model with the three factors.

LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

The study only considered five variables out of many variables that can measure water quality. This limits our basis for generalizations made in this study. Additionally, the modeling data was collected over a short time period which can also limit our temporal generalizations. Besides, most authors suggest that partial regression should be performed on highly controllable variables, an assumption which was not obeyed by this study. Despite these shortcomings, the present study provides a basis for examining the relationship between water quality and fish growth performance in terms of body weight. In order to address the limitations above, this study recommends a future study should involve several predictor variables that can be measured under controlled aquaculture conditions.

CONCLUSION

PLS regressions provide an explicit way of assessing the linear relationship between many variables that it is not possible with Multiple linear regression due to the challenge of meeting assumptions. The method can sufficiently handle multi-co-linearity between independent variables by reducing the multivariate relationship into workable dimensions. Although the method has been widely applied in scientific areas such as computer science, cell biology and chemistry, it is still new in fisheries but can be used to analyze short term experimental data to linearly extract few but relatively important latent variables where there are a large number of manifest variables.

REFERENCES

Abdel-Raheam, H & El-Bassir, A. (2017). Length-Weight Relationship and Condition Factor of Three Commercial Fish Species of River Nile, Sudan". *EC Oceanography* 1(1): 01-07.

Akaike, H. (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, AC- 19, 716-723. http://dx.doi.org/10.1109/TAC.1974.1100705

Alan, G., Bretz, F., Miwa, T., Mi, X. & Hothorn, T. (2020). mvtnorm: Multivariate Normal and t Distributions (version 1.1-1). https://CRAN.R-project.org/package=mvtnorm.

Ayata, S. D., Lévy, M., Aumont, O., Sainte-Marie, J., Tagliabue, A., & Bernard, O. (2013). Archimer Phytoplankton growth formulation in marine ecosystem models : Should we take into account photo-acclimation and variable stoichiometry in. *Journal of Marine Systems*, 125(September), 29–40. https://doi.org/10.1016/j.jmarsys.2012.12.010

Burnham, K. P. & Anderson, D.R. (2002) *Model Selection and Inference: A Practical Information-Theoretic Approach*. 2nd Edition, Springer-Verlag, New York.

Conceicäo, L. E. C. (1997). Growth in early life stages of fishes: an explanatory model. *PhD thesis, Wageningen Agricultural University*, Wageningen, The Netherlands.

Colléter, M., Guitton, J., & Gascuel, D. (2013). An introduction to the EcoTroph R Package: analyzing aquatic ecosystem trophic networks. *R Journal*, 5(1), 98–107. Retrieved from http://journal.r-project.org/archive/2013-1/colleter-guitton-gascuel.pdf

De Graaf, G. J., Dekker, P. J., Huisman, B. & Verreth, J. A. J. (2005). Simulation of *O. niloticus* culture in pond, through individual–based modeling, using a population dynamics approach. *Aquaculture Research*, 36: 455-472.

Epskamp, S. (2019). semPlot: Path Diagrams and Visual Analysis of Various SEM Packages' Output. R package version 1.1.2.https://CRAN.R-project.org/package=semPlot

Faraway, J. J. (2005). Linear Models with R, Chapman & Hall/CRC: *Texts in Statistical Science*. NewYork, USA. ISBN 0-203-59454-1

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models*. Third Edition. Thousand Oaks CA:Sage.

Frank, I. & Friedman, J. (1993), "A statistical view of some chemometrics regression tools," *Technometrics*, 35, 109-135.

Gelman, A., & Unwin, A. (2013). Infovis and Statistical graphics: different goals, different looks. Journal of Computational and Graphical Statistics, 22(1), 2–28.

Horton, N. J. & Kleinman, K. (2015). *Using R and rstudio for data management, statistical analysis, and graphics*, second edition, CRC Press, Boca Raton, 313 Pp. ISBN 978-1-4822-3736-8. DO - 10.1201/b18151. Retrieved from http://www.amherst.edu/~nhorton/r^2/

Ighwela, K. A., Aziz Bin Ahmed, A. B. & Abol-Munafi, A. B. (2011). Condition Factor as an Indicator of Growth and Feeding Intensity of Nile Tilapia Fingerlings (Oreochromis niloticus) Feed on Different Levels of Maltose. *American-Eurasian J. Agric. & Environ. Sci.*, 11 (4): 559-563, ISSN 1818-6769

Jopp, F., Reuter, H., & Breckling, B. (2011). Modelling complex ecological dynamics: An introduction into ecological modelling for students, teachers & scientists. Modelling Complex Ecological Dynamics: An Introduction into Ecological Modelling for Students, *Teachers & Scientists*. https://doi.org/10.1007/978-3-642-05029-9

Juan-Jorda, M. J., Mosqueira, I., Freire, J. & Dulvy, N. K. (2015). Population declines of tuna and relatives depend on their speed of life. *Proc. R. Soc. B* 282: 20150322.

Mevik, B., Wehrens, R. & Liland, K. H. (2020). pls:Partial Least Squares and Principal Component Regression. R package version 2.7-3. https://CRAN.R-project.org/package=pls

Nehemia, A., Maganira, J. D. & Rumisha, C. (2012). Length-Weight relationship and condition factor of Tilapia species grown in marine and fresh water ponds. *Agriculture Biol. J. N. Am.3* (3):117-124.

Okomoda, V. T., Koh I. C. C., Hassan, A., Amornsakun T. & Shahreza, S. M. (2018). Length-weight relationship and condition factor of the progenies of pure and reciprocal crosses

of Pangasianodon hypophthalmus and Clarias gariepinus. *AACL Bioflux* 11(4):980-987.http://www.bioflux.com.ro/aacl

Omweno, J. O., Orina, P. S., Getabu, A. & Outa, N. O. (2020). Growth and aquaculture potential of Tilapia jipe, Oreochromis jipe and Nile tilapia, Oreochromis niloticus. International *Journal of Fisheries and Aquatic Studies*, 2020; 8(3): 395-399. http://www.fisheriesjournal.com

Palmer, P. B. & O'Connell, D. G. (2009). Regression Analysis for Prediction: Understanding the Process. *Cardiopulmonary Physical Therapy Journal*, 20(3): 23-26.

Pinheiro, J., Bates D., DebRoy. S., Sarkar, D. & R Core Team (2020). _nlme: Linear and Nonlinear Mixed Effects Models_. *R package version* 3.1-144. https://CRAN.R-project.org/package=nlme>.

Platikanov, Y. S. (2016). Application of chemometric methods to water quality studies. *PhD Thesis*, University of Barcelona, Barcelona, Spain. 286 Pp.

Rawlings, J. O., Pantula, S. G. & Dickey, D. A. (1998). *Applied Regression Analysis: A research tool, Second Edition. Springer-Verlag*, New York, Inc. ISBN 0-387-98454-2.https://www.springer.com ›Applied Regression Analysis - A Research Tool | John O ... – Springer

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rodionova, O. Y. & Pomerantsev, A. L. (2020). Detection of outliers in projection - based modeling. *Anal. Chem.* (92): 2656−2664

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Softwar*e, 48(2): 1-36. http://www.jstatsoft.org/v48/i02/.

Sparre, P. & Venema, S. C. (1998). Introduction to tropical fish stock assessment. Part 1. Manual. FAO Fisheries *Technical Paper*, (306.1, Rev. 2). 407 Pp.

Ward, B. A. (2009). Marine Ecosystem Model Analysis Using Data Assimilation by October 2009

Zar, J. H. (1984), "*Biostatistical Analysis*", (2nd ed.). Englewood Cliffs, New Jersey: Prentice Hall.