JEE Journal of Ecological Engineering
Volume 19, Issue 3, May 2018, pages 141–145
https://doi.org/10.12911/22998993/86462

Received: 2018.02.01
Accepted: 2018.03.15
Published: 2018.05.01

# Predictive Modelling for Characterisation of Organics in Pit Latrine Sludge from Unplanned Settlements in Cities of Malawi

Khumbo Kalulu[1*], Bernard Thole[1], Theresa Mkandawire[1], Grant Kululanga[1]

[1] University of Malawi, The Polytechnic, P/Bag 303, Chichiri, Blantyre 3, Malawi

* Corresponding author's e-mail: kkalulu@poly.ac.mw

**ABSTRACT**

The limited availability of data on faecal sludge characteristics remains one of the major challenges faced by developing countries in proper management of faecal sludge. In view of the limited financial resources and expertise in these developing countries, there is a need to come up with less-resource-intensive approaches for faecal sludge characterisation. Despite being used substantially in wastewater, there is limited evidence on the use of predictive modelling as a tool for cost-effective characterisation of faecal sludge. In this study, first order multiple linear regression modelling is investigated as a less-resource-intensive approach for accurate prediction of organics (biochemical oxygen demand and chemical oxygen demand) in pit latrine sludge. The predictor variables explored in the modelling include pH, electrical conductivity, total solids, total volatile solids, fixed solids and moisture content. The modelling uses data collected from 80 latrines in unplanned settlements of four cities in Malawi. The study shows that it is possible to reliably predict chemical oxygen demand and biochemical oxygen demand in pit latrine sludge using electrical conductivity and total solids, which require low levels of resources and expertise to determine.

**Keywords:** Akaike Information Criterion, biochemical oxygen demand, chemical oxygen demand, faecal sludge characteristics, multiple linear regression model

## INTRODUCTION

The limited availability of data on faecal sludge characteristics remains one of the major challenges faced by developing countries. This is attributed to lack of financial resources and expertise, among other factors (Strande et al., 2014). The existing body of knowledge presents high spatial and temporal variability necessitating generation of context-specific data (Bassan et al., 2013). Generation of such data using traditional lab-based approach calls for high levels of resourcing. Resource constraint in developing countries, thus, presents a need to generate less-resource-intensive approaches for characterisation of sludge (Strande et al., 2014). One of such approaches is the application of predictive models in characterisation of faecal sludge. Predictive modelling provides a cost-effective way of generating accurate information (Aguado, et al., 2006). Despite substantial use in wastewater, there is lack of literature pointing towards use of predictive modelling as a tool for cost-effective characterisation of faecal sludge (Brdjanovic et al., 2007; Singh et al., 2010; Khataee and Kasiri, 2011; Nasr et al., 2012). This study, therefore, explored the applicability of multiple linear regression modelling as a less-resource-intensive method for accurate prediction of organics (biochemical oxygen demand and chemical oxygen demand) in pit latrine sludge from four cities in Malawi.

## DATA AND METHODOLOGY

The data used for modelling was collected from 80 pit latrines in unplanned settlements of four cities in Malawi (Blantyre, Lilongwe, Mzuzu and Zomba). Predictive models for organics (biochemical oxygen demand and chemical oxygen

demand) in pit latrine sludge were generated using first order multiple linear regression modelling. The generic form for first order regression model with n predictor variables is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (1)$$

The procedure for model building and selection is shown in Figure 1. The model building and selection were carried out in Minitab 17 and Microsoft Excel at a significance level of 0.05. The first stage in the model building and selection process was data preparation during which the data was checked for missing values and outliers. In addition, a visual inspection of the probability plot was performed to check for normality of the untransformed values of the dependent variable. In the cases where normality was not satisfied, the dependent variable was subjected to transformations following the order in the Tukey Ladder of Powers until normality was attained (Barker and Shaw, 2015).

The second stage was the identification of predictor variables and their combinations for building competing models. The latrine sludge parameters requiring low skill and resourcing levels were selected to be predictor variables. These included pH, electrical conductivity (EC), moisture content (MC), total solids (TS), total volatile solids (TVS) and fixed solids (FS). Determination of pH and EC in the pit latrine sludge was done using potentiometric methods. The potentiometric methods require basic skills of dipping meter probes and direct reading of values (for both buffer solution during calibration and sample solution) from the meter. The sludge moisture content and solids were determined using gravimetric methods, the core skills of which include weighing, setting right temperature of the furnace and direct measurement reading from the weighing scale. In order to reduce the probability of multicollinearity in competing models built from these predictor variables, Pearson correlation was used to identify highly correlated predictor variables. Highly correlated predictor variables were those with $|r| \geq 0.7$ (Vatcheva et al., 2016).

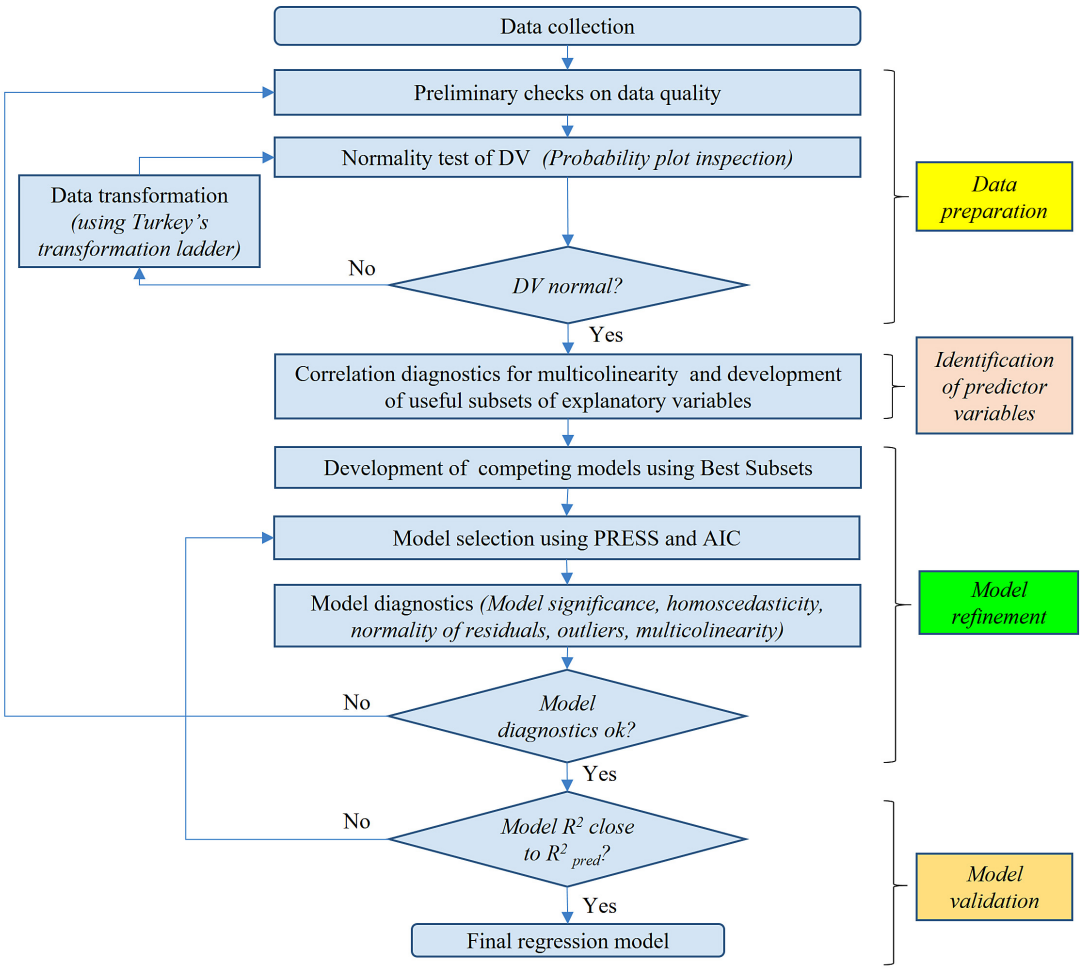


**Figure 1**. Model building and selection flow chart

Subsets of predictor variables were formulated in such a way to ensure that no subset contained highly correlated predictor variables. The Best Subset function in Minitab 17 was used to generate candidate models from the different subsets of the predictor variables. The best model was selected from the list of the candidate models, using the prediction sum of squares (PRESS) statistic. The model with the lowest PRESS was selected. In the instances where competing models had the same PRESS value, Akaike Information Criterion (AIC) values were calculated and the model with the smallest AIC value was selected. AIC statistic was chosen because it aims at achieving parsimony, which fits in well with resource maximisation that is desirable in resource-constrained settings (Bozdogan, 2000). The selected model was then investigated for model significance, homoscedasticity, randomness of residuals, outliers, amount of data for precise estimation of the strength of the regression relationship and multicollinearity of predictor variables. A model was deemed to be significant when its *p*-value was less than a significance level of 0.05. Homoscedasticity, normality of residuals and outliers were checked through the visual inspection of the Residuals vs Fitted Values plots. Specifically, the randomness of points on both sides of zero and large residuals that could have a strong influence on the model were checked. Large residuals and unusual values were identified and investigated back to the untransformed data for their unusual nature. The observation-to-predictor ratio was used to check the sufficiency of data for precise estimation of the strength of the regression relationship. In literature, the minimum observation-to-predictor ratio ranges from 10 to 30 (Pedhazur and Schmelkin, 2013). Variance inflation factor (VIF) was used to check for multicollinearity of the predictor variables fitted in the model. VIF values in the range $0 < VIF < 5$ suggest that there is no multicollinearity problem. VIF values of $5 \leq VIF \leq 10$ show moderate multicollinearity while $VIF \geq 10$ is indicative of significant multicollinearity (Moustris et al., 2012).

Model validation was performed using the predicted r-squared value ($R^2_{pred}$), which measures how well a model predicts responses for new observations. Validation of predictions was conducted by comparing the model $R^2$ and $R^2_{pred}$ values. A model was judged to provide valid predictions if values of $R^2_{pred}$ and $R^2$ were close to each other (Frost, 2013).

## RESULTS AND DISCUSSION

### Predictive model for biochemical oxygen demand (BOD)

The study found that linear regression modelling can be used to reliably predict biochemical oxygen demand (BOD) in latrine sludge from unplanned settlement across the four cities. The predictive model arrived at was:

$$log_{10}\, BOD = 2.4588 - 0.977\, log_{10}\, TS \quad (2)$$

where: *BOD* is biochemical oxygen demand (mg/g TS),
*TS* is total solids (%).

The BOD model statistics are shown in Table 1. The relationship between the model variables is significant ($p<.0001$) and explains about 91% ($R^2$) of the variability that existed in the data. Since $R^2 > 75\%$, the variability explained is substantial enough to have confidence in the model (Hair et al., 2013). There is no effect of multicollinearity, since the VIF (1) for the model is less than 5. The observation-to-predictor ratio for the model (240) is greater than the minimum of 10 to 30. No observable trend was found in the Residuals vs Fitted values plot, implying homoscedasticity and randomness of residuals. The model provides valid predictions as $R^2_{pred}$ (90.9%) is close to $R^2$ (91.0%). This prediction model presents a way of cutting down on the time required to analyse faecal sludge for BOD. It takes at least 5 days to obtain BOD results from the method used in this study while total solids' determination takes less than 24 hours. However, it should be noted that this level of reliability of the model holds for a BOD range 3.65 to 1139.7 mg/g TS within which the model was developed.

### Predictive model for chemical oxygen demand (COD)

Linear regression modelling produced a model that allows a reliable prediction of chemical ox-

**Table 1**. BOD model statistics

| | |
|---|---|
| Model significance | <0.0001 |
| $R^2$ | 91.0% |
| SSE | 6.1717 |
| $R^2_{pred}$ | 90.9% |
| PRESS | 6.3 |
| VIF (Log TS) | 1.0 |
| Observation-to-predictor ratio | 240 |

ygen demand in pit latrine sludge from unplanned settlements across the four cities. The predictive model arrived at was:

$$log_{10}\, COD = 3.668 - \\ - 0.8882\, log_{10}\, TS - 0.0852\, log_{10}\, EC \quad (3)$$

where: $COD$ is chemical oxygen demanding (mg/g TS),
$TS$ is total solids (%),
$EC$ is electrical conductivity (µs/cm).

The COD model statistics are shown in Table 2. The COD model is significant ($p<0.0001$) and explains a substantial part of the variability in the data with $R^2$ (91.8%) > 75% (Hair et al., 2013). No multicollinearity of the fitted predictor variables exists in the model, as both variables have VIF of 1.03, which is less than 5. The observation-to-predictor ratio (160) for the model is greater than the minimum range of 10 to 30.

**Table 2**. COD model statistics

| | |
|---|---|
| Model significance | <0.0001 |
| $R^2$ | 91.8% |
| SSE | 5.3 |
| $R^2_{pred}$ | 91.6% |
| PRESS | 5.5 |
| VIF (Log TS) | 1.03 |
| VIF (Log EC) | 1.03 |
| Observation-to-predictor ratio | 160 |

Valid predictions can be made from the model, since $R^2_{pred}$ (91.6%) and $R^2$ (91.8%) close to each other. The Residuals vs Fitted values plot did not display any observable trend implying homoscedasticity and randomness of residuals. Though COD takes a shorter duration (2 hours) to obtain results, the model still provides a cost-effective way of generating data on COD of pit latrine sludge, since the reagents and expertise required to conduct a COD test outweigh the requirements of gravimetric methods. Just like the BOD model, the level of reliability for this model holds within the COD range 33.8 to 9604.4 mg/g TS.

## CONCLUSIONS

The study has demonstrated that it is possible to reliably predict BOD and COD in pit latrine sludge using electrical conductivity and total solids, which require low levels of resources and expertise to determine. This predictive characterisation seems to be applicable across different spatial settings/localities. Since the models were developed using data from latrines from only four sites, there is a need to evaluate the performance of these models with sludge from other urban areas of Malawi for generalizability at a national level.

## Acknowledgements

Special thanks to Water Research Commission SA for funding the study through the Sanitation Research Fund for Africa (SRFA) Project. The Polytechnic, University of Malawi is acknowledged for providing lab space and equipment as well as administrative support during the research study. Our gratitude extends to the technician team from the Department of Physics and Biochemical Sciences at The Polytechnic, University of Malawi. Consortium for Advanced Research Training in Africa (CARTA) is also recognised for capacity building of the corresponding author in research and publication.

This research was also supported by the Consortium for Advanced Research Training in Africa (CARTA). CARTA is jointly led by the African Population and Health Research Center and the University of the Witwatersrand and funded by the Wellcome Trust (UK) (Grant No: 087547/Z/08/Z), the Carnegie Corporation of New York (Grant No. B 8606.R02), Sida (Grant No. 54100029). The statements made and views expressed are solely the responsibility of the fellow.

## REFERENCES

1. Aguado D., Ferrer A., Seco A., Ferrer J. 2006. Comparison of different predictive models for nutrient estimation in a sequencing batch reactor for wastewater treatment. Chemometrics and Intelligent Laboratory Systems, 84(1), 75–81.
2. Barker L.E. and Shaw K.M. 2015. Best (but oft-forgotten) practices: checking assumptions concerning regression residuals. The American Journal of Clinical Nutrition, 102(3), 533–539.
3. Bassan M., Tchonda T., Yiougo L., Zoellig H., Mahamane I., Mbéguéré M., Strande L. 2014. Characterization of faecal sludge during dry and rainy seasons in Ouagadougou, Burkina Faso. Proc. 36th WEDC International Conference, 1–5.
4. Bozdogan H. 2000. Akaike's information criterion and recent developments in information complexity. Journal of Mathematical Psychology, 44(1), 62–91.

5. Brdjanovic D., Mithaiwala M., Moussa M.S., Amy G., Van Loosdrecht M.C.M. 2007. Use of modelling for optimization and upgrade of a tropical wastewater treatment plant in a developing country. Water Science and Technology, 56(7), 21–31.
6. Frost J. 2013. Multiple regression analysis: Use adjusted R-squared and predicted R-squared to include the correct number of variables. Minitab Blog, 13(6).
7. Hair J.F., Black W.C., Babin B.J., Anderson R.E., Tatham R.L. 2013. Multivariate data analysis). Upper Saddle River, NJ: Prentice Hall.
8. Khataee A.R. and Kasiri M.B. 2011. Modeling of biological water and wastewater treatment processes using artificial neural networks. CLEAN–Soil, Air, Water, 39(8), 742–749.
9. Moustris K.P., Nastos P.T., Larissi I.K. Paliatsos A.G, 2012. Application of multiple linear regression models and artificial neural networks on the surface ozone forecast in the greater Athens area, Greece. Advances in Meteorology, 2012, 1–8.
10. Nasr M.S., Moustafa M.A., Seif H.A., El Kobrosy G. 2012. Application of Artificial Neural Network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT. Alexandria Engineering Journal, 51(1), 37–43.
11. Pedhazur E.J. and Schmelkin L.P. 2013. Measurement, design, and analysis: An integrated approach. Psychology Press.
12. Singh K.P., Basant N., Malik A., Jain G. 2010. Modeling the performance of "up-flow anaerobic sludge blanket" reactor based wastewater treatment plant using linear and nonlinear approaches–a case study. Analytica Chimica Acta, 658(1), 1–11.
13. Strande L., Ronteltap M., Brdjanovic D. 2014. Faecal Sludge Management: Systems Approach for Implementation and Operation. IWA Publishing.
14. Vatcheva K.P., Lee M., McCormick J.B., Rahbar M.H. 2016. Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology (Sunnyvale, Calif.), 6(2).