E-ISSN: 1116 - 3240 P-ISSN: 3026-9334
MPRI-JAIMR | INTERNATIONAL JOURNAL OF AFRICAN INNOVATION & MULTIDSCIPLINARY RES.
VOL. 06 NO. 2 – OCTOBER, 2024

APPLICATION OF SUPPORT VECTOR MACHINE AND ORDINARY KRIGING (SVM-OK) TO SPATIAL INTERPOLATION OF POROSITY ATTRIBUTE OF AGBABU BITUMEN DEPOSIT

[1]OKWUDIRE CHUKWUDI EMMANUEL; [2]UDEOZO FRANCIS CHINONSO; & [3]ONOVO AUGUSTINE CHINEMEREM

[1]Department of Petroleum Engineering, College of Engineering, Covenant University Otta. [2&3] Department of Mechanical/Production Engineering, Faculty of Engineering, Enugu State University of Science and Technology.

DOI Unique ID: 10.70382/mejaimr.v6i2.003

Abstract

The lack of geostatistical modelling of volumetric characteristics of Nigeria's long-discovered heavy oil and bitumen deposits in Agbabu was accountable for the discrepancies in the hydrocarbon-in-place calculations. The absence of spatial variability evaluation in such efforts is responsible for the inconsistency and divergence that define the impacts of the existing attempts to estimate the porosity attribute. Due to its optimum features and simplicity of estimate, Kriging is reported as the most generally used approach to evaluate the connection between variables (inputs and output). However, Kriging reveals that the relationship between variables requires a linear

Keywords:
Variogram, SVM, Kriging, Bitumen Field, Geological Studies.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

form, which is frequently not the case. Thus, the necessity for a procedure with no fixed shape is built based on information acquired directly from the data. Consequently, support vector regression, a form of the support vector machine, is employed as a choice in contrast to kriging. This research is a comparative application of support vector machine and kriging (SVM-OK) to the spatial interpolation of the porosity property of the Agbabu bitumen field. The porosity values from existing data are trained, cross-validated, and verified with the help of the support vector machine, and they are interpolated by ordinary kriging. The geographical patterns of the predictions have been checked visually using graphs based on these three diagnostics: variance of the error terms and correlation of the error terms. The findings of this study illustrate the performance of the combination of the Support vector machine and conventional kriging in the spatial interpolation of volumetric characteristics in the Agbabu bitumen deposit. The approach exhibited great predictive performance. SVM-kriging will be employed in reservoir characterisation with the least mean absolute error (MAE) and mean squared error (MSE).

Introduction

Before the discovery of oil and gas, which comprise over 80% of the country's income, Nigeria's economy was far more diversified. Due to ongoing exploitation and overdependence as a significant source of income, the hydrocarbon potential of the abundantly oil-rich Niger Delta will be reduced or exhausted shortly.

Geostatistical analysis conducted on the bitumen deposit by Falebita (2014) reveals that the bitumen is primarily found in the shallow, unconsolidated sands of the formation, which are intercalated with shale and clay layers. The deposits are heavy oil and bitumen with high viscosity and density. The depositional environment of the bitumen-bearing sands was likely a shallow marine environment with a high rate of organic matter accumulation. The organic matter was then transformed into bitumen through thermal maturation, which occurs due to the burial and heating of the organic matter

over geological time. The formation's geological characteristics affect the reserves' distribution and properties.

Falebita (2014) also highlighted the challenges and uncertainties associated with estimating the reserves' properties, such as porosity, permeability, and oil saturation, and the need for careful consideration of the geological and geostatistical factors affecting the accuracy of the estimates. It is vital to observe and analyse the geological characteristics of the formation, which are essential for understanding the distribution and properties of the bitumen reserves. This understanding is necessary for effective reservoir management and production optimisation. There is a need for accurate characterisation for effective reservoir development and production strategies.

Artificial Intelligence techniques have been used in petroleum engineering to predict various reservoir properties such as porosity, permeability, water saturation, lithofacies and wellbore stability. Artificial intelligence (AI) technologies have recently attracted enormous interest due to their rapid response speed and for promoting conventional problems containing a large spectrum of reservoir engineering. Hence, many advanced machine learning algorithms such as artificial neural networks (ANN), fuzzy logic (FL), genetic algorithm (GA), response surface model (RSM), particle swarm optimisation (PSO), and support vector machine (SVM) have been utilised in many oil and gas reservoirs. (Lim, 2005). The most extensively used of these techniques is Artificial Neural Networks (ANN). More recent techniques, such as Support Vector Machines (SVM), have featured in the literature with better performance indices. However, SVM has not been widely embraced in petroleum engineering as a better alternative to ANN. ANN has been reported to have a lot of limitations, such as its lack of global optimism. On the other hand, SVM has been introduced as a generalisation of the Tikhonov Regularization procedure that ensures its global optima and offers ease of training (Anifowose et al., 2011).

Geological Studies on Nigerian Bitumen Deposits

According to (Onoh, 1983), the Nigerian Bitumen Company, a German Company, started exploration for bitumen in 1908 along the coastal region stretching from Okitipupa in Ondo State to Lagos. Amongst the workers is the geological survey team from the University of Ife (now Obafemi Awolowo University), Ile-Ife, led by Professor O. S. Adegoke between 1974 and 1985. Their efforts have led to several reports and publications. According to Adegoke and his team, the natural bitumen deposits are located within a belt spanning across from east of Ijebu-Ode in Ogun State, through Okitipupa and Ifesowapo areas in Ondo State to the western margins of Delta State. The asphalt-bearing zone is 120 km long and about 6 km wide, and the thickness ranges from 0.5m to over 40m with an average of about 6m (Adedimila, 2004). A conservative estimate of the recoverable asphalt from the bituminous sands was also found to be 270 million metric tonnes, which compares favourably with other significant deposits. A few engineering studies, one of which was conducted by Falebita (2014), have also been conducted to appraise the quality of material from the bitumen-rich areas. Unfortunately, however, the effectiveness and reliability of most of these attempts and the corresponding findings and assessments have been adversely affected by one major constraint or the other, notable among which have been the seemingly recurring inability to obtain sufficient quantity and variety of samples from the bitumen-rich zones and an often severe limitation of specialised equipment for core-sampling, and for in-situ and laboratory testing. (Adedimila, 2004).

Falebita (2014) sought to use ordinary kriging to estimate the bitumen reserves, utilising data on the depth, thickness, and concentration of dried tar obtained from forty boreholes to anticipate bitumen in place (BIP) and recoverable reserves bitumen inside the research field. This was erroneous due to the cursory consideration for spatial variability.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Spatial Interpolation for porosity attributes

The following algorithms were tested because they are some of the most popular algorithms in the literature.

Variogram

The geostatistical tool used most commonly to explain the spatial relationship is the variogram. Geostatistical modelling for computation and modelling starts with semi-variograms. In an unsampled area, it then approximates the desired variable. The semi-variogram is defined in mathematical form as:

$$\gamma(\vec{L}) = \frac{1}{2}\vec{V}[\vec{X}(\vec{U}) - X(\vec{U} + \vec{L})] \qquad (2.4)$$

It is half of the difference between the two values that are divided by $\vec{L}$
The mean square pair difference, roughly separated by L, is a variogram (for lag distance L). It is known, mathematically, as

$$\gamma(\vec{L}) = \frac{1}{2n(\vec{L})}\sum_{i=1}^{n(\vec{L})}[\vec{X}(\vec{U}_1) - X(\vec{U}_1 + \vec{L})]^2 \qquad (2.5)$$

where n($\overrightarrow{L}$)= number of pairs at a lag distance $\vec{L}$;
$\vec{X}(\vec{U}_1)$ and, $(\vec{U}_1 + \vec{L})$ = data values for the pair located at lag distance apart, and $\gamma(\vec{L})$ is calculated based on the sample data value (Mondal, 2015).
The variogram is used to analyse and model spatial variability of Petro-physical properties and lithofacies. Also, 90% of geostatistical reservoir characterisation studies use geostatistical modelling methods based on variograms (Gringarten & Deutsch, 1999). By analysing sample pairs, a Variogram is measured and can be data of well-log, traditional seismic depth-point data or cores (Yarus, 2015). A few measurements would be very close together with a square difference close to zero. Those measures would have a higher value for the square

differences at greater separation distances. However, there is little possibility that the values will take any usual behaviour when separation times get to a specific distance. A single pair of points can have a low or high value or two values of similar magnitude at a considerable lag.

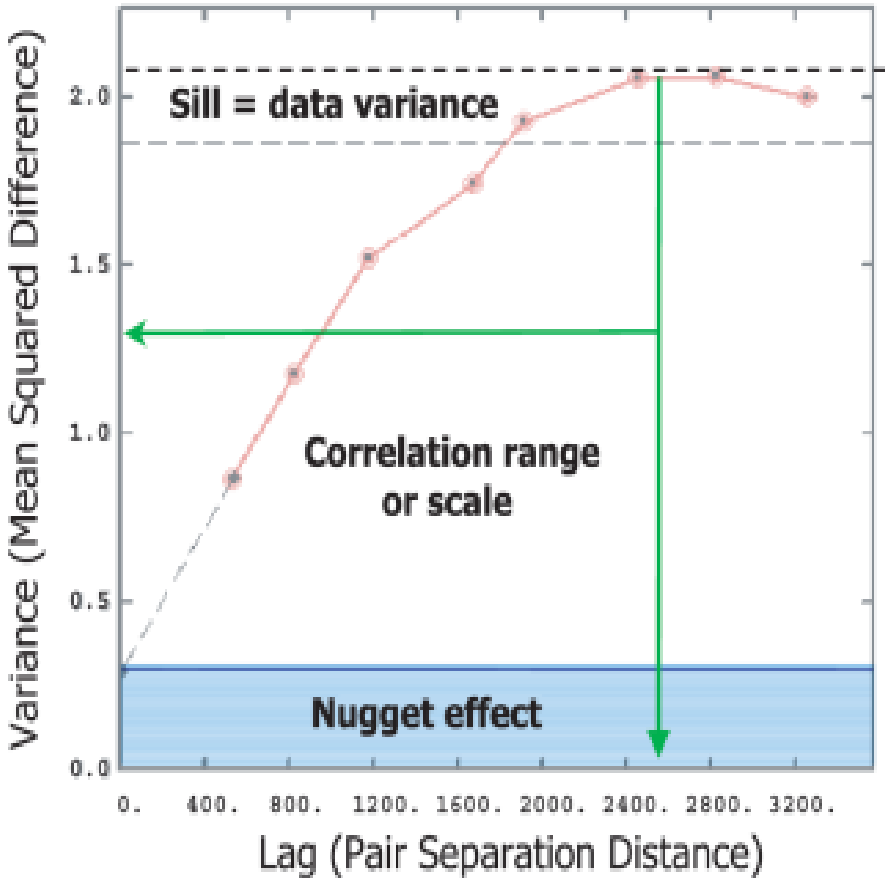

Figure 1.1: Omnidirectional experimental (semi)variogram (Yarus, 2015)

The necessary modelling components of the variogram are shown in Fig. 1.1. Modelling with or without a nugget can affect mapping. The objective of modelling is to have spatial weighting scheme variograms. The variation in correlation range with azimuth can be detected by constructing directional variograms. Evaluating the minimum and maximum continuity directions in practice is tedious, requiring several variograms with several azimuths to be

constructed. Fortunately, by computing the variogram "plot," which reflects variance for all, most

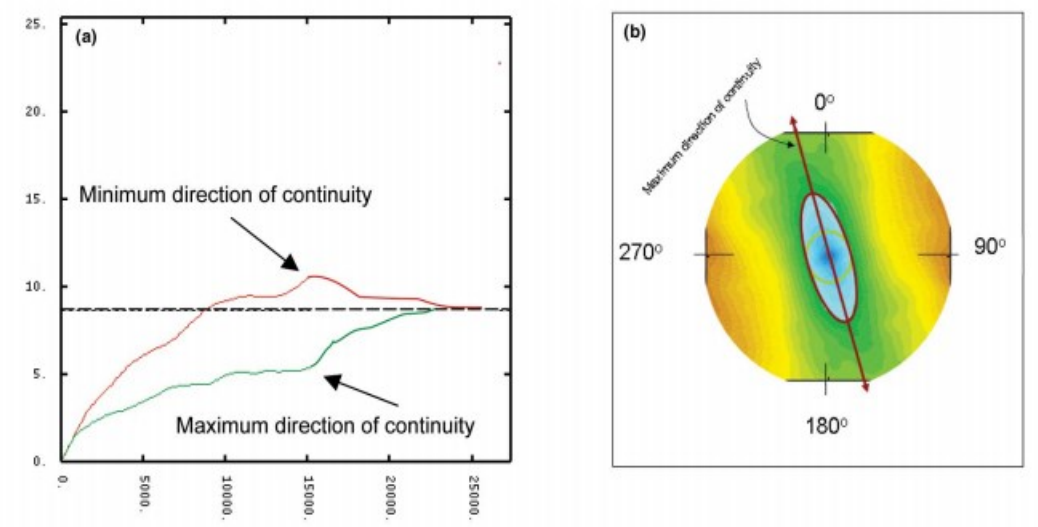

Figure 1.2: Bidirectional variogram: (a) indicates the consistency of the minimum and maximum directions. (b) simultaneously displays all continuity trends; variance rises from light blue (low) to brown (high). (Yarus, 2015).

Ordinary kriging

Kriging (KG) is a geostatistical technique for estimating properties at locations that do not have measured data. In other words, KG is the geostatistical method of predicting values at unsampled points, a form of multi-dimensional interpolation commonly used to build the proxy model in petroleum reservoir studies. It uses a variogram model (a measure of spatial correlation) to infer the weights given to each data point. It is worth mentioning that KG is similar to other interpolation methods, such as radial basis function (RBF) and spline. Besides, it is a combination of a polynomial model, a global function over the entire input space, and a localised deviation model based on the spatial correlation of samples. The main goal of the KG is to predict the values of stationary covariance at the unsampled point concerning the mean squared error. The covariance function is not commonly known and needs to be estimated. Some types of KG, ordinary kriging, simple kriging, universal kriging,

and co-kriging, presented the details and their mathematical derivation. The authors give more information on this technique.

Isaaks & Srivastava., (1990) referred to this method as the Best Linear Unbiased Estimator (BLUE) because it gave a weighted linear combination of data as estimates, its residual mean (mean error) is approximately zero and minimising the variance of error, the distinguishing feature is the aim. In a practical sense, the mean and variance of error are unknown. A random function determines the weight assigned to each unsampled point, accounting for its unbiased nature. These weights and known data points,, would be used to determine the unknown points, $\check{v}$.

$$\check{v} = \sum_{j=1}^{n} w_j . v \tag{2.9}$$

We must limit the weights to one since, in the intense case where all data values are equal to a constant, the estimated value should also correspond to that constant, though weights may vary from one location to another. According to Wackernagel (2003), the commonly used model is ordinary kriging, and it helps estimate a value at a point in a region where a variogram is known by using estimation location data in the neighbourhood. A correlation model (either a variogram or a correlogram) is necessary in this kriging method. This model makes ordinary kriging accountable for qualitative insights such as anisotropy patterns. The variogram is the most common indicator of a random field's spatial variability. However, variogram estimates are obtained at discrete distances between sample point values. In practice, such functions are obtained by fitting the variogram estimates with a smooth curve, known as variogram modelling.

Background of SVM

SVM, first suggested by Cortes and Vapnik, is the supervised learning approach for classification, pattern recognition, and regression. It is based on the

statistical theory and the approach of minimising structural risk. SVM is a computer technology that employs space for high-dimensional functions. It gives prediction functions that are expanded on a subset of vectors that support (Jiang et al., 2017). Introducing Vapnik's ε-insensitive loss function into the SVM algorithm resulted in support vector regression (SVR) answering the non-linear regression problem (Oloso et al., 2009). It was also proven that the regression variation of SVM, called SVR, demonstrates outstanding efficiency. A fundamental advantage of adopting SVR is that it is a non-parametric approach. Unlike SLR, whose effects rely on Gauss-Markov assumptions, SVR's output model is independent of distributions of the dependent and independent variables underlying it. Instead, the SVR approach depends on the kernels. Another advantage of SVR is that it allows a non-linear model to be generated without affecting the explanatory variables, helping to comprehend the final model further. The key downside of this technique is that the data may also be over-fit. (Veronesi & Schillaci., 2019). The modelling objective is constructing a scenario nearest to the underlying feature in the target region (about a particular error) (concerning a specific error). According to Gunn (1998), two circumstances led to inaccuracies in this respect:

a. Approximation Error is the consequence of the judgment that the region of the target area is less and the underlying characteristic is beyond the area of the hypothesis. A poor collection of the sample space leads to a high approximation error and is termed model failure.

b. Estimation Error is a learning error resulting in a strategy of picking the non-optimal model from the space of the hypothesis. It is these two errors that produce the generalisation error.

Assumptions of regression

One of the assumptions of regression is that it has a constant variance of error terms. This may be recognised using scatter plots of the residual versus the fitted values. When there is a funnel-like structure of the residual plot, it is said to have a non-constant variance of error term. The fit line is positioned at y = 0.

Points above the fit line are positive residuals, while those below are 22 harmful residuals. Regression also implies uncorrelated error terms, i.e., lack of trends or patterns. When there are trends or patterns, the error terms are associated. This can be detected via scatter plots (Casella, 2002). Another critical difficulty of spatial interpolation is the occurrence of outliers in the data. In statistics, outliers refer to excessively big or tiny numbers sometimes found on histograms or scatter plots. Outliers result from falling on the extreme, either very high or tiny, relative to the rest of the data. There might be more than one outlier, depending on the dataset. Outliers alter the mean and variance disproportionately by a few extreme values. This may be determined via scatter plots of standardised residual against the fitted values (Casella, 2002).

Methodology

This chapter focuses on the methodology and spatial interpolation techniques used with the machine learning algorithm for estimating spatially continuous data on properties for non-sampled locations, evaluating the estimates as a function of sample location configuration and data values. R statistical programming language is the tool that was used to build algorithms and validate the results of the entire simulation process in this project.

R Programming language

R is a popular statistical programming language widely used in data analysis, statistical modelling, and graphical representation. Developed by Ross Ihaka and Robert Gentleman in the early 1990s, R provides various statistical and graphical techniques, making it a powerful tool for researchers and data scientists. (R Core Team., 2021). Some of the features and capabilities of R are as follows;

i. Extensive Statistical Functions: R offers a comprehensive set of statistical functions for data manipulation, descriptive statistics, inferential statistics, hypothesis testing, regression analysis, time series

analysis, and more. It provides a wide range of built-in functions and packages that are contributed by the R community.

ii. Graphics and Visualization: R excels in data visualisation with its versatile graphical capabilities. It provides numerous packages, such as ggplot2, lattice, and base graphics, that allow users to create high-quality plots, charts, and graphs. These visualisations help in understanding data patterns, relationships, and trends.

R Studio

It is an integrated development environment (IDE) designed explicitly for the R programming language. It provides a user-friendly interface that enhances the productivity and efficiency of R users. RStudio offers a range of features and tools that facilitate coding, data analysis, visualisation, and collaboration. (RStudio Team., 2021)

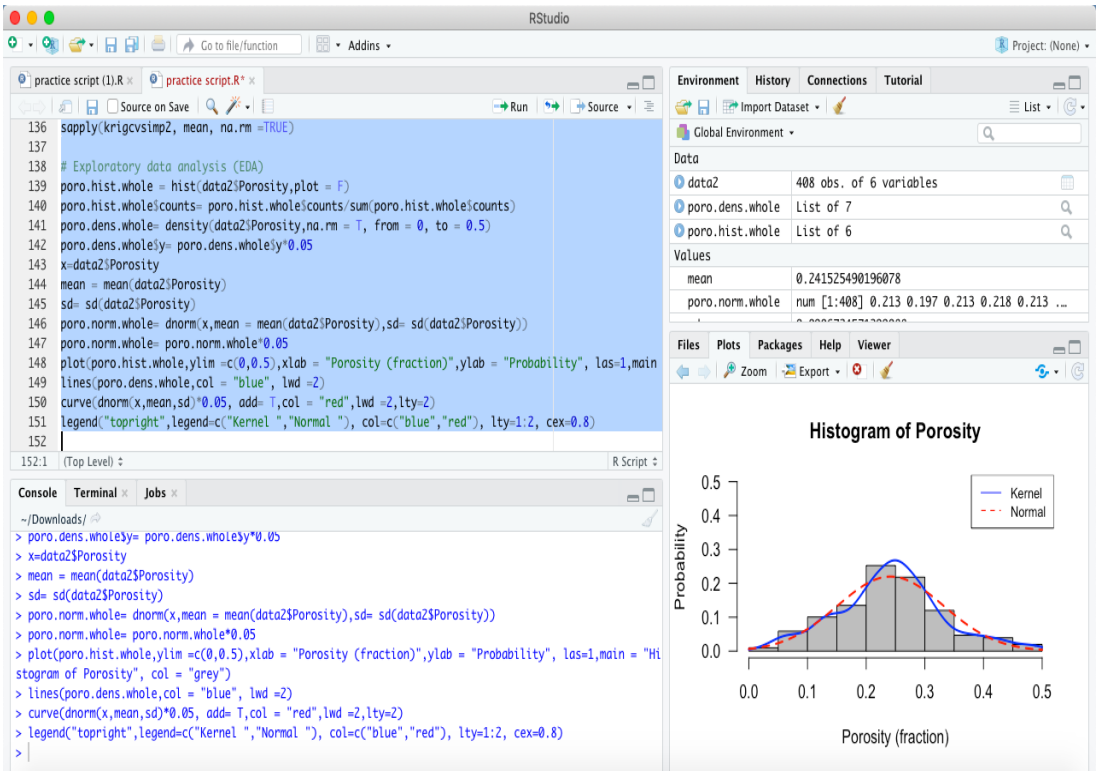

Figure 2.1: RStudio Interface

gstat package

The "gstat" package is an essential tool in R for geostatistical analysis and spatial modeling. It provides various functions for analysing and modelling spatially correlated data, performing geostatistical interpolation, and conducting uncertainty assessments. (Pebesma, 2004). Some of the features and capabilities of the "gstat" package include;

i. Variogram Analysis: The gstat package allows users to calculate and model variograms, which describe the spatial correlation structure of a variable of interest. Variograms help understand the spatial dependency and provide essential information for subsequent geostatistical analyses.
ii. Spatial Interpolation: Using the variograms, the gstat package offers several interpolation methods, including ordinary kriging, universal kriging, and co-kriging. These methods allow for predicting values at unobserved locations based on the observed values at nearby locations while considering the spatial correlation structure.

Data Analysis

Assumptions for Regression Models

Regression analysis is commonly used **to model the relationship between a response and a predictor. These assumptions are primary conditions that should be met before drawing inferences about the estimates of the model or before making predictions using a model. The hypothesis associated with the regression model is:**

1. Variance of Error Terms
2. Correlation of Error Terms
3. Outliers

SVM Implementation

By the optimal SVR model, the predicted values ranged from 0.02 – 0.38 approximately, as seen in Figure 4.2. After standardisation of training data, the

optimal SVM model was built by C and σ setting as 1 and 1, respectively, with the minimum RMSE. Based on the increase in RMSEs, the SVR model showed topographic indicators as the most important predictor for explaining the spatial variations of porosity.

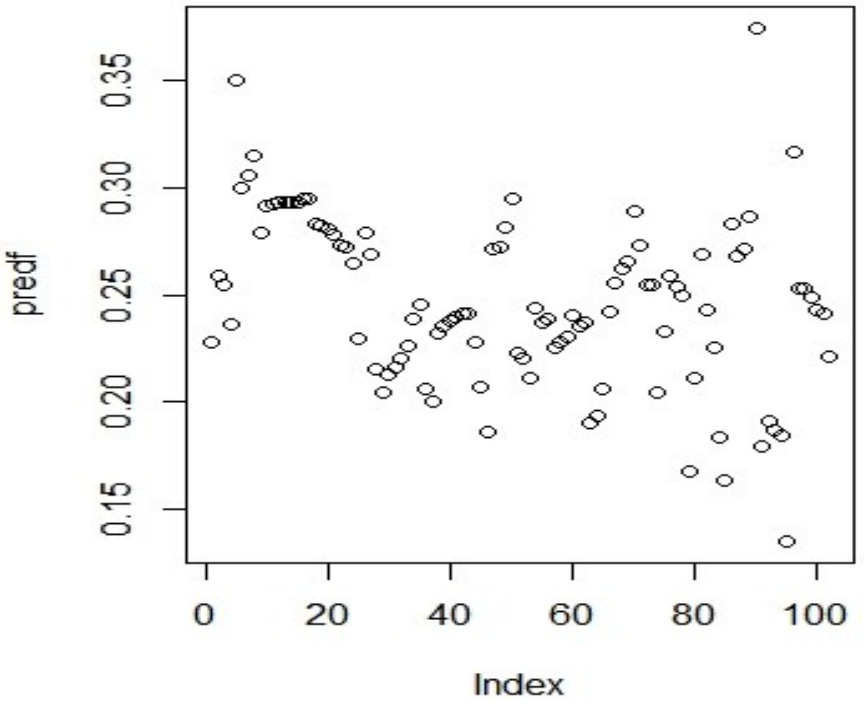

Figure 2.2 Plot showing predicted porosity from SVM Implementation.

Variance of error terms

The residuals are defined as generally distributed as they hover closely around the fit line, as seen in Figure 2.1 to Figure 2.2. Ordinary kriging in Figure 2.1 to Figure 2.2, residuals which are deviations from the actual values, seemed to possess a funnel shape at the beginning but halfway were randomly distributed, did not possess a trend, and variance was reasonably constant across the predictor range (homoscedastic) of about -0.23 to 0.31. From the fit line, the range on both sides is almost equal. (James 2006)

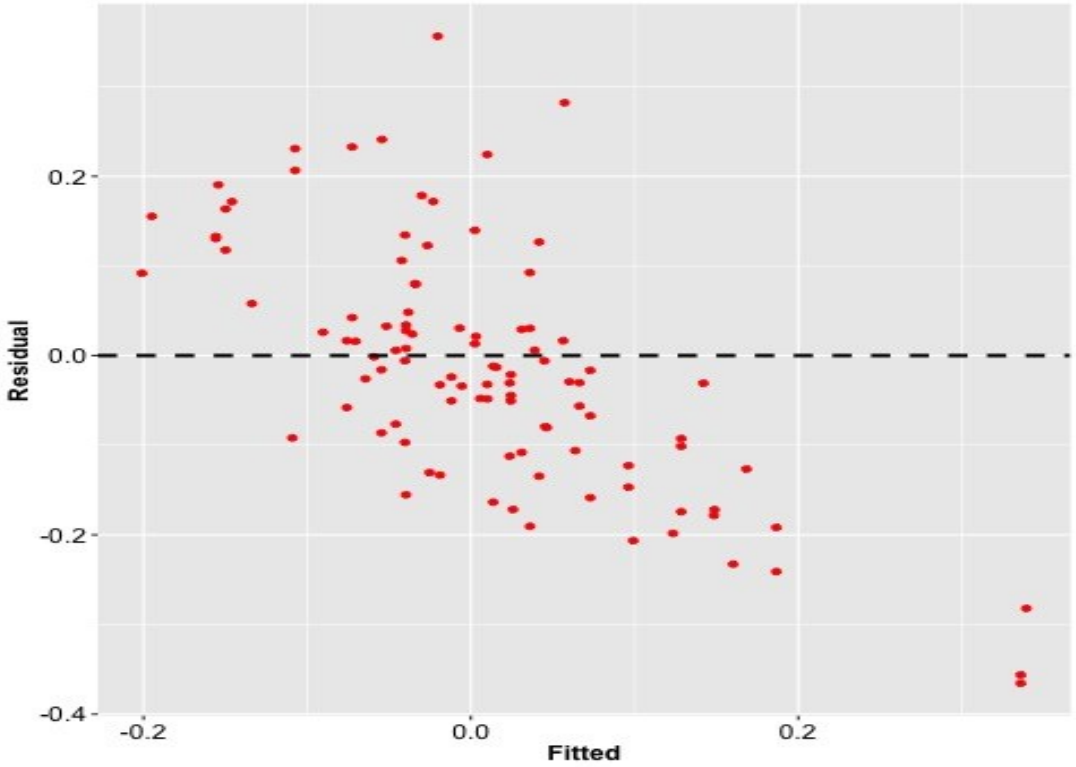

Figure 2.1: Plot showing 5-fold cross-validation of ordinary kriging

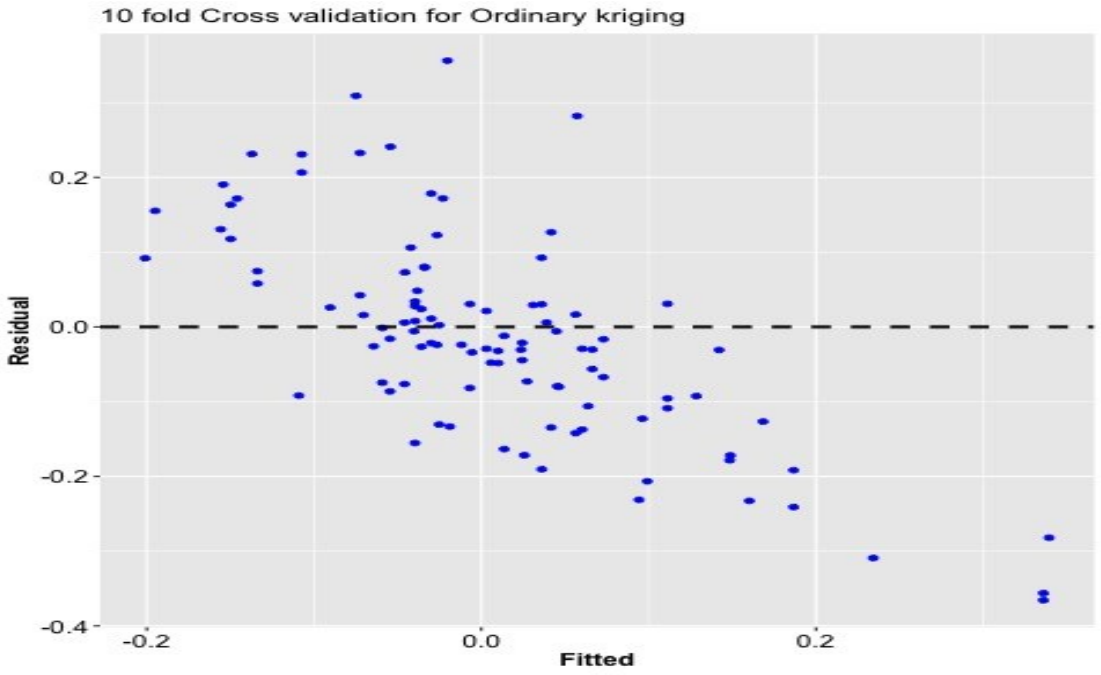

Figure 2.2: Plot showing 10-fold cross-validation of ordinary kriging

Correlation of error terms

The residual plots in Figures 2.1- 2.2) should not show any trend or pattern, as this is known as correlation. The residuals do not demonstrate a systematic fluctuation; they fluctuate arbitrarily around zero. If it does, there is an issue since one of the model's assumptions is that error terms are not correlated (James, 2006). Although they do not own a trend or structure, some plots are more arbitrary.

Conclusion

From the findings of this study, it can be inferred that the support vector machine combined with the ordinary kriging interpolation can be used for spatial interpolation of the porosity attribute of the Agbabu bitumen deposit. These findings also show that SVMOK has high accuracy, a high degree of precision, and higher generalisation performance when combined and is commonly used in reservoir characterisation.

Recommendations

In SVM, the choice of the function of the kernel and the definition of the optimum parameter values have a critical influence on the output of the final system. A systematic process of finding the ideal value of parameters in SVM for the best-predicted outcomes and the influence of other factors, such as the kernel function, should be established.

The second subject for future studies addresses the generalisation of SVM based on the acceptable level of the scale of the training set and a reference for evaluating the performance of the generalisation.

References

Adedimila A. S. (2004). *CHARACTERISATION OF NIGERIA'S NATURAL BITUMEN.* Paper Presented at the National Engineering Conference and Annual General Meeting of the Nigerian Society of Engineers https://api-ir.unilag.edu.ng/server/api/core/bitstreams/7403982e-6c82-4d14-940a-ddffeaa5402d/content.

Anifowose, F. E. (2011). *reduction of oil and gas reservoir properties using support vector machines.* International Petroleum Technology Conference 2011, IPTC 2011. https://doi.org/10.2523/iptc-14514-ms.

Casella, G. (2002). *Statistical inference (2nd ed.) (Vol. 1, Issue September, p. 660). .*

Falebita, D. E. (2014). *A Geostatistical Review of the Bitumen Reserves of the Upper Cretaceous.* Petroleum & Coal, 56(5), 572–581.

Gringarten, E. & Deutsch (1999). *Methodology for Variogram Interpretation and Modeling for Improved Reservoir Characterization.* SPE Annual Technical Conference And Exhibition. https://doi.org/10.2118/56654-ms.

Gunn, S. R. (1998). *Support Vector Machines for classification and regression.* In Analyst (Vol. 135, Issue 2). https://doi.org/10.1039/b918972f.

Isaaks, E. & Srivastava (1990). *Applied geostatistics (Vol. 28, Issue 01)*. Oxford University Press, Inc. https://doi.org/10.5860/choice.28-0304.

James, G. W. (2006). *An Introduction to Statistical Learning with Application in R.* G. Casella, S. Fienberg, & I. Olkin (eds.); Vol. 102). Springer Science+Business Media. https://doi.org/10.1016/j.peva.2007.06.006.

Jiang, H. H. (2017). *Field support vector regression.* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Bioinformatics), 10634 LNCS(10), pp. 699–708. https://doi.org/10.1007/978-3-319-70087-8_72.

Lim, J.-s. (2005). Reservoir properties determination using fuzzy logic and neural networks from healthy data in offshore Korea. *Journal of Petroleum Science and Engineering*, pp. 182–192.

Mondal, S. Z. (2015). *2D geostatistical modelling and volume estimation of an essential part of western inland oil field India.* Spgindia.org. Retrieved 29 August 2020, from https://www.spgindia.org/11_biennial_form/2d-geostatistical-modeling-and-volume-estimation-of-an-important-part-of-western-onland-oil-field-india.pdf.

Oloso, M. A. (2009). *Prediction of Crude Oil Viscosity and Gas/Oil Ratio Curves Using Recent Advances in Neural Networks.* Society of Petroleum Engineers -, 125360.

Onoh, J. K. (1983). *The Nigerian Oil Economy.* St. Martin Press.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences, 30(7)*, 683-691.

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team. (2021). *RStudio: Integrated Development Environment for R.* RStudio, PBC, Boston, MA, USA. URL: https://www.rstudio.com/.

Veronesi, F. & Schillaci (2019). *Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation.* Ecological Indicators, pp. 101, 1032–1044. https://doi.org/10.1016/j.ecolind.2019.02.026.

Wackernagel, H. (2003). *Geostatistical models and kriging.* IFAC Proceedings Volumes (IFAC- 50 PapersOnline), 36(16), 543–548. https://doi.org/10.1016/S1474-6670(17)34818-8.

Yarus, e. (2015). *An Arm Chair Overview for Petroleum Research Engineers.* https://www.researchgate.net/publication/242632297.