11th MALAYSIA STATISTICS CONFERENCE 2024
Data and Artificial Intelligence: Empowering the Future
Sasana Kijang, Bank Negara
19th September 2024

Comparison analysis for variogram models of Teak stands volume specific to Solomon-clone in Tawau, Sabah Malaysia; Effect of bin width

Johannah Jamalul Kiram[1,2], Rossita Mohd Yunus[2], Yani Japarudin[3], Mahadir Lapammu[3], Oliver Monteuuis[4], and Doreen Kim Soh Goh[5]

[1] Preparatory Centre for Science and Technology, Universiti Malaysia Sabah, Jalan UMS, Kota Kinabalu, Sabah, Malaysia
[2] Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, Kuala Lumpur, Malaysia
[3] Sabah Softwood Berhad, Tawau, Sabah, Malaysia
[4] CIRAD-BIOS Department – UMR AGAP, TA A-108/03, Avenue Agropolis, F-34398 Montpellier, Cedex 5 France
[5] YSG Biotech Sdn Bhd, Yayasan Sabah Group, Voluntary Association Complex, Mile 2½, off Tuaran Road, Kota Kinabalu, Sabah, Malaysia

Abstract:

The volume of teak trees, scientifically known as tectona grandis, was examined based on a teak plantation managed by the research and development team at Sabah Softwood Berhad, Brumas camp, Tawau, Sabah, Malaysia. Using the Exponential model as the experimental variogram, different width of bins was applied to obtain different variogram models. These models were graphed for comparison, and the root mean square error was calculated. Cross validation was also done to see how well each model predicts unseen data. The comparison resulted in having bin width 0.003 which is approximately 333 meters apart as the best bin width to model the Exponential model. It showed the least RMSE and best graphical observation. This study proves that the decision to choose the correct bin width to predict a model highly affects its accuracy, despite the size of the sample.

Keywords:

spatial, variogram, bin width

1. Introduction:

This study focuses on the spatial model teak trees, particularly their volume, which is known as a variogram model. However, before a variogram model of the volume of these teak trees can be constructed, selecting the correct bin width is vital as it affects sampling. Despite the number of georeferenced teak trees collected, the correct bin width plays an important role in selecting a proper model to represent the data. The bin width determines the range of lag distances over which pairs of sample points are grouped for variogram calculation (Oliver & Webster, 2015). The lag distance is the distance between a pair of sampled trees. Paired trees depend on sampling too. There may be more trees that can be paired when they are far apart but if the sampling did not cover a certain amount of distance and direction, hence, results will be affected. A variogram's accuracy depends on an adequate amount of data that suits its density, not how large the sample size is, especially if the sample collected does not vary in lag interval. Thus, to ensure reliability, comparisons must be made for every lag interval for better estimating variogram models (Oliver & Webster, 2015).

Scientifically known as *tectona grandis*, planted in Brumas Camp, Tawau, Sabah Malaysia, particularly Solomon Island-derived clones are proven to thrive best in this research. The teak tree itself is eminent in the research field, whether in situ or in vitro, natural stands or plantations, its studies can be found in most tropical parts of the world such as in India, Nepal, Thailand, and Brazil (Ghosh et al., 2019; Kenzo et al., 2020; Koirala et al., 2021; Pelissari et al., 2017; Tewari & Singh, 2018). In spite of the teak tree's eminent profile in the research field, the focus of its spatial importance has been very limited (Kiram et al., 2022, 2023; Pelissari et al., 2017). Studies that focus on the teak trees volume includes effects of different spacings in Brazil (Vendruscolo et al., 2022), a study using artificial neural network and regression to estimate the volume on stems of teak tree (Tavares Júnior et al., 2021), and growth and yield models using linear and multiple linear regression for teak trees in Nigeria (Popoola & Ude, 2024).

This study is based on research in 1994, instigated by Innoprise Corporation Sdn Bhd (ICSB) in an investment for mass cloning of teak trees, within a structure of partnership between the CIRAD Forestry Department and ICSB (D. K. Goh & Galiana, 2000). The success of their research and investment on a monoclonal block in 1997 (D. K. Goh & Monteuuis, 2005) inspired another two provenance-cum-progeny trials within the same year. Detailed reports of its field performances have been published (D. K. Goh & Monteuuis, 2005; D. K. S. Goh et al., 2013; Monteuuis & Goh, 2015, 2017) however never before deepen its spatial statistical research and how it effects the growth. The bole volume of the teak tree derived from previous research (D. K. S. Goh et al., 2013; Monteuuis & Goh, 2015) of similar objectives as in equation (1)

$$V = \frac{1}{10}\left[\left[1.3\pi\left(\frac{D}{2}\right)^2\right] + \left[\pi\left(\frac{D}{2}\right)^2\left(\frac{H-1.3}{3}\right)\right]\right] \tag{1}$$

where V is the volume, given the D which is the diameter at breast height (1.3meters above ground) and H is the height of the tree. This partnership then widens their research and implanted several teak plantations in different districts of Sabah, for research purposes and for yielding. Hence, providing data for this study is the teak plantation managed by a research team at Sabah Softwood Berhad at the Brumas camp in Tawau district, Sabah, East Malaysia. Spatially continuous data are crucial in all ecological system including forests for decision-making (Karahan & Erşahin, 2018).

Small bin widths will produce detailed variogram but noisy due to limited sample pairs, but large bin width will produce smoother variogram, but possibly hiding important spatial patterns, bringing correct bin widths into question. The objective of this study is to determine the correct bin width to obtain the best variogram model for the volume of tectona grandis, specific to the Solomon Island-derived clone. Identifying the relationship between the physical parameters and the spatial information of the tree has previously been executed (Kiram et al., 2022, 2023). Existing experimental variogram models, specifically exponential model will be used to examine the effect of bin widths. These models are then graphed for comparison, and the root mean square error will be calculated. Cross validation will also be done to see how well each model predicts unseen data. This study focuses on small plantation sites as it helps to give estimation and understanding of how spatial dependence will affect bigger plantations(Kiram et al., 2022).

2. Methodology:

2.1 Data

This study is a continuance of previous study (Kiram et al., 2022) it focuses on one block dedicated to *tectona grandis* that is managed by the research and development team of Sabah Softwood Berhad, at Brumas camp, Tawau, Sabah, Malaysia whereby its data have been observed throughout the span of 12 years, located on the coordinates 4°37’23.85’’N and 117°47’05.12’’ E. Figure 1 below shows the topological features of the site, provided by the Sabah Softwood Berhad research team.

These plots were designed using randomized complete block with four contiguous replications, comprised two rows each of 30 plants of the 15 different genotypes. The plots were spaced 4 x 4 m with 625 stems per hectare, resulting in over 4000 trees. Assessed rows were only the 11th to 20th plants of each row, corresponding to 80 plants per clone in all. Block 96G situated on slopy land with an altitude ranging from 180 to 370 meters above sea level. The region’s climate was classified as tropical rainforest (Köppen), while its soil is classified as Tg Lipat soil with low content of nitrogen, potassium, and magnesium.

The sample data of 1200 trees were collected including their height, and diameter at breast height (DBH). These data are then used to calculate its volume according to equation (1). However, specific to Solomon Island-derived clone, and not considering those that are undergrowth, or those that have fallen, the total samples collected are 451 trees. The statistical analyses in this study were executed by using R Studio 4.0.5 and Microsoft Excel. Out of these samples, there are 432 georeferenced individual tree points obtained for the 6th year plot. The georefencing points were accurate up to six decimal points. Further geostatistical analysis was done using ArcGIS 10.8.1 and R Studio 4.0.5.

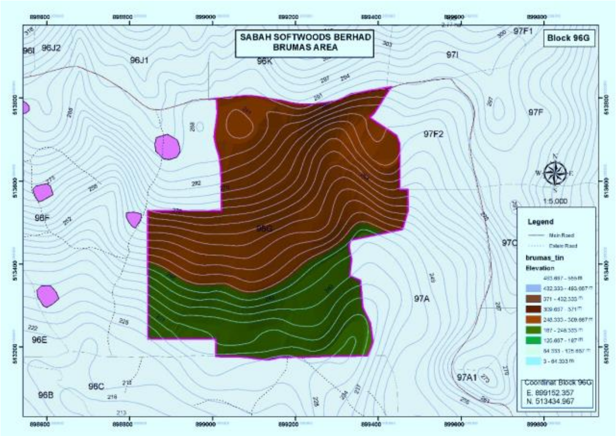

Fig. 1. Topologic map of block 96G at Brumas camp, Tawau, Sabah.

2.2 The Experimental variogram

The semivariogram, frequently referred to as just variogram, is a statistic that assesses the average of how the similarity between two random variables lessens, as the distance between the variables grows, and it leads to an applications in exploratory data analysis

(Olea, 1999). A variogram model assumes that the samples that are in a nearby location tend to behave in a similar way compared to the things that are farther away. In this study, two random variables $Z(x_i)$ and $Z(x_i + h)$, where it is the volume of the tree is assessed, and the difference between the volume of two trees are taken for an $N(h)$ number of pairs, the average is then taken. In this context, x_i and $x_i + h$ are the spatial positions separated by a vector, h. In other words, it measures the spatial dependence between two observations as a function of distance. The distance between two points, h, will not be equal to 0, because there cannot be two trees at the same exact spot. Thus, the variogram function will vary from more than zero, up until the highest value of h, where the points will be farthest away from each other. To conduct variogram, the data must be approximately normal, and anisotropy must also be checked before assuming isotropy to avoid high error. In the equation (3) below, $\gamma(h)$ is the semivariance of $Z(x_i)$ variable, $N(h)$ is the number of pairs of plots for each lag, and distance h (Olea, 1999).

Bringing into the next step, which is the experimental variogram, where it predicts the value of the targeted variable on the unsampled location within the study area. When the spatial autocorrelation is modelled using the spatial data, the experimental model is accomplished. This study focuses on Exponential theoretical variogram to test the effect of bin-width towards a model's accuracy. The variogram models are as shown in Table 1.

Table 1. Empirical variogram and theoretical variogram.

Model	**VARIOGRAM**	
Empirical	$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2$	
Theoretical-Exponential	$\gamma(h) = C_0 + C\left[1 - e^{(\frac{-h}{a})}\right]$	if, $0 \leq h \leq a$. Otherwise, $C_0 + C$.

The bin-width that are being compared are 555 metres apart, 333 metres apart, 277.5 metres apart and 222 metres apart. This is using latitude conversion where 1 degree of latitude is approximately 111 kilometers, thus using bin-widths of 0.005, 0.003, 0.0025 and 0.002 respectively on the Rstudio command for fit.variogram under the package 'gstat'.

Cross validation is then carried out to ensure reliability and the variogram represents the spatial structure better, graphical observations are done and the root mean square error(RMSE) of each model is calculated.

3. Results:

Figure 2a, 2b, 2c and 2d shows the fitted Exponential variogram to the empirical variogram with different bin widths. As observed, the graph with the bin width of 0.002 in Figure 2d shows the smoothest fit, whilst the bin width of 0.005 in figure 2a shows clear and concise fittings. However, this makes concluding on graphical observations alone difficult as the main aim of modelling is always reliability and precision.

Cross-validation is then carried out as shown in figure 3a, 3b, 3c, and 3d. The prediction errors for all four different bin widths are relatively symmetric around zero, which indicates balanced errors without systematic bias. All four graphs suggest small errors overall, due to the narrow histograms. However, close observations do show that the histogram for bin-width 0.003 in figure 3b has a slightly higher concentration of errors around zero, which implies slightly better overall prediction accuracy, indicating that is could be the more accurate model.

The final calculations of RMSE, nugget, psill and range is shown in Table 2, suggest that the most accurate model is the model with bin width 0.003 with the smallest RMSE calculated of 0.05361541. Thus, making the exponential model of bin width 0.003 (333metres apart) to be the best model compared to the others.

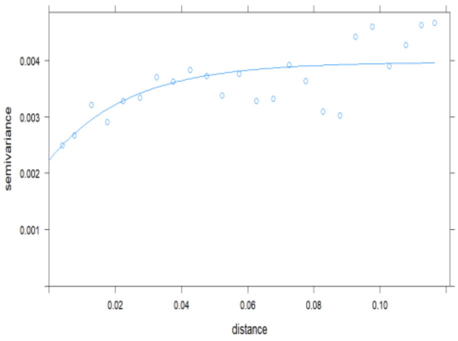

Fig. 2a. Bin-width 0.005

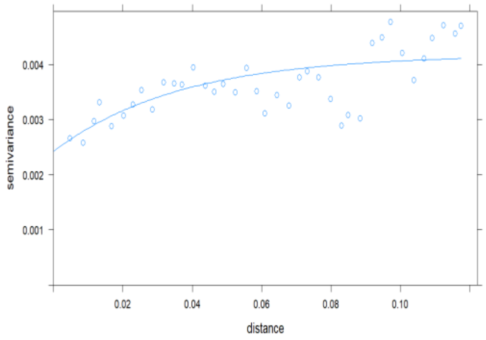

Fig. 2b. Bin-width 0.003

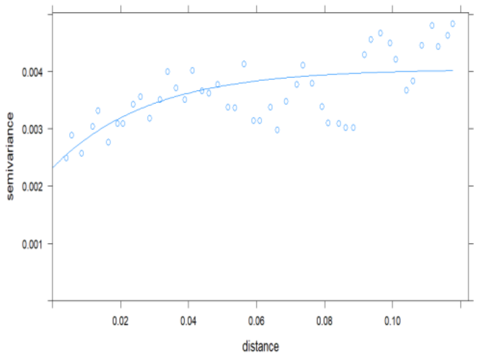

Fig. 2c. Bin-width 0.0025

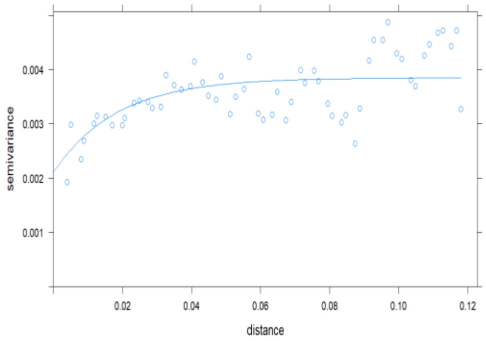

Fig. 2d. Bin-width 0.002

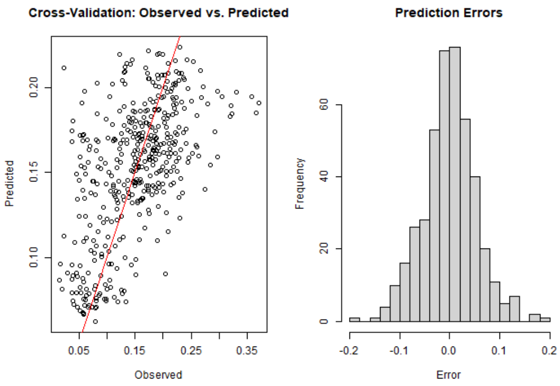

Fig. 3a. Cross-validation of model with bin-width 0.005

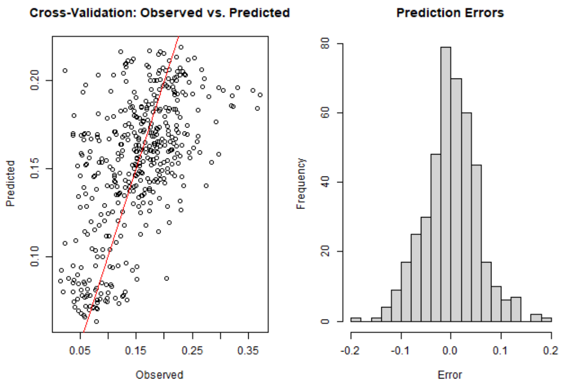

Fig. 3b. Cross-validation of model with bin-width 0.003

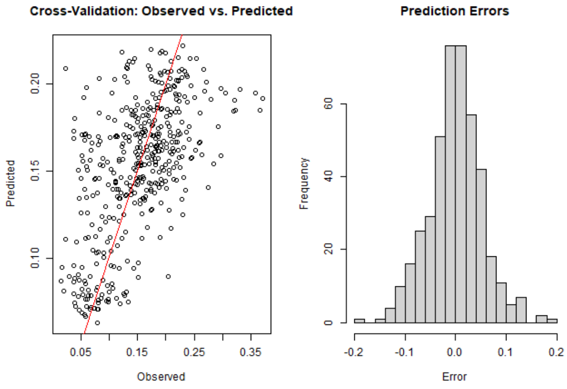

Fig. 3c. Cross-validation of model with bin-width 0.0025

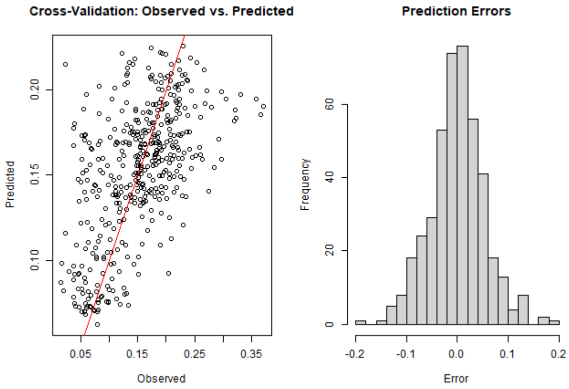

Fig. 3d. Cross-validation of model with bin-width 0.002

Table 2. Calculated RMSE, Nugget, psill and Range for all models

Bin Width	pairs	RMSE	Nugget	Psill	Range
0.005	24	0.05383854	0.002230198	0.001740518	0.02396327
0.003	39	0.05361541	0.002424124	0.001753697	0.03648523
0.0025	47	0.05372381	0.002320105	0.001727215	0.02840659
0.002	59	0.05404312	0.002111044	0.001736450	0.01845881

4. Discussion and Conclusion:

This study proves that the decision to choose the correct bin width to predict a model highly affects its accuracy. Comparison with different bin width must be made first before proceeding with selecting a model. In this study, the best bin width is 0.003, approximately 333 meters apart in lag distance between the tree pairs, which yielded 39 pairs of trees. It gave the least amount of error and its prediction errors shown in figure 3 suggested higher concentration around zero, which indicates better prediction accuracy.

While most statistical models suggest that the bigger the sample size, the better the model will be, it appears different when it comes to spatial modelling. Bin width and lag distance plays a huge role in determining the reliability and accuracy of the spatial model. If the data is sampled in every direction possible, with a variety of lag distances enough to cover the whole research site, then this may be the case. However, when it comes to spatial data, it could be costly, difficult and may even be dangerous to ensure that the sample collected is wide enough to cover what is needed. Thus, amongst reason why spatial modelling is also crucial is to be able to predict without having the need to go through difficult and dangerous sampling tasks especially when it involves data that are within a forest area where it is the habitat for so much wildlife.

References:

Ghosh, S., Nandy, S., Mohanty, S., Subba, R., & Kushwaha, S. P. S. (2019). Are phenological variations in natural teak (Tectona grandis) forests of India governed by rainfall? A remote sensing based investigation. *Environmental Monitoring and Assessment*, *191*(S3), 786. https://doi.org/10.1007/s10661-019-7680-0

Goh, D. K., & Galiana, A. (2000). *Vegetative propagation of teak*.

Goh, D. K., & Monteuuis, O. (2005). *Rationale for developing intensive teak clonal plantations, with special reference to Sabah*.

Goh, D. K. S., Japarudin, Y., Alwi, A., Lapammu, M., Flori, A., & Monteuuis, O. (2013). *Growth differences and genetic parameter estimates of 15 teak (Tectona grandis Lf) genotypes of various ages clonally propagated by microcuttings and planted under humid tropical conditions*.

Karahan, G., & Erşahin, S. (2018). Geostatistical analysis of spatial variation in forest ecosystems. *Eurasian Journal of Forest Science*, *6*(1), 9–22.

Kenzo, T., Himmapan, W., Yoneda, R., Tedsorn, N., Vacharangkura, T., Hitsuma, G., & Noda, I. (2020). General estimation models for above-and below-ground biomass of teak (Tectona grandis) plantations in Thailand. *Forest Ecology and Management*, *457*, 117701.

Kiram, J. J., Mohamad Yunus, R., Japarudin, Y., & Lapammu, M. (2022). Specifying Spatial Dependence for Teak Stands Specific to Solomon Island-Derived Clones in Tawau, Sabah, Malaysia: A Preliminary Study. *Sustainability*, *14*(10), 6005. https://doi.org/10.3390/su14106005

Kiram, J. J., Yunus, R. M., Japarudin, Y., & Lapammu, M. (2023). *Volumetric model estimation using regression and geostatistical application for the Tectona grandis stands in Sabah*. 020016. https://doi.org/10.1063/5.0110076

Koirala, A., Montes, C. R., Bullock, B. P., & Wagle, B. H. (2021). Developing taper equations for planted teak (Tectona grandis Lf) trees of central lowland Nepal. *Trees, Forests and People*, *5*, 100103.

Monteuuis, O., & Goh, D. K. S. (2015). Field growth performances of teak genotypes of different ages clonally produced by rooted cuttings, in vitro microcuttings, and meristem culture. *Canadian Journal of Forest Research*, *45*(1), 9–14.

Monteuuis, O., & Goh, D. K. S. (2017). *Origin and global dissemination of clonal material in planted teak forests*.

Olea, R. A. (1999). The Semivariogram. In R. A. Olea, *Geostatistics for Engineers and Earth Scientists* (pp. 67–90). Springer US. https://doi.org/10.1007/978-1-4615-5001-3_5

Oliver, M. A., & Webster, R. (2015). The Variogram and Modelling. In M. A. Oliver & R. Webster, *Basic Steps in Geostatistics: The Variogram and Kriging* (pp. 15–42). Springer International Publishing. https://doi.org/10.1007/978-3-319-15865-5_3

Pelissari, A. L., Roveda, M., Caldeira, S. F., Sanquetta, C. R., Corte, A. P. D., & Rodrigues, C. K. (2017). Geostatistical modeling of timber volume spatial variability for Tectona grandis LF precision forestry. *Cerne*, *23*, 115–122.

Popoola, V. D., & Ude, N. G. (2024). Growth and yield models for teak in Shangev-Tiev Plantation, Konshisha Local Government Area Benue State, Nigeria. *Journal of Research in Forestry, Wildlife and Environment*, *16*(1), 115–123.

Tavares Júnior, I. D. S., De Souza, J. R. M., Lopes, L. S. D. S., Fardin, L. P., Casas, G. G., Oliveira Neto, R. R. D., Leite, R. V., & Leite, H. G. (2021). Machine learning and regression models to predict multiple tree stem volumes for teak. *Southern Forests: A Journal of Forest Science*, *83*(4), 294–302. https://doi.org/10.2989/20702620.2021.1994345

Tewari, V. P., & Singh, B. (2018). Total wood volume equations for Tectona grandis Linn F. stands in Gujarat, India. *Journal of Forest and Environmental Science*, *34*(4), 313–320.

Vendruscolo, D. G. S., Cunha Neto, F. V., & Fraga, I. M. (2022). Bark volume and thickness in teak trees with different spacings. *Pesquisa Florestal Brasileira*, *42*, 1–9. https://doi.org/10.4336/2022.pfb.42e201902067

NOTE: THE REQUIRED NUMBER OF PAGES FOR PAPER IS SIX PAGES