Brazilian Journal of Biometrics, **42**, 213–224 (2024)
doi:10.28951/bjb.v42i3.679

BRAZILIAN JOURNAL OF
BIOMΣTRICS
ISSN:2764-5290

ARTICLE

Analysis of stunting in East Java, Indonesia using random forest and geographically weighted random forest regression

Yuliani Setia Dewi[1],⋆ Silvia Hastuti[1], and Mohamat Fatekurohman[1]

Department of Mathematics, University of Jember, East Java, Indonesia
⋆Corresponding author. Email: yulidewi.fmipa@unej.ac.id

(Received: July 29, 2023; Revised: October 23, 2023; Accepted: November 21, 2023; Published: August 30, 2024)

Abstract

Stunting is one of the problems that the world focuses on today to be resolved immediately. World Health Organization (WHO) stipulates that a country's public health problems are said to be chronic if the stunting prevalence rate reaches more than 20%.The prevalence rate of stunting in Indonesia in 2021 reached 24.4%. This study aims to analyze factors that correlate with the prevalence of stunting in East Java Province using machine learning methods: Random Forest Regression (RFR) and Geographically Weighted Random Forest (GWRF) methods. The results of this research are the factors that correlate with the prevalence of stunting based on the RFR method, namely the number of babies who get early breastfeeding initiation, the number of malnourished toddlers, and the number of active integrated health posts. The RFR method results in RMSE values of 3.014, MAPE 11.69%, and R^2 0.8168. The factors that correlate with the prevalence of stunting based on the GWRF method are divided into six groups according to the similarity of factors that correlate with stunting in the regency/city. The GWRF method gives better results than the RFR indicated by the resulting RMSE values of 1.023, MAPE 4.45%, and R^2 0.9788.

Keywords: Stunting; Random forest regression; Geographically Weighted random forest.

1. Introduction

Stunting is a condition of stunted physical growth in children due to a chronic lack of nutritional intake (Abdullah *et al.*, 2021; Kemenkes, 2021; Roediger *et al.*, 2020). The World Health Organization (WHO) stipulates that a country's public health problems are said to be chronic if the stunting prevalence rate reaches more than 20% (De Onis *et al.*, 2019; Kadir, 2021). Indonesia is one of the countries with chronic public health problems because the national stunting prevalence in 2021 reached 24.4% (Kemenkes, 2021). East Java Province is one of the priority provinces for handling stunting, according to BKKBN, because it is included in the five provinces with the highest number of stunting cases in Indonesia. In 2021, Bangkalan Regency had the highest stunting prevalence in East Java, reaching 38.9%. The lowest stunting is in Mojokerto City, with a percentage of 6.9% (Kemenkes, 2021). The relationship between stunting prevalence and influential factors is not always linear (Bitew *et al.,* 2022; Sisimayi *et al.,* 2021), and multicollinearity sometimes be found (Chilyabanyama *et al.,* 2022; Sisimayi *et al.,* 2021).

A method that can be used to analyze non-linear data and data with multicollinearity is Random Forest Regression (RFR). RFR consists of a collection of decision trees for regression problems and can model non-linear relationships between dependent and independent variables (James *et al.,* 2013). The research about a factor analysis that affects nitrate concentration using RFR (Ouedraogo *et al.,* 2018) concludes that the RFR Method gives better results than linear regression models, indicated by the larger R^2 values of the RFR method of 0.97 and R^2 of the linear regression model of 0.64. RFR for poverty estimation in Bangladesh (Zhao *et al.,* 2019) results an R^2 value of 0.70. However, the RFR is a machine learning method that works without regard to the spatial aspects of the data.

Differences in stunting prevalence between observation areas can be influenced by different geographical (Ahmed *et al.,* 2021; Menon *et al.,* 2018; Muche *et al.,* 2021), environmental (Budge *et al.,* 2019; Kwami *et al.,* 2019; Titaley *et al.,* 2019), and social factors (Kwami *et al.,* 2019; Mohammed *et al.,* 2019) between regions. Therefore, the factors that influence the prevalence of stunting in each regency in East Java Province are thought to be influenced by spatial heterogeneity. Spatial heterogeneity is found due to the presence of characteristic differences that occur between regions.

One machine learning method that considers spatial heterogeneity in the data is Geographically Weighted Random Forest (GWRF). It is a combined method between GWR and RFR that can analyze data spatially with non-linear relationships between independent and dependent variables. The main difference between RFR and GWRF is that the RFR method analyzes data without considering spatial aspects, while the GWRF method analyzes data by considering it. Some of the previous studies related to the GWRF method include being used in population modeling problems (Georganos *et al.,* 2021), risk factor analysis of COVID-19 mortality rates (Luo *et al.,* 2021), analysis of socioeconomic and environmental factors against poverty in China (Luo *et al.,* 2022), and evaluation of the causes of changes in the amazon forest in Northern Ecuador (Santos *et al.,* 2019). Using GWRF, the result of Georganos *et al.,* 2021 gives an RMSE value of 0.648 in analyzing the population distribution, while Luo *et al.,* 2021 obtained an R^2 of 0.78 in analyzing the distribution of the COVID-19 mortality rate.

In this research, we investigate the factors that correlate with the prevalence of stunting in East Java Province, Indonesia. The difference in stunting prevalence rates in each region shows that it is thought to be influenced by spatial aspects. GWR is a method that works by paying attention to the spatial aspects of data. Previous research on analyzing factors causing stunting using the GWR method was done by Al Azies *et al.,* 2019. However, the method was only limited to linear and non-multicollinearity data. Based on these problems, this research aims to analyze stunting data in East Java Province, Indonesia, using the machine learning method (GWRFR and RFR), with and without considering spatial heterogeneity. The methods are more flexible and can be used for linear or non-linear, and with or without multicollinearity data. Furthermore, we discuss comparing the two approaches in analyzing the stunting data.

2. Materials and Methods

The data in this research are from the publications of the Ministry of Health Indonesia (Kemenkes, 2021), the East Java Provincial Health Office (Dinkes, 2021), and the Central Agency on Statistics of East Java Province (BPS, 2022). The data consists of stunting prevalence data along with factors that are suspected of affecting stunting in nine cities and 29 regencies in East Java Province, Indonesia, obtained in 2021. The variables consist of one dependent variable (Y) and eight independent variables (X), with details in Table 1.

Table 1. Research variables

Variables
Prevalence of stunting cases (Y)
Percentage of poor population (X_1)
Percentage of families with access to proper sanitation (X_2)
Percentage of low birth weight babies (X_3)
Number of active integrated health posts (X_4)
Percentage of babies who obtain exclusive breastfeeding (X_5)
Number of newborns who get early initiation of breastfeeding (X_6)
Number of underweight toddlers (X_7)
Number of pregnant women who received blood-added tablets (X_8)

2.1 Random Forest Regression (RFR)

RFR is a method that consists of a collection of decision trees used for regression problems. The Random Forest algorithm (Schonlau & Zou, 2020) is given below:

1. Tunning *mtry* and *ntree* parameters.
2. Randomly retrieve D_i sample data from dataset D with returns.
3. Build a tree by using D_i sample data.
4. Repeat steps 2-3 as many as k (the desired number of trees).

2.1.1 Parameter Tuning

Parameter tuning is performed to determine the optimal parameters to be used in constructing a Random Forest. Parameters in a Random Forest include *mtry* and *ntree*. The parameter tuning process is carried out with k-fold cross-validation (Probst *et al.,* 2019).

2.2 Geographically Weighted Random Forest (GWRF)

GWRF is a combined method of GWR and Random Forest that can be used to analyze spatial data with non-linear relationships between dependent and independent variables. GWRF method procedures are given below (Luo *et al.,* 2021, 2022)

1. Define weights using kernel functions
2. Determine the optimum bandwidth value
3. Select all neighbors from region i according to the bandwidth value
4. Build a local Random Forest (RF) with the input of region i and its neighbors
5. Analyze the value of variable importance of region i
6. Repeat steps 3-5 for each region i, so that the variable importance value of each region is obtained.

2.2.1 Optimum Weight and Bandwidth

The weight indicates the range of the region to be involved in the modeling. There are two weight functions in the GWRF method, namely fixed kernel and adaptive kernel. A fixed kernel is a kernel that refers to the maximum distance from the region to be involved in constructing the local RF. The adaptive kernel is a kernel that refers to the maximum number of neighbors that will be involved in building a local RF (Georganos *et al.,* 2021). After determining the type of kernel, the next step is determining the optimum bandwidth value. GWRF selects the optimum bandwidth based on the highest R^2 OOB value. The selection of the nearest neighbor that falls within the bandwidth range is determined based on the Euclid distance between the i-th region and the j-th region defined by the formula in equation 1.

$$d_{ij} = \sqrt{\left(u_i - u_j\right)^2 + \left(v_i - v_j\right)^2} \tag{1}$$

where (u_i,v_i) is the latitude and longitude value of region i.

2.2.2 Variable Importance

The degree of importance of an independent variable can be known by using the formula increasing the average error value (%IncMSE). %IncMSE indicates the percentage increase in MSE value when randomization is performed on an independent variable (Liang *et al.,* 2020). The formula of %IncMSE can be seen in Equation 2.

$$\%IncMSE = \left(\frac{MSE_{permuted} - MSE}{MSE}\right) \times 100\% \tag{2}$$

where MSE is mean square error and $MSE_{permuted}$ is MSE value obtained when permutations are randomly performed on an independent variable.

2.3 RMSE, MAPE, and R-Square

The results obtained from the RFR and GWRF methods are then compared to choose which method gives better results in this study. The selection of the best approach is determined based on the values of Root Mean Square Error, Mean Absolute Percentage Error, and R-Square (Feng *et al.,* 2021).

2.3.1 Root Mean Square Error (RMSE)

RMSE is the square root value of the mean squared error (MSE). The formula can be written as below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

where n is the number of observations, y_i is a value of observation result, and $\widehat{y_i}$ denotes the predicted result value.

2.3.2 Mean Absolute Percentage Error (MAPE)

The MAPE value indicates the average percentage of error of the prediction value compared to the observation value

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{4}$$

2.3.3 R-Square

R-Square or coefficient of determination is a value that indicates the magnitude of the influence of the independent variable on the dependent variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{5}$$

where $\bar{y}$ represents the average value of the observation result.

2.4 Research Steps

We analyze the data using the machine learning method, RFR, and GWRF based on the steps below:

1. Descriptive statistics of the data
2. Performing spatial heterogeneity tests using Bruce Pagan (Breusch & Pagan, 1979)
3. Analyzing the Random Forest Regression method (Schonlau & Zou, 2020)
   (a) This step performs tunning mtry and ntree parameters. The parameter tuning process is carried out with *k*-fold cross-validation (Probst *et al.,* 2019). The mtry value option to be used in this research is 1,2,3,4,5,6,7,8, and the ntree value option to be used is 25,50,100,500,1000.
   (b) Analyzing variable importance uses increasing the average error value (Liang *et al.,* 2020).
4. Analyzing the GWRF method
   (a) We use the adaptive kernel function to determine the optimum bandwidth and the increment of the mean square error to find the variable importance of each region.
   (b) We map variable importance based on the similarity of factors correlating with each region's stunting.
5. Comparison of RFR and GWRF based on RMSE, MAPE, and R^2 indicators

3. Results and Discussion

East Java, Indonesia, consists of 29 regencies and nine cities and is geographically located between 111°0' – 114°4' East Longitude and 7°12' – 8°48' South Latitude. In 2021, stunting was more than 20% in East Java. Table 2 presents the descriptive analysis of the dependent and independent research variables. The distribution map of stunting prevalence categories in East Java Province can be seen in Figure 1. The highest stunting locations tend to happen in Madura Island and the east part of East Java.

The stunting data contain multicollinearity. Some variables have high variance inflation factor (VIF) values (Table 3). When we analyze the data using a linear model, it performs poorly, with an R^2 of 0.3608. Based on the Breusch-Pagan test, we reject H_0; the resulting p-value is 0.0192<0.05. It means there is an indication of spatial heterogeneity in the data.

Table 2. Descriptive analysis results.

Variable	Minimum	Mean	Maximum
Y	6.9	21.7	38.9
X_1	4.09	11.32184	23.76
X_2	70.7	95.91579	100
X_3	0.6	6.384211	66.4
X_4	153	975.7632	2803
X_5	42.1	72.81211	92.2
X_6	1664	10432.11	33289
X_7	275	3694.947	1864
X_8	1811	13476.26	41226

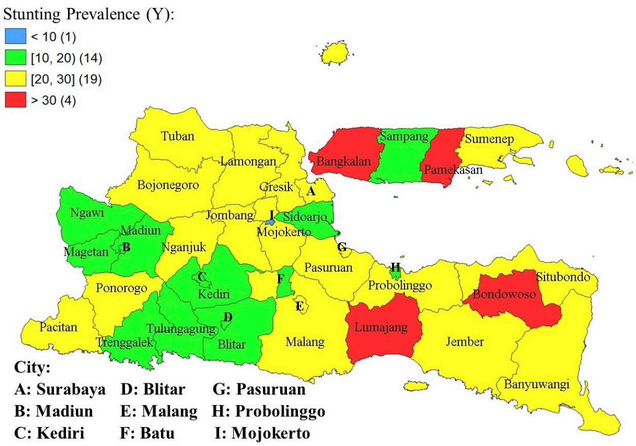

Figure 1. Distribution of stunting prevalence categories.

Table 3. The VIF value.

Variable	VIF
X_1	1.338
X_2	1.667
X_3	1.220
X_4	4.816
X_5	1.337
X_6	25.868
X_7	3.220
X_8	26.950

3.1 Random Forest Regression

We simulate and select the best *mtry* and *ntree* parameters for building the model. The parameters are chosen based on the *RMSE* value. The factors related to stunting in East Java are ranked based on variable importance.

3.1.1 Parameter Tuning

The parameter tuning process in the RFR method is done to find the best mtry and ntree parameter values used to build the Random Forest model. We use eight options 1,2,3,4,5,6,7,8 of the *mtry* parameter and five options 25,50,100,500,1000 of *ntree* parameter. The best parameters are selected using 10-fold cross-validation and determined based on the smallest *RMSE* value produced and presented in Table Table 4.

We find that the *ntree* = 500, *mtry* = 2 has the smallest *RMSE* value. Therefore, we use the parameters for further analysis.

Table 4. RMSE values resulting from tuning mtry and ntree parameters

Number			Number of *ntree*		
of *mtry*	25	50	100	500	1000
1	6.5215	6.0850	6.6361	5.9227	6.1496
2	6.4299	5.8936	6.3754	5.8675	6.1119
3	6.5342	6.0220	6.6765	5.8738	6.2429
4	6.3556	6.1438	6.3914	5.9024	6.2951
5	6.8258	5.9953	6.5812	5.9405	6.3439
6	6.6979	6.0726	6.8340	5.9769	6.3603
7	6.8482	6.1219	6.4916	5.8941	6.3734
8	6.2763	6.0458	6.5332	5.9853	6.3739

3.1.2 Variable Importance

We rank independent variables importance using the increment of Mean Square Error (%IncMSE). The order of the independent variables with the value of %IncMSE from the highest to the lowest is X_6, X_7, X_4, X_1, X_8, X_2,X_5, and X_3, respectively (Table 5). The three variables that most correlate with the prevalence of stunting include the number of babies who get early breastfeeding initiation (X_6), the number of malnourished toddlers (X_7), and the number of active integrated health posts (X_4). %IncMSE values of low birth weight babies (X_3) and percentage of babies who obtained exclusive breastfeeding (X_5) show negative results. A negative value in %IncMSE indicates that the MSE value obtained after random permutation in a predictor variable is smaller than the MSE value before permutation. It means that the response variable has a low correlation with the predictor variable (Du *et al.,* 2019; Meador, 2020; Roth *et al.,* 2021). The percentage of low birth weight babies (X_3) and the percentage of babies who obtain exclusive breastfeeding (X_5) have a low correlation with stunting in East Java.

Table 5. Variable importance value of RFR

Independent Variable	%IncMSE
X_1	5.255670
X_2	2.118345
X_3	-3.106665
X_4	8.262264
X_5	-0.465226
X_6	9.430673
X_7	8.824693
X_8	4.837198

3.2 Geographically Weighted Random Forest

We find the bandwidth based on R^2 OOB value using the adaptive kernel function. Like RFR, the factors related to stunting are ordered based on variable importance. We cluster locations that have similar characteristics.

3.2.1 Optimum Weight and Bandwidth

This research determines the optimum bandwidth value based on the R^2 OOB value of the local Random Forest and uses an adaptive kernel to find the weight (Figure 2).

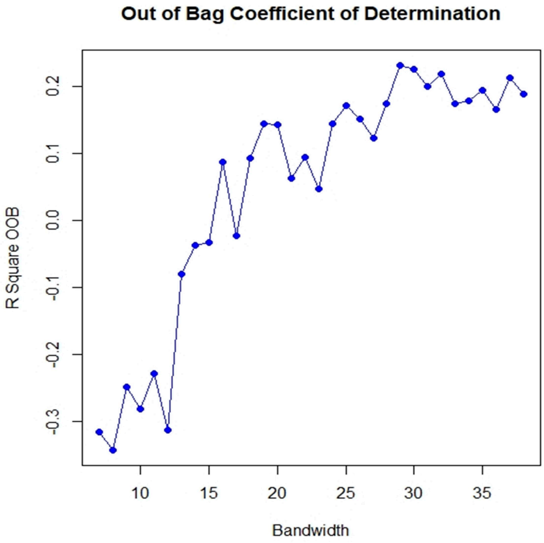

Figure 2. R^2 OOB value on optimum bandwidth search.

Based on the graph in Figure 2, it can be seen that the highest R^2 value of 0.2317 is obtained when the bandwidth is worth 29. This value indicates that the construction of local Random Forests in each regency/city will involve 29 nearby areas. As an illustration, the construction of the local RFR model in Pacitan Regency only involves 29 locations closest to Pacitan Regency, such as

Pacitan, Ponorogo, Trenggalek, Magetan, Madiun City, Ngawi, Madiun, Tulungagung, Kediri City, Nganjuk, Bojonegoro, Blitar, Blitar City, Kediri, Jombang, Batu City, Mojokerto, Mojokerto City, Malang City, Tuban, Lamongan, Malang, Sidoarjo, Gresik, Pasuruan, Pasuruan City, Surabaya City, Lumajang, and Bangkalan.

3.2.2 Variable Importance

The value of the variable importance for each area varies. Based on the three most important variables, we group the locations. Table 6 exhibits this grouping

Table 6. Variable importance of location grouping

Group	Variable importance	regencies/cities
1	X_1,X_4, and X_7	Ponorogo, Lumajang, Jember, Banyuwangi, Situbondo, Probolinggo, Pasuruan, Bangkalan, Sampang, Pasuruan City.
2	X_1,X_6, and X_7	Pacitan, Nganjuk, Magetan, Ngawi, Probolinggo City, Madiun City.
3	X_1,X_7, and X_8	Bondowoso, Malang City.
4	X_4,X_6, and X_7	Malang, Madiun, Sumenep.
5	X_4,X_7, and X_8	Sidoarjo, Gresik, Pamekasan, Surabaya City.
6	X_6,X_7,and X_8	Trenggalek, Tulungagung, Blitar, Kediri, Mojokerto, Jombang, Bojonegoro, Tuban, Lamongan, Kediri City, Blitar City, Mojokerto City, Batu City.

Group four consists of Malang, Madiun, and Sumenep Regency. Factors that correlate with the prevalence rate of stunting in these three regions include the number of active integrated health posts (X_4), the number of babies who get early initiation of breastfeeding (X_6), and the number of malnourished toddlers (X_7). The same explanation also applies to the other five groups. The distribution map of each regency/city based on the grouping of variable importance is presented in Figure 3.

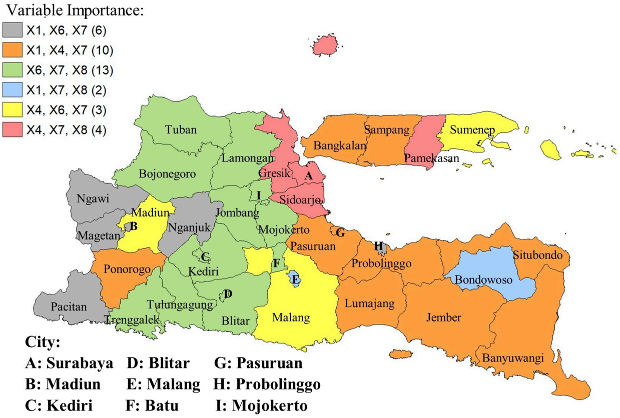

Figure 3. Variable importance grouping distribution map.

Figure 3 shows that Bondowoso Regency is included in a different group from the surrounding areas, such as Banyuwangi, Jember, Probolinggo, and Situbondo Regencies. The difference in groups is due to the stunting prevalence rate in Bondowoso Regency showing a higher percentage than the surrounding regencies. The stunting prevalence rate in Bondowoso Regency shows a percentage of 37%. Unlike Bondowoso Regency, the stunting prevalence rate in Banyuwangi, Jember, Probolinggo, and Situbondo Regencies is 20%-24%.

3.3 Methods Comparison

The comparison of the goodness of the RFR and GWRF methods is based on the RMSE, MAPE, and R^2 values (Table 7).

Table 7. RMSE, MAPE, and R^2

Methods	RMSE	MAPE	R^2
RFR	3.014867	11.69245	0.8168135
GWRF	1.023974	4.452734	0.9788683

The GWRF method is better than RFR; it has smaller MAPE and RMSE values and R^2 values close to 1. The MAPE value of the GWRF method is 7.24% smaller compared to the MAPE value of the RFR method. The R^2 value of the GWRF method also shows a better result of 0.9789.

4. Conclusions

The research results show that East Java, Indonesia's stunting data contains multicollinearity and spatial heterogeneity. Some independent variables have large VIF values. Based on the Breusch-Pagan test, the resulting p-value is 0.0192<0.05. When we analyze it using a linear model, it performs poorly ($R^2 = 0.3608$). Based on the machine learning approach, the RFR method yields an R^2 of 0.8168, RMSE of 3.014, and MAPE value of 11.69%. The GWRF method has better goodness of fit than RFR. It gives R^2 of 0.9788, RMSE of 1.023, and MAPE value of 4.45%. The RFR method results that the factors that correlate with the prevalence of stunting in East Java Province are the number of babies who receive early initiation of breastfeeding (X_6), the number of malnourished toddlers (X_7), and the number of active integrated health posts (X_4). From the GWRF, the factors are different between areas. There are six groups according to the similarity of them, and Bondowoso regency is in different group from the surrounding areas, it has a high stunting prevalence rate.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Abdullah, A. Z., Thaha, R. M., Hidayanty, H., Sirajuddin, S., Syafar, M., *et al.* Risk factor and interventions of behavioral changing strategy in acceleration of stunting prevention: A systematic review. *Enfermería Clínica* **31,** S636–S639 (2021).
2. Ahmed, K. Y., Agho, K. E., Page, A., Arora, A., Ogbo, F. A., Maternal, G. & (GloMACH), C. H. R. C. Mapping geographical differences and examining the determinants of childhood stunting in Ethiopia: a Bayesian geostatistical analysis. *Nutrients* **13,** 2104 (2021).
3. Al Azies, H., Cholid, F. & Trishnanti, D. Pemetaan Faktor-Faktor yang Mempengaruhi Stunting pada Balita dengan Geographically Weighted Regression (GWR). *semnaskes,* 156–165 (2019).
4. Bitew, F. H., Sparks, C. S. & Nyarko, S. H. Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia. *Public health nutrition* **25,** 269–280 (2022).
5. BPS, J. T. *Provinsi Jawa Timur dalam Angka 2022* (BPS Provinsi Jawa Timur, 2022).
6. Breusch, T. S. & Pagan, A. R. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the econometric society,* 1287–1294 (1979).

7. Budge, S., Parker, A. H., Hutchings, P. T. & Garbutt, C. Environmental enteric dysfunction and child stunting. *Nutrition reviews* **77,** 240–253 (2019).

8. Chilyabanyama, O. N. *et al.* Performance of machine learning classifiers in classifying stunting among under-five children in Zambia. *Children* **9,** 1082 (2022).

9. De Onis, M. *et al.* Prevalence thresholds for wasting, overweight and stunting in children under 5 years. *Public health nutrition* **22,** 175–179 (2019).

10. Dinkes, J. T. *Profil Kesehatan Jawa Timur Tahun 2021* (Dinas Kesehatan Provinsi Jawa Timur, 2021).

11. Du, Y., Deng, F. & Liao, F. A model framework for discovering the spatio-temporal usage patterns of public free-floating bike-sharing system. *Transportation Research Part C: Emerging Technologies* **103,** 39–55 (2019).

12. Feng, L., Wang, Y., Zhang, Z. & Du, Q. Geographically and temporally weighted neural network for winter wheat yield prediction. *Remote Sensing of Environment* **262,** 112514 (2021).

13. Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E. & Kalogirou, S. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International* **36,** 121–136 (2021).

14. James, G., Witten, D., Hastie, T., Tibshirani, R., *et al. An introduction to statistical learning* (Springer, 2013).

15. Kadir, S. Nutritional needs of fish to prevent stunting in early childhood. *Journal of Xi'an Shiyou University, Natural Science Edition* **17,** 477–484 (2021).

16. Kemenkes. *Buku Saku Hasil Studi Status Gizi Indonesia (SSGI) Tingkat Nasional, Provinsi, Kabupaten/Kota Tahun 2021* (Badan Penelitian dan Pengembangan Kesehatan, 2021).

17. Kwami, C. S., Godfrey, S., Gavilan, H., Lakhanpaul, M. & Parikh, P. Water, sanitation, and hygiene: linkages with stunting in rural Ethiopia. *International journal of environmental research and public health* **16,** 3793 (2019).

18. Liang, H., Guo, Z., Wu, J. & Chen, Z. GDP spatialization in Ningbo City based on NPP/VIIRS night-time light and auxiliary data using random forest regression. *Advances in Space Research* **65,** 481–493 (2020).

19. Luo, Y., Yan, J. & McClure, S. Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis. *Environmental Science and Pollution Research* **28,** 6587–6599 (2021).

20. Luo, Y., Yan, J., McClure, S. C. & Li, F. Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model. *Environmental Science and Pollution Research,* 1–13 (2022).

21. Meador, M. R. Historical changes in fish communities in urban streams of the south-eastern United States and the relative importance of water-quality stressors. *Ecology of Freshwater Fish* **29,** 156–169 (2020).

22. Menon, P., Headey, D., Avula, R. & Nguyen, P. H. Understanding the geographical burden of stunting in India: A regression-decomposition analysis of district-level data from 2015–16. *Maternal & child nutrition* **14,** e12620 (2018).

23. Mohammed, S. H., Muhammad, F., Pakzad, R. & Alizadeh, S. Socioeconomic inequality in stunting among under-5 children in Ethiopia: a decomposition analysis. *BMC research notes* **12,** 1–5 (2019).

24. Muche, A., Melaku, M. S., Amsalu, E. T. & Adane, M. Using geographically weighted regression analysis to cluster under-nutrition and its predictors among under-five children in Ethiopia: evidence from demographic and health survey. *PloS one* **16,** e0248156 (2021).

25. Ouedraogo, I., Defourny, P. & Vanclooster, M. Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. *Hydrogeology Journal* (2018).

26. Probst, P., Wright, M. N. & Boulesteix, A.-L. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* **9,** e1301 (2019).

27. Roediger, R., Hendrixson, D. T. & Manary, M. J. *A roadmap to reduce stunting* 2020.

28. Roth, M., Michiels, H.-G., Puhlmann, H., Sucker, C. & Hauck, M. Multiple soil factors explain eutrophication signals in the understorey vegetation of temperate forests. *Journal of Vegetation Science* **32,** e13063 (2021).

29. Santos, F., Graw, V. & Bonilla, S. A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon. *PLoS One* **14,** e0226224 (2019).

30. Schonlau, M. & Zou, R. Y. The random forest algorithm for statistical learning. *The Stata Journal* **20,** 3–29 (2020).

31. Sisimayi, C., Mupandawana, M., Mutambwa, M., Sisimayi, T. & Njovo, H. Assessing the Multi-Dimensional Risk of Stunting Amongst Children Under Five Years in Zimbabwe (2021).

32. Titaley, C. R., Ariawan, I., Hapsari, D., Muasyaroh, A. & Dibley, M. J. Determinants of the stunting of children under two years old in Indonesia: A multilevel analysis of the 2013 Indonesia basic health survey. *Nutrients* **11,** 1106 (2019).

33. Zhao, X., Yu, B., Liu, Y., Chen, Z., Li, Q., Wang, C. & Wu, J. Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh. *Remote Sensing* **11,** 375 (2019).