

## ESTIMATION OF JAVA GRDP IN REGENCY/CITY LEVEL: SATELLITE IMAGERY AND MACHINE LEARNING APPROACHES

Anak Agung Gede Rai Bhaskara Darmawan Pemayun<sup>1</sup>, M Ziko Azizi<sup>1</sup>, Nur Ainun Daulay<sup>1</sup>,  
Nur Hidayah Apriliani<sup>1</sup>, Fitri Kartiasih<sup>2,\*</sup>

<sup>1</sup>Statistical Computing, Politeknik Statistika STIS

<sup>2</sup>Statistics, Politeknik Statistika STIS

email: \*fkartiasih@stis.ac.id

**Abstract:** Gross Regional Domestic Product (GRDP) is one of the most important socio-economic indicators. In order to gain a more comprehensive understanding of the current economic situation and regional differences, estimating GRDP using integration of satellite imagery and official statistics data can provide valuable information. This research estimates the GRDP value in 2022 by using data in 2019 to 2021 related to two aspects, agriculture and non-agriculture. Soil adjusted vegetation index (SAVI), enhanced vegetation index (EVI), and land cover (LC) used as agriculture aspect, while nighttime light (NTL), human settlement index (HSI), land area, and population per regency/city used as non-agriculture aspect. GRDP estimation are produced with machine learning approach using support vector machine (SVM) and random forest (RF) method. Correlation test on each variable shows only land area that does not have a significant correlation with GRDP. RF model then chosen as the best model with RMSE, MSE, MAE, and  $R^2$  value of 0.2549; 0.5049; 0.7727; and 0.2543, respectively. The estimated values acquired in several regencies/cities have rather near, some even very close to the official statistics values.

**Keywords:** GRDP; satellite imagery; machine learning; random forest; support vector machine

**Abstrak:** Produk Domestik Regional Bruto (PDRB) merupakan salah satu indikator sosio-ekonomi yang penting. Penghitungan nilai PDRB dengan pendekatan yang melibatkan kombinasi data citra satelit dan statistik resmi dapat memberikan informasi serta pemahaman yang lebih komprehensif. Penelitian ini melakukan estimasi nilai PDRB pada tahun 2022 menggunakan data tahun 2019 hingga 2021 dengan melibatkan dua aspek, agrikultur dan non-agrikultur. Data *soil adjusted vegetation index* (SAVI), *enhanced vegetation index* (EVI), dan tutupan lahan (*land cover/LC*) digunakan sebagai aspek agrikultur, sementara data citra cahaya malam (NTL), *human settlement indeks* (HSI), luas wilayah kabupaten/kota, dan jumlah populasi per kabupaten/kota digunakan sebagai aspek non-agrikultur. Estimasi PDRB dihasilkan dengan menggunakan pendekatan *machine learning* berupa *support vector machine* (SVM) dan *random forest* (RF). Pengecekan korelasi antarvariabel menunjukkan bahwa hanya variabel luas wilayah tidak berpengaruh signifikan terhadap nilai PDRB. Model *random forest* kemudian dipilih sebagai model terbaik dengan nilai evaluasi RMSE, MSE, MAE, dan  $R^2$  berturut-turut sebesar 0.2549, 0.5049, 0.7727, dan 0.2543. Nilai estimasi yang diperoleh di beberapa kabupaten/kota cukup mendekati, bahkan ada yang sangat dekat dengan nilai statistik resmi.

**Kata kunci:** PDRB; citra satelit; machine learning; random forest; support vector machine

## INTRODUCTION

The success of economic development in a region can be seen from its economic growth [1], [2]. One of the indicators that can be used is Gross Domestic Product (GDP). GDP is a standard measure of the added value created through the production of goods and services in a certain area over a certain time period. GDP can be calculated in national and regional area (Gross Regional Domestic Area or GRDP). The higher the GDP of an area, the better the economic performance in that area [3].

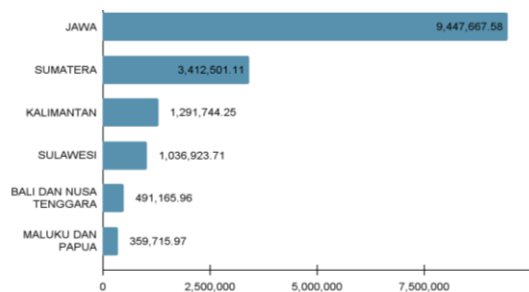


Image 1. Total GRDP Value by Island in Indonesia in 2019 (billion rupiah)

Source: BPS Indonesia [4]

In Indonesia, for example, when viewed by island, the total GRDP value for Java Island in 2019 is the highest among other islands. Java Island, which consists of 119 regencies and cities, has a GRDP value that continues to show an increasing trend. With a dense population for around 56% (154 million people) of the total population in Indonesia [5], Java Island is a place where various economic activities are concentrated. Economic growth in Java also tends to be positive and stable which reached 3.66% in 2021 [4].

Currently, GDP calculation in many countries still using traditional method. The data used comes from data collection through census or surveys. However, this method requires large cost or resources and takes quite a long time. Therefore, an alternative data sources can be used to calculate or at least, estimate GDP and GRDP value in an efficient and effective ways.

The existence of big data nowadays can give a more comprehensive understanding in many aspects. For instance, the use of satellite imagery or geospatial data allows spatial analysis to show patterns of economic growth, land use, infrastructure and social change in various regions. With this source of data, it is possible to analyze and monitor economic activities such as agriculture, fishing, and infrastructure development in certain areas.

Satellite imagery has been widely used to predict GDP values. One of the indices that are often used are NTL (*Nighttime Light*) intensity which reflects the economic prosperity of a country or region. Some numbers of studies have shown that NTL intensity has a strong correlation with economic indicators. Research conducted by [6] in China found that NTL from NPP-VIIRS imagery has a correlation with regency-level GDP with a  $R^2$  value of 0.8544. Then, research by [7] found that high-resolution VIIRS light data provides a better prediction of metropolitan statistical area (MSA) GDP than state GDP, indicating that light data may be more closely related to the urban sector than to the rural sector.

However, many other studies have shown that the use of NTL only is not enough to measure GDP accurately at the regional level. This include inability to explain areas without observable night light and distinguish certain types of economic activity, such as agriculture and forestry, which are not always associated with an increase in night light during growth.

Research by [8] stated that NTL has a rough, blooming or overglow, and saturation spatial image which can result in a loss of detail and an accurate representation of the image. Each pixel covers a larger area in the image so that small objects or features cannot be identified in detail. Thus, an over-estimation value in bright areas can increase the distribution of GRDP estimation error.

Several studies then use other data that are also considered to estimate GDP with NTL data. Research by [9] presented a model for estimating agricultural and non-agricultural economic growth at the national and subnational levels using land cover and NTL data. Research by [10] conduct a regression analysis between GDP and built-up area normalized difference built-up index (NDBI) and normalized difference vegetation index (NDVI) extracted from Landsat satellite imagery.

Unfortunately, among those previous studies, research to estimate the value of GRDP by using satellite imagery is still rarely carried out in Indonesia. Thus, the main objective of this study is to estimate GRDP value in Java Island, Indonesia with the previous variables used as well as adding other supporting data, both from satellite imagery and

official statistics data. The GRDP value to be estimated is at the regency/city level in 2022 through machine learning approach with using data from 2019 to 2021 since many regencies and cities in Java Island experienced a decrease in GRDP as a result of the Covid-19 pandemic [11].

## METHOD

### Research Locus

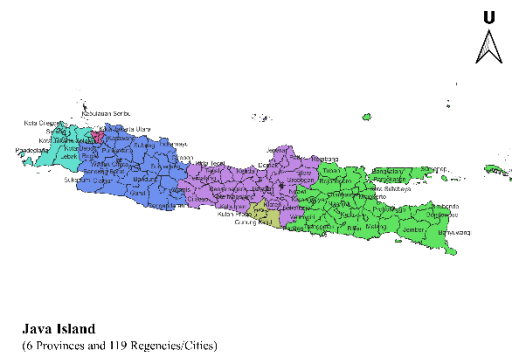


Image 2. Java Island Territory

Java Island has various characteristics and fairly good representation of Indonesia's diversity in terms of social, culture, and economy. Java Island also plays a significant role in Indonesia's economic growth since it is home to most of the provinces with high GRDP in Indonesia. The high population concentration in this area indicate various economic activities are located here. On the other hand, some areas in Java Island are also include rural areas with rich variety in agriculture, such as East Java, which has many coffee, tea and rubber plantations. Research in this area can provide insight into sustainable agriculture, natural resource management, and the challenges and opportunities in Indonesia's agricultural sector.

Table 1. Data Summary and Data Sources

Source	Spatial Resolution	Variable	Spectral Band	Description	Unit	Reference
Landsat-8 OLI/TIRS	30 m	SAVI	B4 (Red) dan B5 (NIR)	Annual data	index	[13]
Terra-MODIS (MOD13Q1)	250 m	EVI	EVI	16-day data collected over a year span	index	[14]
Terra-MODIS (MCD12Q1)	500 m	Land Cover (LC)	LC_Type1	Land classification divided into 17 classes	index	[15]
NOAA-VIIRS	750 m	Night-time Light (NTL)	avg_rad	Night light image data per month	nano watts/cm <sup>2</sup> sr	[15]
Combination of EVI and NTL data	-	HSI	-	Residential density in a certain area	index	[14]
BPS	-	Land Area (AREA)	-	Total area per regency/ city	km <sup>2</sup>	[16]
BPS	-	Population (POP)	-	Population per regency/ city	people	[16]
BPS	-	Regency/ City GRDP	-	Regency/City GRDP Value in Java Island	billion rupiahs	[4]

### Data, Data Sources, and Tools

Research by [12] stated that areas with the most illuminated by night light have a fairly high correlation to GDP. However, according to BPS data, the agricultural sector, which is generally located in areas with low light intensity, is one of the highest contributors to GDP in Indonesia [4]. This makes researchers interested in combining data related to agriculture and non-agriculture in estimating the GRDP value in Java Island. The data summary and sources in this study are presented in the following table.

This study includes 8 variables to estimate the GRDP value of regencies/cities in Java Island. Some

variables that related to agriculture aspect are *Soil Adjusted Vegetation Index* (SAVI), *Enhanced Vegetation Index* (EVI), and land cover (LC). These variables collected from satellite imagery data using *Google Earth Engine*, a free cloud based geospatial data processing and analysis platform cloud which allows users to visualize and analyze satellite imagery data for various purposes. Then, the rest of variables are related to non-agriculture aspect (except GRDP).

All variables are collected in the range period of 2019 to 2022. The data is processed using Microsoft Excel for getting numerical data (in form of image tabular results) and geospatial software QGIS for mapping visualization. The

results of the tabular data then used to construct machine learning model using R language in Google Colaboratory.

### Research methodology

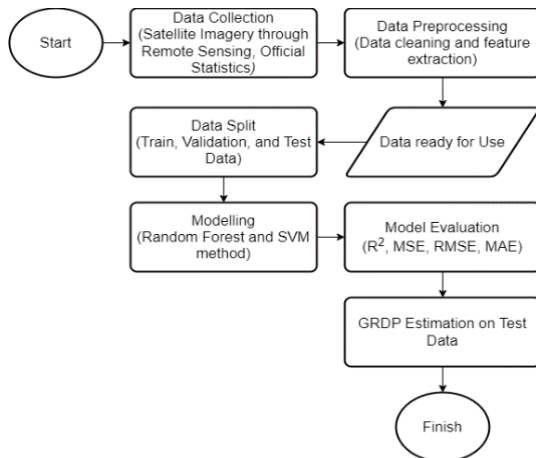


Image 3. Research Flowchart

This study uses a quantitative research method, which is used to acquire knowledge or solve problems by using numerical data. The selection of the research timeframe from 2019 to 2022 is based on the time before, during and after the Covid-19 pandemic so that it is able to cover all current conditions.

The estimation of GRDP values at the regency/city level is conducted using machine learning approach with Support Vector Machine (SVM) and Random Forest (RF) method. This method has been previously used by [17] to estimate GRDP microregional, which then obtains best model with RF method. In machine learning, the data must be separated into

two subsets, which is training data and test data. In this study, data from 2019 to 2021 will be used as training data, while data from 2022 will be used as test data.

Model evaluation is performed to obtain the best GRDP estimation model. The evaluation uses several indicators such as MSE, RMSE,  $R^2$ , and MAE. The formula for the indicators used described as follows

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2 \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (3)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}| \quad (4)$$

## RESULT AND DISCUSSION

### Correlation Test on Variables

Before the variables used to create the model, all variables are tested with the dependent variable (GRDP) to check the correlation of each variable. Only significant variables are used in the model. Based on the result in table 2, land area (AREA) does not have a significant impact with the GRDP at a significance level of 5%. Therefore, all variables are used to perform the model, except AREA.

Table 2. Correlation of Variables with GRDP, 2019-2021

Variables	2019			2020			2021		
	<i>r</i>	<i>p-value</i>	Significant	<i>r</i>	<i>p-value</i>	Significant	<i>r</i>	<i>p-value</i>	Significant
EVI	-0.573	0.000*	Yes	-0,518	0.000*	Yes	-0,578	0.000*	Yes
SAVI	-0.597	0.000*	Yes	-0,526	0.000*	Yes	-0,564	0.000*	Yes
LC	0.275	0.002*	Yes	0,189	0,039*	Yes	0,268	0,003*	Yes
NTL	0.581	0.000*	Yes	0,451	0.000*	Yes	0,535	0.000*	Yes
HSI	0.784	0.000*	Yes	0,624	0.000*	Yes	0,767	0.000*	Yes
AREA	-0.158	0.085	No	-0,145	0,115	No	-0,137	0,134	No
POP	0.431	0.000*	Yes	0,424	0.000*	Yes	0,522	0.000*	Yes

\*significant at 5% significance level

### Formation and Selection of the 2022 GRDP Estimation Model

Table 3. Model Evaluation Results for Estimated GRDP Value in 2022

Evaluation Indicator	SVM Model	RF Model
MSE	0.3303	0.2549
RMSE	0.5747	0.5049
$R^2$	0.7331	0.7727
MAE	0.2942	0.2543

After choosing the significant variables to be used in the model, all training observation of the significant variables are assigned in each method. Each training model then used for the test dataset to be evaluated using some indicators (MSE, RMSE,  $R^2$ , and MAE). The best model is expected to have small MSE, RMSE, and MAE values followed by high  $R^2$  score.

Evaluation results of the two models in table 3, it appears that the MSE, RMSE, and MAE values for Random Forest model is smaller than the SVM model. Looking at the  $R^2$  value, Random Forest model also has higher  $R^2$  value than the SVM model. Thus, the best model for estimating the GRDP value of regencies/cities in 2022 is the model built using the Random Forest method. The model chosen is in line with the model research results by [17].

### Results of Estimating GRDP at the Regency/City Level in 2022

After the best model is selected based on the evaluation indicators, GRDP value estimation is performed at the regency/city level in 2022. The estimated results have been inversed (previously transformed during modeling) in order to obtain an exact estimated value and match with the official statistical data.

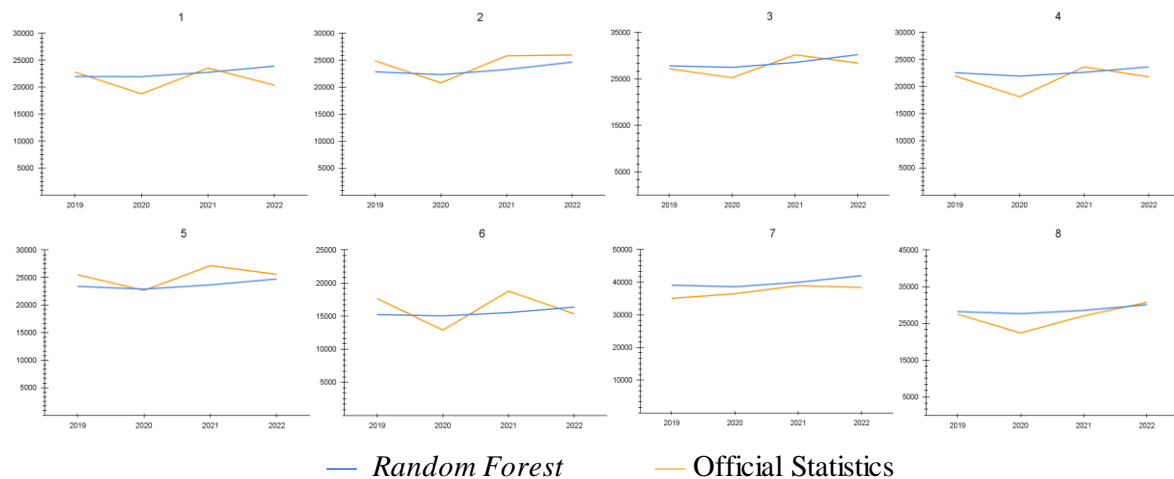


Image 4. Estimated GRDP Value Graph of Some Regencies/Cities in Java Island 2019-2022 (billion rupiah)

Where: 1. Ciamis Regency, 2. Cimahi City, 3. Klaten Regency, 4. Lumajang Regency, 5. Probolinggo Regency, 6. Banjarnegara Regency, 7. Garut Regency, 8. Jombang Regency

The following image presents a graph of GRDP estimation result in some regencies compared to the official GRDP statistics produced by BPS. Some regions have very close estimation result, such as Jombang, Banjarnegara, Probolinggo, and Cimahi. The model also used to estimate the GRDP values from 2019 to 2021. The results obtained are quite satisfactory because there are several regencies/cities that have close estimated value to the actual, such as Garut and Probolinggo.

close to the real GRDP value in official statistical data. Satellite imagery can provide many information through vary of indices. Adding more variables in creating GRDP estimation model can be conducted in future research since GRDP is complex and includes many sectors. The same applies to the method used. Thus, the GRDP estimation will be more accurate.

## BIBLIOGRAPHY

- [1] R. Hawari and F. Kartiasih, "Kajian Aktivitas Ekonomi Luar Negeri Indonesia Terhadap Pertumbuhan Ekonomi Indonesia Periode 1998-2014," *MEDIA Stat.*, vol. 9, no. 2, pp. 119–132, Dec. 2016, doi: 10.14710/medstat.9.2. 119-132.
- [2] I. P. Pertiwi, N. D. Camalia, R. Rega, and F. Kartiasih, "Mencermati Pengaruh Utang Dan Variabel Makroekonomi Terhadap Sustainability Fiskal Di Indonesia," *J. Ilm. Ekon. Dan Bisnis*, vol. 20, no. 2, pp. 185–

## CONCLUSION

This research succeeded in creating model to estimate GRDP values at the regency or city level in 2022. Through the machine learning approach, the best model chosen is the Random Forest model with the highest  $R^2$  and lowest MSE, RMSE, and MAE. Some estimated values are fairly close, some even very

- 193, Oct. 2023, doi: 10.31849/jieb.v20i2.12259.
- [3] A. Kusumasari and F. Kartiasih, "Agglomerasi Industri dan Pengaruhnya terhadap Pertumbuhan Ekonomi Jawa Barat 2010-2014," *J. Apl. Stat. Komputasi Stat.*, vol. 9, no. 2, Art. no. 2, Dec. 2017, doi: 10.34123/jurnalaks.v9i2.143.
- [4] Badan Pusat Statistik, "Produk Domestik Regional Bruto Kabupaten Kota di Indonesia 2017-2021," *07100.2203*, p. xx+180, 2022.
- [5] B. P. S. Indonesia, "Statistik Indonesia 2021." Accessed: Mar. 23, 2024. [Online].
- [6] X. Li, H. Xu, X. Chen, and C. Li, "Potential of NPP-VIIRS Nighttime Light Imagery for Modeling the Regional Economy of China," *Remote Sens.*, vol. 5, no. 6, pp. 3057–3081, Jun. 2013, doi: 10.3390/rs5063057.
- [7] X. Chen and W. D. Nordhaus, "VIIRS Nighttime Lights in the Estimation of Cross-Sectional and Time-Series GDP," *Remote Sens.*, vol. 11, no. 9, p. 1057, May 2019, doi: 10.3390/rs11091057.
- [8] X. Cao, Y. Hu, X. Zhu, F. Shi, L. Zhuo, and J. Chen, "A simple self-adjusting model for correcting the blooming effects in DMSP-OLS nighttime light images," *Remote Sens. Environ.*, vol. 224, pp. 401–411, Apr. 2019, doi: 10.1016/j.rse.2019.02.019.
- [9] S. Keola, M. Andersson, and O. Hall, "Monitoring Economic Development from Space: Using Nighttime Light and Land Cover Data to Measure Economic Growth," *World Dev.*, vol. 66, pp. 322–334, Feb. 2015, doi: 10.1016/j.worlddev. 2014.08.017.
- [10] K. Faisal, A. Shaker, and S. Habbani, "Modeling the Relationship between the Gross Domestic Product and Built-Up Area Using Remote Sensing and GIS Data: A Case Study of Seven Major Cities in Canada," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 3, p. 23, Feb. 2016, doi: 10.3390/ijgi5030023.
- [11] N. S. Harum, M. Aini, M. A. Risxi, and F. Kartiasih, "Pengaruh Sosial Ekonomi dan Kesehatan terhadap Pengeluaran Konsumsi Pangan Rumah Tangga Provinsi Jawa Tengah Tahun 2020," *Semin. Nas. Off. Stat.*, vol. 2023, no. 1, Art. no. 1, Oct. 2023, doi: 10.34123/semnasoffstat.v2023i1.1919.
- [12] C. D. Elvidge, M. Zhizhin, T. Ghosh, F.-C. Hsu, and J. Taneja, "Annual Time Series of Global VIIRS Nighttime Lights Derived from Monthly Averages: 2012 to 2019," *Remote Sens.*, vol. 13, no. 5, p. 922, Mar. 2021, doi: 10.3390/rs13050922.
- [13] M. Hu and B. Xia, "A significant increase in the normalized difference vegetation index during the rapid economic development in the Pearl River Delta of China," *Land Degrad. Dev.*, vol. 30, no. 4, pp. 359–370, 2019, doi: 10.1002/ldr.3221.
- [14] W. Yue, J. Gao, and X. Yang, "Estimation of Gross Domestic Product Using Multi-Sensor Remote Sensing Data: A Case Study in Zhejiang Province, East China," *Remote Sens.*, vol. 6, no. 8, pp. 7260–7275, Aug. 2014, doi: 10.3390/rs6087260.
- [15] J. Sun, L. Di, Z. Sun, J. Wang, and Y. Wu, "Estimation of GDP Using Deep Learning With NPP-VIIRS Imagery and Land Cover Data at the County Level in CONUS," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 1400–1415, 2020, doi: 10.1109/JSTARS.2020.2983331.
- [16] G. C. McCord and M. Rodriguez-Heredia, "Nightlights and Subnational Economic Activity: Estimating Departmental GDP in Paraguay," *Remote Sens.*, vol. 14, no. 5, p. 1150, Feb. 2022, doi: 10.3390/rs14051150.
- [17] R. Zulkarnain, "Estimasi PDB Mikroregional: Studi Kasus di Pulau Jawa," *Semin. Nas. Off. Stat.*, vol. 2022, no. 1, pp. 885–894, Nov. 2022, doi: 10.34123/semnasoffstat.v2022i1.1257.