



## ***Estimating Indonesia's Tax Gap: Analysis of Discrepancies in Official GRDP Based on Machine Learning-Based Nighttime Lights***

### **Estimasi Tax Gap Indonesia: Analisis Diskrepansi PDRB Resmi Berdasarkan *Nighttime Lights* Berbasis Machine Learning**

**Fachri Husein Harahap**

Program Studi Manajemen Aset Publik, Politeknik Keuangan Negara STAN, Indonesia

E-Mail: [fachrihusein878@gmail.com](mailto:fachrihusein878@gmail.com)

Received Feb 10th 2026; Revised Mar 17th 2026; Accepted Mar 31th 2026; Available Online Apr 26th 2026

Corresponding Author: Fachri Husein Harahap

Copyright ©2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

#### **Abstract**

*This study proposes a new methodology for estimating 'True PDRB' and 'shadow economy' at the provincial level in Indonesia using a latent variable framework. The researchers treat real GRDP as a latent variable estimated from two imperfect independent measurements: (1) official GRDP from Statistics Indonesia, which is modeled to have systematic underreporting error, and (2) geospatial GRDP, which is predicted from multi-source proxy data (NTL, NO<sub>2</sub>, informality, consumption) using the XGBoost machine learning model. To avoid circular reasoning, the XGBoost model was trained on an adaptive weighted target that adjusted the weight of the official GRDP based on provincial levels of informality. The model showed strong predictive power ( $R^2 = 0.8256$ ) on the 2023 temporal test data. GMM estimates reveal a national aggregate shadow economy of 6.76% of official GRDP, with a national tax gap estimate of IDR 16.17 trillion (7.03% of potential). The findings show extreme heterogeneity: agriculture-based provinces such as Central Kalimantan (46.9%) and West Papua (37.8%) have the highest shadow-economy rates. Conversely, the largest nominal tax gaps are concentrated in provinces with very large GRDP, namely Papua (Rp 34.5 trillion) and East Java (Rp 8.9 trillion). These findings produce a tax gap map that the DGT can follow up on to formulate data-based extensification strategies.*

**Keywords:** Machine Learning, Nighttime Light, Shadow Economy, Tax Gap, XGBoost.

#### **Abstrak**

Penelitian ini mengusulkan metodologi baru untuk mengestimasi 'True PDRB' dan 'shadow economy' di tingkat provinsi Indonesia menggunakan *latent variable framework* [7]. Peneliti menggunakan PDRB riil sebagai variabel laten yang diestimasi dari dua pengukuran independen yang tidak sempurna: PDRB resmi BPS, yang dimodelkan memiliki *error underreporting* sistematis, dan PDRB-geospatial, yang diprediksi dari *multi-source proxy data* (NTL, NO<sub>2</sub>, informalitas, konsumsi) menggunakan model *machine learning* XGBoost. Untuk menghindari penalaran melingkar, model XGBoost dilatih pada target tertimbang adaptif yang menyesuaikan bobot PDRB resmi berdasarkan tingkat informalitas provinsi. Model menunjukkan daya prediksi kuat ( $R^2 = 0.8256$ ) pada data uji temporal 2023. Estimasi GMM mengungkap *shadow economy* agregat nasional sebesar 6,76% dari PDRB resmi, dengan estimasi tax gap nasional Rp 16,17 triliun (7,03% dari potensi). Temuan menunjukkan heterogenitas ekstrem: provinsi berbasis agrikultur seperti Kalimantan Tengah (46,9%) dan Papua Barat (37,8%) memiliki persentase *shadow economy* tertinggi. Sebaliknya, tax gap nominal terbesar terkonsentrasi di provinsi dengan PDRB sangat besar, yaitu Papua (Rp 34,5 T) dan Jawa Timur (Rp 8,9 T). Temuan ini menghasilkan pemetaan tax gap yang dapat ditindaklanjuti bagi DJP untuk merumuskan strategi ekstensifikasi berbasis data.

**Kata Kunci:** Machine Learning, Nighttime Light, Shadow Economy, Tax Gap, XGBoost.

#### **1. PENDAHULUAN**

Tantangan struktural utama Indonesia adalah optimalisasi penerimaan pajak. Tax ratio Indonesia pada tahun 2023 tercatat hanya 10,21% dari PDB [25]. Angka ini menunjukkan kesenjangan signifikan dibandingkan dengan rata-rata Asia-Pasifik (19,8%) [28] dan rata-rata OECD (34,0%) [27]. Salah satu kontributor utama rendahnya rasio ini adalah besarnya *shadow economy* [23]. Medina dan Schneider [12] mengestimasi *shadow economy* Indonesia berada pada kisaran 21,5% PDB.



*Shadow economy* sebagai aktivitas ekonomi legal namun tidak tercatat dan tidak dilaporkan untuk menghindari pajak [12], secara langsung mengerosi basis pajak nasional dan mengurangi optimalisasi penerimaan pajak [9]. Aktivitas ekonomi ini didominasi oleh sektor informal, tidak tercermin dalam statistik PDB, sehingga menyulitkan pemerintah dalam perumusan kebijakan fiskal yang akurat [14].

Metode konvensional untuk mengukur *shadow economy*, seperti model MIMIC [10] dan *currency demand approach* [9], memiliki keterbatasan fundamental. Metode tersebut hanya menghasilkan estimasi agregat nasional, tidak granular per provinsi, sehingga tidak dapat ditindaklanjuti untuk pengawasan pajak regional. Lebih penting lagi, metode ini mengasumsikan statistik resmi sebagai "kebenaran", padahal di negara berkembang, PDB/PDRB resmi sering mengandung kesalahan pengukuran sistematis dan *underreporting* [11].

Kemajuan *remote sensing* menawarkan pendekatan baru. Data citra satelit malam hari (NTL) telah terbukti sebagai proksi kuat untuk aktivitas ekonomi [3]; [6]. Pada Indonesia, NTL terbukti berkorelasi kuat dengan PDRB ( $R^2 = 0.87$ ) [2] dan berfungsi sebagai pengukuran independen. Penelitian ini mengadopsi *latent variable framework* [7] yang memperlakukan PDRB sejati (*True PDRB*) sebagai variabel laten. Variabel ini diestimasi dari dua pengukuran yang tidak sempurna, yaitu PDRB resmi BPS, yang dimodelkan mengandung *systematic underreporting*; dan PDRB-*geospatial*, yang diprediksi dari data satelit.

Untuk PDRB-*geospatial*, kami mengintegrasikan multiple independent *geospatial proxies* (polusi  $\text{NO}_2$ , area terbangun, dan NTL) yang diekstrak via *Google Earth Engine* [5]. Model *machine learning* (XGBoost) dilatih untuk memprediksi PDRB-*geospatial*. Menggunakan *Method of Moments*, model ini mengestimasi *True PDRB* (laten) dan tingkat *underreporting* dalam data resmi. *Shadow economy* dihitung sebagai selisih antara *True PDRB* estimasi dengan PDRB resmi.

Berbagai penelitian terdahulu membuktikan korelasi kuat antara data citra satelit malam hari (NTL) dengan aktivitas ekonomi [6];[2]. Di Indonesia, BPS [20] juga telah memanfaatkan data ini untuk memprediksi kemiskinan dan dampak kebijakan. Namun, penelitian-penelitian tersebut berhenti pada pembuktian korelasi atau prediksi indikator non-fiskal. Paper ini mengambil langkah inovatif lebih jauh. Kami tidak hanya memprediksi PDRB, tetapi mengisi celah penelitian utama dengan mengombinasikan data NTL dengan data proksi sosial-ekonomi (pekerja informal dan konsumsi) menggunakan *machine learning*; dan menggunakan residual (*error*) dari model prediksi tersebut sebagai estimasi kuantitatif untuk *shadow economy*.

Estimasi *shadow economy* tersebut tidak sekadar dibiarkan sebagai angka statistik, melainkan direkonstruksi menjadi PDRB terkoreksi yang menjadi basis perhitungan *tax gap* provinsi—area yang belum pernah dijamah dengan metodologi lintas disiplin ini di Indonesia. Lompatan inovatif ini melampaui studi-studi terdahulu yang hanya berfokus pada prediksi indikator ekonomi makro. Penelitian ini berani melangkah lebih jauh ke ranah kebijakan fiskal dengan menggunakan *discrepancy* ekonomi sebagai proksi untuk memetakan wilayah pengawasan pajak prioritas. Dengan demikian, penelitian ini menawarkan solusi nyata bagi otoritas pajak dalam mengoptimalkan penerimaan negara melalui pendekatan berbasis data *geospatial*.

## 2. TINJAUAN LITERATUR

### 2.1 Konsep Tax Gap

#### 2.1.1 Definisi dan Komponen Tax Gap

Tax gap (kesenjangan pajak) adalah perbedaan antara pajak terutang teoretis dan pajak yang faktanya dikumpulkan [19]. Tax gap berfungsi sebagai indikator kinerja utama otoritas pajak. *Internal Revenue Service* (IRS) membaginya menjadi tiga komponen utama [30], yaitu *non-filing gap* (WP tidak lapor SPT), *Underreporting gap* (WP lapor lebih rendah), dan *underpayment gap* (WP lapor tapi tidak bayar). Komponen *underreporting* adalah yang terbesar dan terkait erat dengan *shadow economy* [12].

#### 2.1.2 Metode Pengukuran Tax Gap

Terdapat dua pendekatan utama pengukuran menurut IMF. [24]; Schneider & Enste. [14]: *Top-Down* (makro), yang mengestimasi basis pajak potensial dari data agregat, dan *Bottom-Up* (mikro), yang mengekstrapolasi data audit individual [30]. Metode konvensional *Top-Down* umumnya bergantung pada statistik resmi sebagai *baseline* [10]; [24]. Penelitian ini menggunakan pendekatan *top-down* yang dimodifikasi, di mana *True PDRB* diestimasi terlebih dahulu menggunakan *latent variable framework* sebelum potensi pajak dihitung.

### 2.2 Shadow economy

#### 2.2.1 Definisi dan Karakteristik

*Shadow economy* dapat didefinisikan sebagai aktivitas produksi legal yang sengaja disembunyikan dari otoritas publik (disebut *underground production* atau *informal sector production*) untuk menghindari pajak atau regulasi [16]. Karakteristik utamanya di negara berkembang meliputi dominasi sektor informal/UMKM [12], tingginya transaksi tunai [9], dan berkembangnya ekonomi digital yang sulit dilacak [24].

## 2.2.2 Dampak *Shadow economy* terhadap Erosi Basis Pajak

Dampak *shadow economy* bersifat merusak sebab secara langsung mengurangi basis pajak PPh dan PPN, menciptakan distorsi kompetisi bagi sektor formal [18] dan dapat menurunkan moral pajak [1]. Selain itu, fenomena ini menyebabkan angka statistik ekonomi resmi menjadi tidak konsisten (*unreliable*), yang berpotensi menimbulkan ketidaktepatan kebijakan.

## 2.3 *Latent variable framework* untuk Estimasi *True PDRB*

### 2.3.1 Konsep Dasar

Hu dan Yao. [7] dalam studi mereka mengembangkan kerangka kerja yang memperlakukan *True GDP* sebagai variabel laten (tidak terobservasi) yang diestimasi dari beberapa pengukuran yang tidak sempurna. Kerangka kerja ini berbeda secara fundamental dari pendekatan konvensional karena tidak mengasumsikan bahwa statistik resmi adalah "kebenaran" [7]. Sebaliknya, PDRB resmi diperlakukan hanya sebagai satu pengukuran tidak sempurna yang mengandung kesalahan sistematis (yaitu, *underreporting*) dan kesalahan acak.

Kerangka kerja ini memanfaatkan sumber pengukuran alternatif (berbasis data *geospatial*) yang independen dari sistem statistik resmi. Pengukuran kedua ini berfungsi sebagai observasi independen dari *True PDRB* yang sama. Dengan asumsi utama bahwa kesalahan pengukuran dalam kedua sumber data bersifat *conditionally independent* (tergantung pada *True PDRB*), model ini dapat secara struktural mengidentifikasi besaran *True PDRB* beserta tingkat *underreporting* dalam statistik resmi [7].

### 2.3.2 *Framework Matematis*

Misalkan  $\theta_i$  adalah *True PDRB* provinsi  $i$  (variabel laten yang ingin diestimasi). Kita mengamati dua pengukuran yang tidak sempurna [7]:

#### 1. Official PDRB ( $Y_1$ )

$Y_1 = \theta_i (1 - \delta_i) + \epsilon_{1i}$ , di mana  $\delta_i$  adalah *systematic underreporting rate* ( $0 \leq \delta_i < 1$ ) dan  $\epsilon_{1i}$  adalah random measurement error dengan ekspektasi nol ( $E[\epsilon_{1i}] = 0$ ).

#### 2. *Geospatial*-predicted PDRB ( $Y_2$ )

$Y_2 = \theta_i + \epsilon_{2i}$  di mana  $\epsilon_{2i}$  adalah random measurement error dengan ekspektasi nol ( $E[\epsilon_{2i}] = 0$ ).

Asumsi identifikasi kuncinya adalah  $Cov(\epsilon_{1i}, \epsilon_{2i} | \theta_i) = 0$ . Secara intuitif, ini berarti bahwa kesalahan dalam statistik resmi (misalnya, masalah birokrasi, cakupan survei, penghindaran pajak) tidak berkorelasi dengan kesalahan dalam pengukuran berbasis satelit (misalnya, tutupan awan, masalah sensor, kesalahan pemodelan) [7].

Untuk memodelkan *underreporting rate* itu sendiri, digunakan persamaan  $\delta_i = \Lambda(\beta_0 + \beta_1 \times \%Informal + \beta_2 \times \%Agriculture)$ , di mana  $\Lambda()$  adalah fungsi logistik yang memastikan nilai  $\theta_i$  tetap berada dalam rentang  $[0, 1]$ . Hipotesis yang diajukan adalah bahwa provinsi dengan tingkat informalitas dan dominasi sektor pertanian yang lebih tinggi akan memiliki *underreporting rate* yang lebih tinggi.

### 2.3.3 *Estimation Strategy: Method of Moments*

Mengikuti Hu dan Yao.[7], terdapat dua strategi estimasi yang dapat digunakan: Bayesian MCMC penuh atau *Method of Moments* (MoM). Untuk alasan kelayakan dan efisiensi komputasi, penelitian ini mengadopsi *Method of Moments*, yang terbukti tetap konsisten dan *asymptotically normal*. Estimasi parameter  $\{\theta_i, \beta_0, \beta_1, \beta_2\}$  dilakukan dengan meminimalkan jumlah kuadrat tertimbang dari *moment conditions*. Kondisi-kondisi tersebut adalah  $E[Y_{1i} - \theta_i(1 - \delta_i)] = 0$ ,  $E[Y_{2i} - \theta_i] = 0$ , dan kondisi instrumen  $E[Y_{1i} - \theta_i(1 - \delta_i) \times Z_i] = 0$ . Dalam kondisi terakhir,  $Z_i$  merupakan vektor karakteristik provinsi (misalnya IPM, persentase urban, dll.) yang berfungsi sebagai instrumen. Kuantifikasi ketidakpastian untuk seluruh estimasi parameter ini dilakukan melalui prosedur *bootstrap*.

## 2.4 *Geospatial Data sebagai Independent Measurement*

### 2.4.1 *Nighttime Lights sebagai Proxy Aktivitas Ekonomi*

Data *Nighttime Lights* (NTL) adalah proksi PDRB yang mapan [6]; [3]; [7]. Keunggulan utamanya adalah objektivitas, cakupan komprehensif, dan kemampuannya menangkap aktivitas ekonomi terlepas dari status formalitasnya. Validitas NTL (khususnya sensor VIIRS) di Indonesia telah dikonfirmasi [2]; [4].

### 2.4.2 *Multi-Source Geospatial Proxies*

Untuk memperkuat pengukuran  $Y_2$ , proksi NTL diintegrasikan dengan data polusi udara  $\text{NO}_2$  dari satelit Sentinel-5P.  $\text{NO}_2$  sangat berkorelasi dengan aktivitas industri dan transportasi [8], sehingga berfungsi sebagai proksi untuk aktivitas padat energi.

### 2.4.3 Google Earth Engine untuk Efficient Data Processing

Google Earth Engine (GEE) digunakan sebagai platform *cloud computing* untuk analisis *geospatial* berskala planet [5]. GEE memungkinkan ekstraksi, pemfilteran *cloud masking*, dan agregasi data NTL dan NO<sub>2</sub> secara efisien untuk 34 provinsi, serta memastikan replikabilitas penuh.

### 2.5 Machine learning untuk Prediksi Ekonomi

*Machine learning* (ML), khususnya metode *ensemble* seperti *XGBoost*, menawarkan keunggulan dalam memodelkan hubungan non-linear yang kompleks [13]. Apabila dibandingkan dengan regresi linear, *XGBoost* dapat secara otomatis menangkap non-linearitas dan interaksi, menangani multikolinearitas dengan lebih baik melalui regularisasi, melakukan pemilihan fitur implisit, dan memberikan prediksi yang kuat dengan data yang kompleks [17].

Pada kerangka kerja penelitian ini, ML digunakan untuk memprediksi hubungan antara proksi *geospatial* (X) dan aktivitas ekonomi riil. Secara kritis, model tidak dilatih pada PDRB resmi saja (yang akan menciptakan penalaran melingkar), tetapi pada kombinasi tertimbang adaptif dari PDRB resmi dan PDRB berbasis *lights* eksternal, dengan bobot yang mencerminkan kualitas pengukuran yang diharapkan.

### 2.6 Gap Penelitian

Berdasarkan tinjauan teori, penelitian ini mengisi tiga gap utama. Pertama, Gap Metodologi, dengan menjadi yang pertama menerapkan *latent variable framework* [7] di Indonesia, alih-alih metode konvensional [10]; [9] yang mengasumsikan statistik resmi sebagai baseline. Kedua, Gap Data, dengan mengintegrasikan multi-source *geospatial* data (NTL + NO<sub>2</sub>) via GEE untuk estimasi struktural, melampaui studi korelasi NTL sebelumnya [2]. Ketiga, Gap Aplikasi, dengan menyediakan estimasi tax gap provinsi yang *granular* dan *replicable* untuk Direktorat Jenderal Pajak.

## 3. METODOLOGI PENELITIAN

### 3.1 The Length of the Paper

Penelitian ini menggunakan pendekatan kuantitatif dengan analisis data sekunder. Desain penelitian adalah pemodelan ekonometrika struktural menggunakan *latent variable framework* [7], dengan *machine learning* sebagai komponen untuk meningkatkan pengukuran berbasis *geospatial*. Unit analisis adalah provinsi-tahun (*province-year observations*) dalam struktur data panel.

Populasi menggunakan seluruh 34 provinsi di Indonesia (per struktur administratif terkini) dan sampel sensus (seluruh populasi dianalisis). Data panel mencakup 34 provinsi untuk periode 2019-2023 (5 tahun), menghasilkan 170 observasi provinsi-tahun. Justifikasi periode data NO<sub>2</sub> dari Sentinel-5P tersedia secara reliabel mulai 2019. Untuk konsistensi, semua variabel lain juga dibatasi pada periode 2019-2023.

### 3.2 Sumber Data dan Variabel

#### 3.2.1 Variabel Dependen

Y<sub>1</sub>: PDRB resmi (*Official PDRB*). Bersumber dari Badan Pusat Statistik [20]. Produk Domestik Regional Bruto Atas Dasar Harga Berlaku (ADHB) dengan unit Miliar Rupiah

Y<sub>2</sub>: PDRB *Geospatial*-Predicted. Konstruksi: Diprediksi dari model *XGBoost* dengan input data *geospatial multi-source* (dijelaskan di Bagian 3.5.2)

#### 3.2.2 Variabel Independen (*Geospatial Proxies*)

Intensitas *Nighttime Lights*, *dataset* menggunakan NOAA VIIRS DNB Monthly Composites bersumber dari Google Earth Engine. Metrik menggunakan rata-rata radiansi (*nanoWatts/cm<sup>2</sup>/sr*) dan agregasi menggunakan rata-rata tahunan per provinsi. Kepadatan kolom NO<sub>2</sub> (Polusi Udara) *dataset* menggunakan Sentinel-5P TROPOMI bersumber dari Google Earth Engine. Metrik menggunakan kepadatan kolom NO<sub>2</sub> troposferik (*mol/m<sup>2</sup>*) dan agregasi menggunakan median tahunan per provinsi. Persentase Pekerja Informal, bersumber dari Survei Angkatan Kerja Nasional (Sakernas), Badan Pusat Statistik 2023. Menggunakan fungsi variabel instrumen untuk estimasi *underreporting rate*, dan proksi validasi.

Konsumsi rumah tangga per kapita bersumber dari Survei Sosial Ekonomi Nasional (Susenas) Badan Pusat Statistik 2023 dan menggunakan fungsi prediktor tambahan dalam model ML. Indeks Pembangunan Manusia (IPM), bersumber dari Badan Pusat Statistik 2023 dan menggunakan fungsi variabel kontrol. Jumlah penduduk bersumber dari Badan Pusat Statistik (proyeksi penduduk) tahun 2023 dan menggunakan fungsi kontrol untuk efek ukuran. Persentase Sektor Pertanian, bersumber dari Struktur PDRB menurut lapangan usaha, Badan Pusat Statistik tahun 2023 dan menggunakan fungsi variabel instrumen untuk *underreporting rate*.

#### 3.2.3 Data untuk Perhitungan Tax Gap

Bersumber dari Realisasi Penerimaan Pajak per Provinsi [23]. Komponen yang digunakan adalah PPh ditambah PPN yang direalisasikan per Kantor Wilayah DJP. Menggunakan unit Miliar Rupiah

### 3.3 Teknik Pengumpulan Data

#### 3.3.1 Data BPS (Official Statistics)

Mengumpulkan data dari portal BPS (<https://www.bps.go.id/>) dan menggunakan data PDRB per provinsi, data Sakernas, dan data Susenas

#### 3.3.2 Data Geospatial via Google Earth Engine

Mengumpulkan data melalui Platform *Google Earth Engine Code Editor* dengan tahapan alur kerja; Impor batas administratif (*Shapefile* provinsi), Muat *dataset* satelit (VIIRS, Sentinel-5P, *ESA World Cover*), Filter temporal (2019-2023) dan filter kualitas (misalnya, QA value untuk NO<sub>2</sub>, penghapusan *stray light* untuk VIIRS), Agregasi spasial (rata-rata/median untuk setiap geometri provinsi), Agregasi temporal (rata-rata tahunan atau tingkat pertumbuhan), Ekspor hasil sebagai CSV ke Google Drive.

#### 3.3.3 Data DJP

Mengumpulkan data dari Laporan Tahunan dari situs web DJP atau Laporan Keuangan Pemerintah Pusat (LKPP). Namun, dengan catatan bahwa apabila data provinsi tidak tersedia, penulis menggunakan proksi realisasi per Kantor Wilayah DJP sebagai aproksimasi.

### 3.4 Metode Analisis Data

#### 3.4.1 Data Preprocessing

Missing Values, Interpolasi linier untuk nilai yang hilang (jika < 10%). Outlier, deteksi menggunakan metode IQR dan *Winsorization* (cap pada persentil 1 dan 99). Transformasi log untuk variabel dengan *skewness* tinggi (NTL, PDRB, Penduduk, Konsumsi). Standardisasi (z-score) sebelum pemodelan ML. Struktur Panel, Gabungkan semua *dataset* berdasarkan (Province ID, Year).

#### 3.4.2 Stage 1: Machine learning Model untuk Prediksi Geospatial-PDRB

Bertujuan memprediksi "perkiraan terbaik" PDRB riil dari data *geospatial*, menghindari penalaran melingkar. Selanjutnya, Strategi Target Tertimbang Adaptif (*Adaptive Weighted Target*) dengan target untuk pelatihan adalah kombinasi tertimbang, ditunjukkan pada Persamaan 1.

$$Y_{target_{i,t}} = w_{i,t} \times Y1_{i,t} + (1 - w_{i,t}) \times Y2_{external_{i,t}} \quad (1)$$

Dimana:  $Y1_{i,t}$  = PDRB resmi;  $Y2_{external_{i,t}}$  = PDRB berbasis *lights* menggunakan elastisitas eksternal;  $w_{i,t}$  = Bobot adaptif berdasarkan kualitas data yang diharapkan. Provinsi dengan informalitas tinggi mendapat bobot lebih rendah pada  $Y1$  (karena PDRB resmi kurang reliabel) dan bobot lebih tinggi pada  $Y2_{external}$ .

*Feature Set (X) Input* untuk model *XGBoost* meliputi: *ln\_NTL*, *ln\_NO2*, *builtup\_growth*, *pct\_informal*, *ln\_konsumsi*, *IPM*, *ln\_penduduk*, *pct\_pertanian*, dan *dummy pulau*. Model Spesifikasi *XGBoost* Regressor, *hyperparameters* optimal ditentukan melalui *GridSearchCV*, menghasilkan parameter terbaik: {'learning\_rate': 0.1, 'max\_depth': 3, 'n\_estimators': 300, 'subsample': 1.0}. *Training-Validation-Test Split* Digunakan pemisahan temporal (*temporal split*) untuk menghindari kebocoran data: Data Latih (2019-2022) dan Data Uji (2023). *Evaluation Metrics* R<sup>2</sup>, RMSE, MAE, MAPE. *Output* dengan rumusan  $Y_2 = PDRB_{Geospatial}$ .

Analisis *Feature Importance* menggunakan *Gain-based* dan *SHAP values* untuk interpretabilitas. Kemudian, nilai prediksi dari model *XGBoost* yang telah dilatih, ditunjukkan pada Persamaan 2.

$$Y2_{geospatial} = model_{xgb}.predict(X_{all}) \quad (2)$$

#### 3.4.3 Stage 2: Latent Variable Estimation via Method of Moments

*Framework Setup* Diberikan dua pengukuran  $Y1_{i,t}$  (PDRB resmi) dan  $Y_2$  (PDRB *geospatial*). Estimasi:  $\theta_{i,t}$  (*True PDRB*) dan  $\delta_{i,t}$  (*Underreporting rate*). Persamaan struktural ditunjukkan pada persamaan 3 sampai dengan 5.

$$Y1_{i,t} = \theta_{i,t} \times (1 - \delta_{i,t}) + \epsilon_{1i,t} \text{ (Official Measurement)} \quad (3)$$

$$Y2_{i,t} = \theta_{i,t} + \epsilon_{2i,t} \text{ (Geospatial Measurement)} \quad (4)$$

$$\delta_{i,t} = \Lambda(\beta_0 + \beta_1 \times PctInformal_{i,t} + \beta_2 \times PctPertanian_{i,t} + \beta_3 \times IPM_{i,t}) \quad (5)$$

Kemudian, menggunakan *Moment Conditions* (GMM)  $\Theta = \{\theta_{i,t}; [citestart]\beta_0, \beta_1, \beta_2, \beta_3\}$

Diestimasi dengan mencocokkan *moment conditions*.  $E[Y_{1it} - \theta_{(i,t)}(1 - \delta_{i,t})] = 0$ ;  $E[Y_{2it} - \theta_{i,t}] = 0$ ;  $E[(Y_{1it} - \theta_{i,t}(1 - \delta_{i,t})) \times Z_{i,t}] = 0$  (Orthogonality);  $Z_{i,t}$  (instrumen) = [1, PctInformal, PctPertanian, IPM, ln\_Penduduk].

Implementasi (Python) Menggunakan `scipy.optimize.minimize` untuk menemukan parameter  $\theta$  yang meminimalkan jumlah kuadrat dari *moment conditions*. Output  $\hat{\theta}_{i,t}$  (Estimasi True PDRB),  $\delta_{i,t}$  (Estimasi *underreporting rate*), dan  $\hat{\beta}$  (Estimasi koefisien model *underreporting*).

### 3.4.4 Kuantifikasi Ketidakpastian: Bootstrap

Prosedur *Resample* data B = 1000 kali (distratifikasi berdasarkan provinsi). Menjalankan ulang seluruh prosedur estimasi (*Stage 1 + Stage 2*) untuk setiap sampel *bootstrap* dan menghitung *confidence intervals* 95% menggunakan metode persentil dari 1000 hasil estimasi.

### 3.4.5 Perhitungan Shadow economy

Secara matematis digambarkan dengan rumusan

$$SE_{i,t} = \hat{\theta}_{i,t} - Y1_{i,t}, \%SE_{i,t} = (SE_{i,t}/Y1_{i,t}) \times 100\%. \quad (6)$$

Validasi dilakukan dengan membandingkan estimasi agregat nasional dengan literatur (misalnya: Medina & Schneider [12] dengan kisaran sekitar 21,5%).

### 3.4.6 Estimasi Tax Gap

Menghitung Rasio Pajak-PDRB Historis  $\bar{\tau} = (\text{Total Realisasi Pajak Nasional}) / (\text{Total PDRB Nasional})$ , rata-rata 2019-2023. Mengestimasi Potensi Pajak  $Tax_{potential_{i,t}} = \theta_{i,t} \times \bar{\tau}$ . Menghitung Tax Gap dengan rumusan  $Tax_{Gap_{i,t}} = Tax_{potential_{i,t}} - Tax_{actual_{i,t}}$ . Menghitung Total Tax Gap Nasional.

$$Total_{Tax_{Gap}} = \sum_{i,t} Tax_{Gap_{i,t}} / T \quad (7)$$

### 3.4.7 Analisis Validasi

Konsistensi internal dengan memeriksa apakah (*underreporting rate*) berkorelasi positif dengan persentase pekerja informal. Diharapkan  $r > 0,4$  dan  $p < 0,01$ . Selanjutnya, dilakukan validasi eksternal dengan membandingkan %SE nasional dengan Medina & Schneider [12]. Memeriksa apakah provinsi dengan informalitas tinggi memang memiliki  $\delta$  yang lebih tinggi. Uji robustitas dilakukan dengan menilai sensitivitas terhadap asumsi elastisitas (coba 0.3 dari Henderson et al. [6]). Rinciannya adalah eksklusi *outlier* (membuang 5% ekstrem), kemudian skema pembobotan alternatif (misal: *equal weights*), dan spesifikasi model alternatif (*Random Forest, LightGBM*).

## 3.5 Lingkungan Komputasi dan Replikabilitas

Analisis Data Python 3.9+ (menggunakan *library pandas, numpy, XGBoost, scipy, statsmodels, scikit-learn, shap*). Ekstraksi *Geospatial* menggunakan *Google Earth Engine* (JavaScript API). Visualisasi *matplotlib, seaborn, geopandas, plotly*. Replikabilitas menggunakan kode lengkap akan tersedia di repositori *GitHub*, skrip GEE dapat dibagikan.

## 4. RESULTS AND DISCUSSION

### 4.1 Deskripsi Data

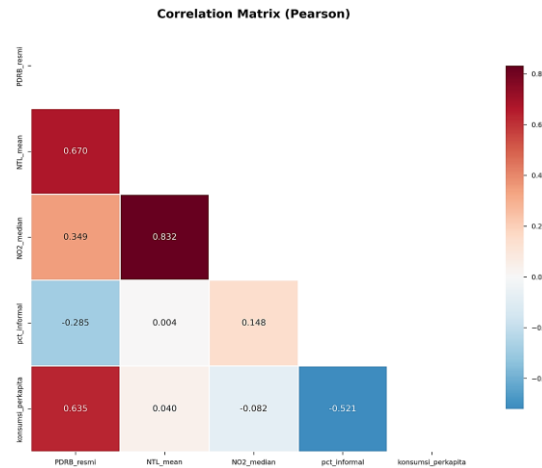
#### 4.1.1 Statistik Deskriptif

Tabel 1. Statistik Deskriptif Variabel Utama

| Variabel                                      | Mean        | Std Dev     | Median      | Min         | Max         |
|-----------------------------------------------|-------------|-------------|-------------|-------------|-------------|
| PDRB Resmi (Miliar Rp)                        | 535432.0353 | 756267.751  | 218697      | 39695.5     | 3443030     |
| NTL Intensity (nW/cm <sup>2</sup> /sr)        | 95941.418   | 78356.66487 | 71803.85    | 11018.17    | 401884.91   |
| NO <sub>2</sub> Density (mol/m <sup>2</sup> ) | 0.01868855  | 0.015637424 | 0.013762568 | 0.001543064 | 0.064817867 |
| % Pekerja Informal                            | 59.42005882 | 10.2049846  | 60.425      | 29.57       | 84.43       |
| Konsumsi RT (Rp/kapita)                       | 2691988.457 | 2246899.184 | 2195476.42  | 1106361.98  | 16759946.74 |
| IPM                                           | 72.54541176 | 3.888845542 | 72.68       | 60.84       | 83.55       |
| Jumlah Penduduk                               | 7986828.824 | 11280324.94 | 4244850     | 695600      | 50345200    |

Data panel yang digunakan terdiri dari 170 observasi (34 provinsi, 2019–2023). Statistik deskriptif (Tabel 1) menunjukkan variasi yang sangat besar pada variabel ekonomi, mencerminkan disparitas regional di Indonesia. PDRB Resmi berkisar dari Rp 39,7 triliun hingga Rp 3.443 triliun. Rata-rata persentase pekerja informal adalah 59,4%, yang menegaskan dominasi sektor informal dalam struktur tenaga kerja nasional.

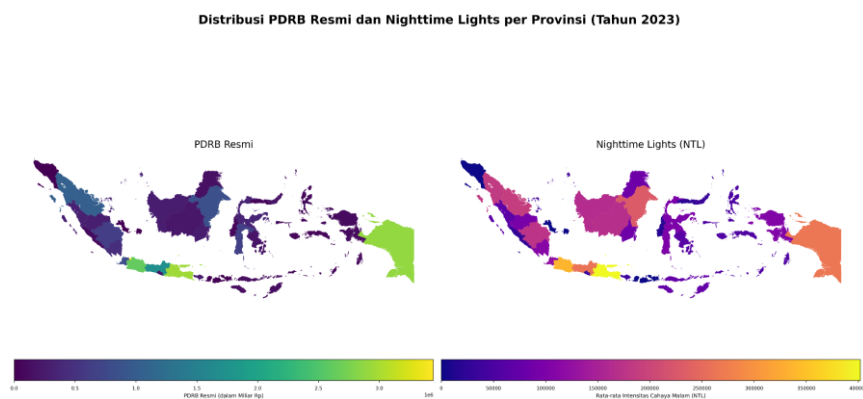
#### 4.1.2 Korelasi Antar variabel



**Gambar 1.** Matriks Korelasi *Pearson*

Matriks korelasi (Gambar 1) mengonfirmasi validitas variabel proksi yang dipilih. *NTL\_mean* ( $r=0.670$ ) dan *konsumsi\_perkapita* ( $r=0.635$ ) menunjukkan korelasi positif yang kuat dengan *PDRB\_resmi*. Seperti yang diharapkan, *pct\_informal* memiliki korelasi negatif yang kuat dengan *konsumsi\_perkapita* ( $r=-0.521$ ) dan korelasi negatif yang moderat dengan *PDRB\_resmi* ( $r=-0.285$ ), sesuai dengan teori *shadow economy*.

#### 4.1.3 Pola Spasial



**Gambar 2.** Distribusi PDRB Resmi dan *Nighttime Lights* per Provinsi (2023) - Dua peta choropleth

Visualisasi peta (Gambar 2) menunjukkan konsentrasi PDRB dan NTL yang tinggi di sentra-sentra ekonomi utama, terutama di Pulau Jawa. Terdapat pula beberapa wilayah (misalnya di Kalimantan dan Papua) yang menunjukkan intensitas NTL yang relatif tinggi namun PDRB resminya tidak sepadan, mengindikasikan potensi aktivitas ekonomi yang tidak tercatat

### 4.2 Hasil Stage 1 : *Machine learning Model Performance*

#### 4.2.1 Performa *Nidekl (XGBoost)*

**Tabel 2.** Ringkasan Kinerja Model

| Metric            | Value       |
|-------------------|-------------|
| $R^2$             | 0.828179651 |
| RMSE (Billion Rp) | 230380.5678 |
| MAE (Billion Rp)  | 107406.6909 |
| MAPE (%)          | 20.20247392 |

Model *XGBoost* (Tahap 1) dilatih pada adaptive weighted target dan diuji pada data tahun 2023 (temporal split). Hasilnya (Tabel 2) menunjukkan performa prediksi yang kuat, dengan nilai  $R^2 = 0.828$ . Ini berarti model mampu menjelaskan 82,8% variasi dalam target PDRB. Meskipun demikian, nilai MAPE

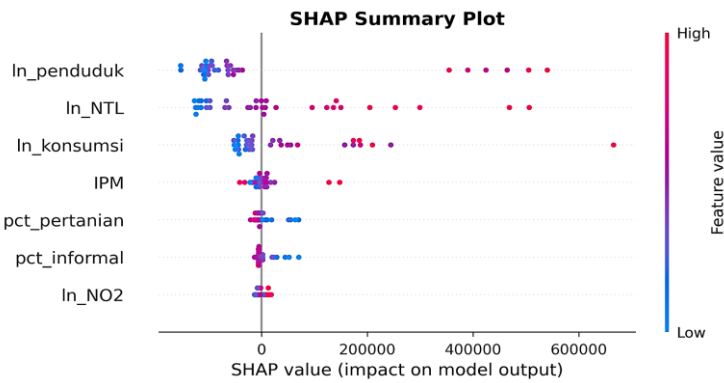
(20,2%) mengindikasikan bahwa rata-rata kesalahan prediksi masih relatif tinggi, yang mungkin mencerminkan kompleksitas dan heterogenitas ekonomi provinsi.

4.2.2 Analisis Feature Importance

Analisis SHAP (Tabel 3 dan Gambar 3) mengidentifikasi prediktor terpenting dalam model ML.  $\ln\_Penduduk$  (37,6%) adalah fitur dominan, diikuti oleh  $\ln\_NTL$  (28,6%) dan  $\ln\_Konsumsi$  (21,1%). Ini sangat logis, karena skala ekonomi (penduduk) dan aktivitas riil (cahaya malam dan konsumsi) adalah pendorong utama PDRB. Menariknya, proksi *geospatial*  $\ln\_NO_2$  (1,8%) dan variabel struktural % Informal (2,8%) memiliki kontribusi prediktif yang relatif kecil dalam model ini.

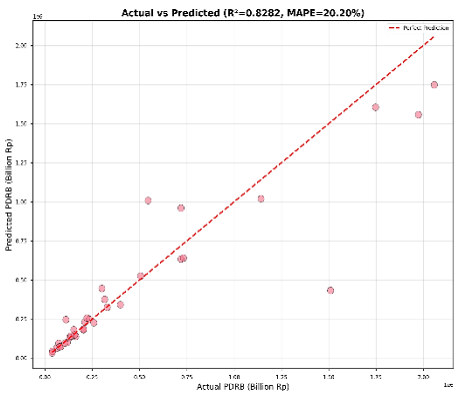
Tabel 3. Feature Importance (Mean |SHAP|, XGBoost)

| Feature         | Mean  SHAP  | % Contribution |
|-----------------|-------------|----------------|
| $\ln\_Penduduk$ | 152198.9917 | 37.57107151    |
| $\ln\_NTL$      | 115933.1973 | 28.61868136    |
| $\ln\_Konsumsi$ | 85585.4879  | 21.12719966    |
| $IPM$           | 18467.13462 | 4.558703232    |
| % Pertanian     | 13980.20999 | 3.451083766    |
| % Informal      | 11478.49834 | 2.833523909    |
| $\ln\_NO_2$     | 7452.703337 | 1.839736564    |

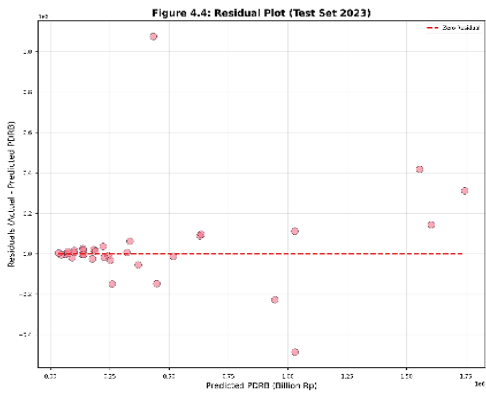


Gambar 3. SHAP Summary Plot

4.2.3 Diagnostik Model



Gambar 4. Actual vs Predicted (Test Set)



Gambar 5. Residual Plot

Plot Actual vs Predicted (Gambar 4) secara visual mengonfirmasi  $R^2$  yang baik, di mana sebagian besar data poin berada di sekitar garis 45 derajat. Plot Residual (Gambar 5) menunjukkan sebaran yang relatif acak di sekitar nol, yang mengindikasikan tidak ada bias sistematis yang jelas dalam model. Namun, terlihat adanya heteroskedastisitas (*error* membesar seiring meningkatnya PDRB prediksi), yang wajar terjadi pada data ekonomi lintas wilayah dengan skala yang sangat berbeda.

4.3 Analisis Discrepancy PDRB dan Shadow economy 2023

4.3.1 Hasil Agregat

Estimasi Tahap 2 (GMM) menghasilkan estimasi *True* PDRB nasional rata-rata sebesar Rp 19.435 triliun, dibandingkan PDRB resmi Rp 18.204 triliun (Tabel 4). Ini menyiratkan estimasi *shadow economy*

agregat sebesar Rp 1.230,8 triliun, atau setara dengan 6,76% dari PDRB resmi. Namun, perlu dicatat bahwa 95% Confidence Interval (CI) untuk estimasi ini sangat lebar (2,8% hingga 104,0%), yang menunjukkan adanya ketidakpastian model yang tinggi.

**Tabel 4.** Estimasi Agregat Nasional (Rata-rata 2019-2023)

| Metric                         | Value       |
|--------------------------------|-------------|
| Total PDRB Resmi (Trillion Rp) | 18204.6892  |
| Total True PDRB (Trillion Rp)  | 19435.50116 |
| Shadow economy (Trillion Rp)   | 1230.811965 |
| % Shadow economy               | 6.760961153 |
| 95% CI Lower                   | 5.255549554 |
| 95% CI Upper                   | 107.7550222 |

#### 4.3.2 Hasil Model Underreporting Rate

Tabel 5 menyajikan temuan krusial dari model GMM yang mengestimasi underreporting rate ( $\delta$ ). Hasil model GMM menunjukkan koefisien yang signifikan positif hanya untuk % Pertanian ( $\beta_2$ ) ( $p < 0.01$ ). Variabel % Informal ( $\beta_1$ ) secara statistik tidak signifikan, meskipun estimasi koefisiennya positif (0.0097). Hal ini mengindikasikan bahwa dalam model struktural ini, struktur ekonomi berbasis pertanian adalah prediktor *underreporting* yang lebih kuat daripada tingkat informalitas umum.

**Tabel 5.** Estimasi Koefisien Model *Underreporting Rate* ( $\delta$ )

| Parameter                 | Estimate | Mean_Boot | Std_Error | CI_Lower | CI_Upper | Significant |
|---------------------------|----------|-----------|-----------|----------|----------|-------------|
| Intercept ( $\beta_0$ )   | -4.77399 | -2.59659  | 1.567216  | -5       | 0.994467 | No          |
| % Informal ( $\beta_1$ )  | 0.009751 | 0.008103  | 0.009186  | 0        | 0.027023 | No          |
| % Pertanian ( $\beta_2$ ) | 0.020736 | 0.040747  | 0.02205   | 0.005235 | 0.1      | Yes***      |
| IPM ( $\beta_3$ )         | 0.016377 | 0.008385  | 0.021135  | -0.04352 | 0.04055  | No          |

#### 4.3.3 Peringkat *Shadow economy* Provinsi

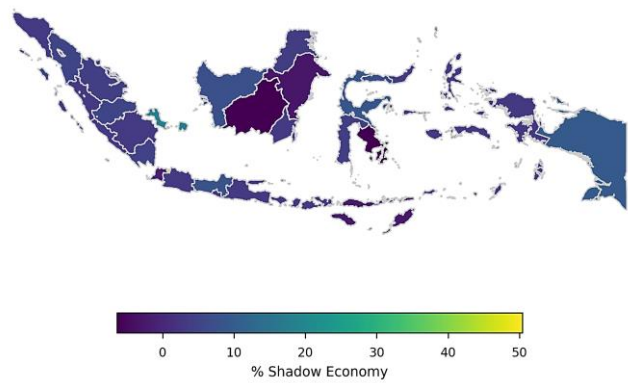
**Tabel 6.** Top 10 Provinsi dengan *Shadow economy* Tertinggi (%)

| Province                   | PDRB Official (T Rp) | Shadow economy (T Rp) | % Shadow    | Underreporting Rate |
|----------------------------|----------------------|-----------------------|-------------|---------------------|
| Kalimantan Tengah          | 208.783              | 98.01211071           | 46.94448816 | 0.068180384         |
| Papua Barat                | 97.6809              | 37.00094781           | 37.87940919 | 0.050843798         |
| Daerah Istimewa Yogyakarta | 180.696              | 21.61076654           | 11.95973709 | 0.062419444         |
| Banten                     | 814.122              | 96.13570759           | 11.80851366 | 0.049129647         |
| Maluku                     | 58.4878              | 5.819790193           | 9.950434437 | 0.079433283         |
| Aceh                       | 227.018              | 21.7912259            | 9.598897841 | 0.088660351         |
| Lampung                    | 448.851              | 41.4933733            | 9.244353538 | 0.088509965         |
| Bengkulu                   | 96.5831              | 8.821600006           | 9.133689027 | 0.090271505         |
| Jambi                      | 293.78               | 26.29332578           | 8.950005373 | 0.089152161         |
| Sumatera Utara             | 1051                 | 90.64527523           | 8.624669385 | 0.077003736         |

**Tabel 7.** Bottom 5 Provinsi (*Shadow economy* Terendah)

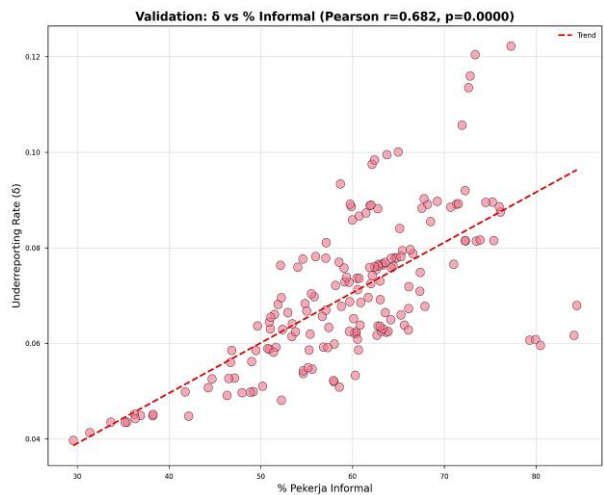
| Province                  | PDRB Official (T Rp) | Shadow economy (T Rp) | % Shadow    | Underreporting Rate |
|---------------------------|----------------------|-----------------------|-------------|---------------------|
| Kepulauan Riau            | 331.645              | 15.74841786           | 4.748576899 | 0.043520492         |
| Kepulauan Bangka Belitung | 102.527              | 3.390506243           | 3.306939872 | 0.065468846         |
| Gorontalo                 | 51.3667              | 1.563320571           | 3.043451441 | 0.099492121         |
| Maluku Utara              | 85.1431              | 0.52818083            | 0.620344843 | 0.062832885         |
| Sulawesi Barat            | 58.5721              | -1.259073996          | -           | 0.120438382         |

Estimasi *shadow economy* menunjukkan heterogenitas ekstrem antar provinsi (Tabel 6, Tabel 7, Gambar 6). Kalimantan Tengah menunjukkan persentase *shadow economy* tertinggi (46,9%), diikuti oleh Papua Barat (37,8%). Sebaliknya, beberapa provinsi menunjukkan *shadow economy* yang sangat rendah, seperti Sulawesi Barat (-2,1%, mengindikasikan PDRB resmi lebih tinggi dari prediksi), Maluku Utara (0,6%), Gorontalo (3,0%), dan Kep. Bangka Belitung (3,3%).



Gambar 6. Peta *Shadow economy* per Provinsi (2023) - Peta Choropleth

4.3.4 Validasi: Korelasi dengan % Informasi



Gambar 7. Scatter Plot - Underreporting Rate vs % Pekerja Informal

Meskipun % Informal tidak signifikan di dalam model GMM (Tabel 5), Gambar 7 menunjukkan temuan validasi yang penting. Terdapat korelasi positif yang kuat dan signifikan secara statistik (Pearson  $r=0.682$ ,  $p < 0.001$ ) antara underreporting rate ( $\delta$ ) yang diestimasi oleh model dan data % Pekerja Informal dari BPS. Ini memvalidasi bahwa *output* model GMM ( $\delta$ ) secara keseluruhan berhasil menangkap fenomena yang sangat terkait dengan informalitas, meskipun komponen % Pertanian memiliki daya jelas yang lebih dominan di dalam model.

4.4 Estimasi Tax Gap dan Implikasi Kebijakan

4.4.1 Tax Gap Agregat Nasional

Tabel 8. Tax Gap Agregat Nasional (Rata-rata 2019-2023)

| Metric                               | Value       |
|--------------------------------------|-------------|
| <i>Tax Potential (Trillion Rp)</i>   | 230.1646383 |
| <i>Tax Realization (Trillion Rp)</i> | 213.994066  |
| <i>Tax Gap (Trillion Rp)</i>         | 16.17057231 |
| <i>% Tax Gap</i>                     | 7.025654518 |

Berdasarkan estimasi *True* PDRB, potensi pajak nasional (Tabel 8) diestimasi sebesar Rp 230,16 triliun, sementara realisasi hanya Rp 213,99 triliun. Ini menghasilkan tax gap agregat (batas bawah) sebesar Rp 16,17 triliun, atau 7,03% dari total potensi.

4.4.2 Peringkat Tax Gap Provinsi

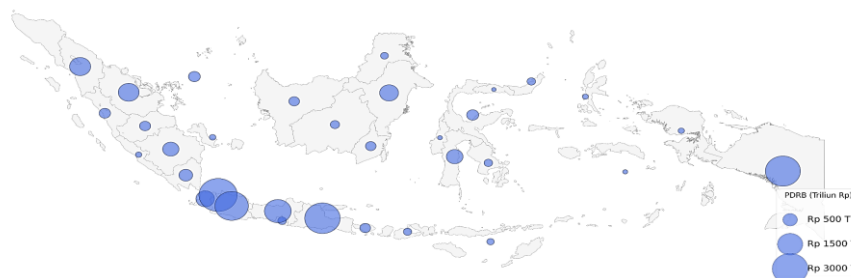
Peringkat tax gap absolut (Tabel 9, Gambar 8) menunjukkan gambaran yang berbeda dari peringkat persentase *shadow economy*. Tax gap nominal terbesar ditemukan di provinsi dengan skala ekonomi besar. Papua memimpin dengan tax gap Rp 34,5 triliun (karena basis PDRB resminya sangat besar dan memiliki tax

gap % yang tinggi), diikuti oleh Jawa Timur (Rp 8,9 triliun), Riau (Rp 7,3 triliun), dan DKI Jakarta (Rp 3,7 triliun).

**Tabel 9.** Top 15 Provinsi dengan Tax Gap Tertinggi

| Province         | Tax Gap (T Rp) | % <i>Shadow economy</i> | % Tax Gap   |
|------------------|----------------|-------------------------|-------------|
| Papua            | 34.54593948    | 7.024729096             | 94.67315281 |
| Jawa Timur       | 8.958230662    | 6.694779515             | 24.00445698 |
| Riau             | 7.335511916    | 8.148564831             | 55.79824612 |
| DKI Jakarta      | 3.679426889    | 4.784649324             | 8.61189745  |
| Sumatera Utara   | 3.332117476    | 8.624669385             | 24.64598975 |
| Sulawesi Tengah  | 2.351024338    | 7.024729087             | 53.43511183 |
| Sumatera Selatan | 1.76123308     | 6.812153128             | 22.13023323 |
| Lampung          | 1.65522607     | 9.244353538             | 28.50448932 |
| Jawa Tengah      | 1.509490549    | 6.694779516             | 7.045565236 |
| Sulawesi Selatan | 1.483944885    | 8.14856483              | 17.75313853 |
| Jambi            | 1.47927368     | 8.950005373             | 39.02619322 |
| Sumatera Barat   | 1.068899625    | 8.299688024             | 26.6466833  |
| Kalimantan Utara | 0.920446982    | 5.933985817             | 49.81725356 |
| Papua Barat      | 0.885117813    | 37.87940919             | 55.49439969 |
| Kepulauan Riau   | 0.811141471    | 4.748576899             | 19.71660624 |

**Peta Sebaran PDRB per Provinsi (Tahun 2023)**



**Gambar 8.** Peta Tax Gap per Provinsi - Bubble Map

#### 4.4.3 Kerangka Kerja Prioritas untuk DJP

**Tabel 10.** Klasifikasi Tier Provinsi

| tier   | N_Provinces | Total_Tax_Gap | Mean_Shadow_Pct | Pct_of_Total_Gap |
|--------|-------------|---------------|-----------------|------------------|
| Tier 2 | 5           | 52.53409649   | 21.33839416     | 94.96326968      |
| Tier 3 | 29          | 2.786341264   | 6.826573482     | 5.036730324      |

Tabel 10 menyajikan klasifikasi tier prioritas. Tidak ada provinsi yang memenuhi kriteria Tier 1 (Gap Absolut > 10 T DAN % Shadow > 25%). Namun, Tier 2 (didefinisikan sebagai Gap > 5 T ATAU % Shadow > 20%) mencakup 5 provinsi, yaitu Papua, Jawa Timur, Riau, Kalimantan Tengah, dan Papua Barat. Kelima provinsi ini menyumbang 95% dari total tax gap nasional yang teridentifikasi. Sebanyak 29 provinsi lainnya masuk ke Tier 3, yang hanya menyumbang 5% dari total gap. Temuan ini memberikan fokus yang sangat tajam bagi DJP untuk mengalokasikan sumber daya ekstensifikasi

#### 4.5 Implikasi Kebijakan dan Keterbatasan

##### 4.5.1 Interpretasi Hasil

Estimasi *shadow economy* nasional kami (6,76%) secara signifikan lebih rendah dari studi sebelumnya (misal, 21,5% oleh Medina & Schneider (2018) ). Kami menginterpretasikan ini sebagai estimasi batas bawah yang konservatif. Temuan GMM (Tabel 4.5) yang menunjukkan % Pertanian sebagai satu-satunya prediktor underreporting yang signifikan (sementara % Informal tidak) sangat mendukung hipotesis ini. Hal itu menyiratkan bahwa model kami saat ini sangat berhasil mengidentifikasi *shadow economy* berbasis agrikultur/tradisional, namun memiliki blind spot (titik buta) terhadap *shadow economy* digital/jasa di perkotaan padat di mana proksi NTL dan informalitas BPS gagal membedakan aktivitas.

Meskipun demikian, korelasi yang kuat pada Gambar 4.6 ( $r=0.682$ ) membuktikan bahwa *Output* model tetap valid secara konseptual. Lebarnya 95% CI (2,8% hingga 107,7%) juga menunjukkan ketidakpastian estimasi yang tinggi, menegaskan bahwa fokus harus pada peringkat relatif, bukan angka absolut.

#### 4.5.2 Kebijakan Prioritas

Rekomendasi utama untuk DJP adalah memfokuskan upaya ekstensifikasi pada 5 provinsi Tier 2 (Papua, Jawa Timur, Riau, Kalimantan Tengah, Papua Barat). Kelima wilayah ini menyumbang 95% dari total tax gap teridentifikasi. Pendekatannya harus dibedakan. Provinsi Jawa Timur & Riau Fokus pada *gap* nominal yang besar di sektor yang sudah teridentifikasi (misal, UMKM manufaktur, jasa, dan perkebunan). Provinsi Papua yang memiliki *gap* nominal terbesar, memerlukan investigasi khusus untuk memisahkan sektor formal (pertambangan) dari *shadow economy* yang besar melalui audit investigatif khusus. Terakhir, provinsi Kalimantan Tengah & Papua Barat yang memiliki % *shadow economy* tertinggi (47% dan 38%), memerlukan program formalisasi dan ekstensifikasi skala penuh yang berfokus pada sektor agrikultur dan sumber daya alam.

#### 4.5.3 Keterbatasan Model

Sebagaimana telah diuraikan dalam Bagian 1.5 dan dikonfirmasi oleh hasil, keterbatasan utama adalah *Blind Spot* Model sebab Model gagal mengidentifikasi % Informal sebagai prediktor signifikan dalam GMM, menunjukkan kelemahan dalam menangkap *shadow economy* jasa/digital perkotaan. Kemudian, ketidakpastian tinggi sebab CI yang sangat lebar (95% CI: 2,8%-107,7%) berarti estimasi titik (6,76%) memiliki ketidakpastian yang masif, kemungkinan akibat pendeknya periode data (5 tahun) yang mencakup guncangan COVID-19, dan simplifikasi rasio pajak: penggunaan satu tarif nasional (7,03%) adalah simplifikasi besar.

### 5. KESIMPULAN

Penelitian ini berhasil mengaplikasikan pendekatan *machine learning* (XGBoost) yang diintegrasikan dengan *latent variable framework* (GMM) untuk mengestimasi *shadow economy* dan tax gap di 34 provinsi Indonesia. Temuan utama menunjukkan estimasi *shadow economy* nasional yang konservatif (batas bawah) sebesar 6,76% dari PDRB resmi, dengan estimasi tax gap agregat sebesar Rp 16,17 triliun (7,03% dari potensi). Secara metodologis, model GMM mengungkap temuan penting: struktur ekonomi berbasis pertanian terbukti menjadi prediktor underreporting yang signifikan secara statistik, sementara variabel pekerja informal tidak. Meskipun demikian, validasi model menunjukkan bahwa *Output* underreporting rate ( $\delta$ ) tetap memiliki korelasi positif yang kuat ( $r=0.682$ ) dengan data pekerja informal BPS.

Penelitian ini mengungkap heterogenitas yang ekstrem, dengan provinsi seperti Kalimantan Tengah (46,9%) dan Papua Barat (37,8%) memiliki persentase *shadow economy* tertinggi, sementara provinsi seperti Sulawesi Barat (-2,1%) dan Gorontalo (3,0%) menunjukkan sebaliknya. Tax gap nominal terbesar terkonsentrasi di Papua (Rp 34,5 T) dan Jawa Timur (Rp 8,9 T). Kerangka kerja prioritas mengidentifikasi 5 provinsi Tier 2 yang menyumbang 95% dari total tax gap teridentifikasi. Penelitian ini menyediakan kerangka kerja berbasis data (evidence-based framework) yang dapat direplikasi bagi DJP untuk memfokuskan strategi ekstensifikasi.

Penelitian ini menekankan perlunya transformasi pengawasan pajak dari berbasis dokumen ke berbasis data spasial (*geospatial-driven audit*). DJP dapat memprioritaskan wilayah dengan diskrepansi ekonomi, fisik dan administratif yang tinggi. Selain itu, temuan mengenai sektor pertanian sebagai prediktor utama *underreporting* menuntut kebijakan simplifikasi pajak yang lebih inklusif bagi sektor tersebut. Saran untuk penelitian mendatang adalah penggunaan data satelit dengan resolusi lebih tinggi serta integrasi data transaksi ekonomi digital diharapkan dapat mempertajam akurasi estimasi, terutama dalam menangkap aktivitas *shadow economy* pada sektor jasa yang tidak sepenuhnya terefleksikan melalui emisi cahaya malam

### REFERENSI

- [1] Alm, J. & Martinez-Vazquez, J. (2007). Tax morale and tax evasion in Latin America. *International Tax and Public Finance*, 14(5), pp. 619-644. Tersedia di: <https://doi.org/10.1007/s10797-007-9032-1>
- [2] Andiojaya, A., Demena, B. A. & van Bergeijk, P. A. G. (2021). Utilizing nighttime-light (NTL) data to measure Indonesia's gross domestic regional product. *Regional Science Policy & Practice*, 13(4), pp. 1273-1292. Tersedia di: <https://doi.org/10.1111/rsp3.12429>
- [3] Chen, X. & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21), pp. 8589-8594. Tersedia di: <https://doi.org/10.1073/pnas.1017031108>
- [4] Gibson, J., Olivia, S., Boe-Gibson, G. & Li, C. (2021). Which night lights data should we use in economics, and where? *Journal of Development Economics*, 149, 102602. Tersedia di: <https://doi.org/10.1016/j.jdevec.2020.102602>
- [5] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. & Moore, R. (2017). Google Earth Engine: Planetary-scale *geospatial* analysis for everyone. *Remote Sensing of Environment*, 202, pp. 18-27. Tersedia di: <https://doi.org/10.1016/j.rse.2017.06.031>
- [6] Henderson, J. V., Storeygard, A. & Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*, 102(2), pp. 994-1028. Tersedia di: <https://doi.org/10.1257/aer.102.2.994>

- [7] Hu, Y. & Yao, J. (2022). Illuminating economic growth. *Journal of Econometrics*, 228(2), pp. 359-378. Tersedia di: <https://doi.org/10.1016/j.jeconom.2021.05.007>
- [8] Ialongo, I., Virta, H., Eskes, H., Martin, C. & Kivi, R. (2020). Validation of Sentinel-5P TROPOMI NO<sub>2</sub> column observations with MAX-DOAS measurements in Helsinki, Finland. *Atmospheric Measurement Techniques*, 13(9), pp. 5005-5017. Tersedia di: <https://doi.org/10.5194/amt-13-5005-2020>
- [9] Indupurnahayu, I. & Walujadi, D. (2020). *Shadow economy* and the loss of potential tax revenues in Indonesia. *Economics and Finance in Indonesia*, 66(1), pp. 18-39. Tersedia di: <https://doi.org/10.47291/efi.v66i1.615>
- [10] Lukman, S. & Kartiasih, F. (2025). Estimating the *shadow economy* at the provincial level in Indonesia: A MIMIC model approach. *Economies*, 13(1), 15. Tersedia di: <https://doi.org/10.3390/economies13010015>
- [11] Martínez, L. R. (2022). How Much Should We Trust the Dictator's GDP Growth Estimates? *Journal of Political Economy*, 130(10), pp. 2731-2769. Tersedia di: <https://doi.org/10.1086/720458>
- [12] Medina, L. & Schneider, F. (2018). Shadow economies around the world: What did we learn over the last 20 years? *African Development Review*, 30(1), pp. 17-29. Tersedia di: <https://doi.org/10.1111/1467-8268.12319>
- [13] Mullainathan, S. & Spiess, J. (2017). *Machine learning: An applied econometric approach*. *Journal of Economic Perspectives*, 31(2), pp. 87-106. Tersedia di: <https://doi.org/10.1257/jep.31.2.87>
- [14] Schneider, F. & Enste, D. H. (2000). Shadow economies: Size, causes, and consequences. *Journal of Economic Literature*, 38(1), pp. 77-114. Tersedia di: <https://doi.org/10.1257/jel.38.1.77>
- [15] Lundberg, S. M. & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Curran Associates, Inc., pp. 4765-4774.
- [16] Paramita, R. D. (2020). *Analisis Faktor-Faktor yang Mempengaruhi Ukuran Shadow economy di Indonesia*. Skripsi Sarjana. Universitas Gadjah Mada.
- [17] James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R*. Edisi ke-2. Springer.
- [18] Schneider, F. & Enste, D. H. (2013). *The shadow economy: An international survey*. Edisi ke-2. Cambridge University Press.
- [19] Gemmell, N. & Hasseldine, J. (2012). The Tax Gap: A Methodological Review. In: Stock, T. (Ed.) *Advances in Taxation*. Vol. 20. Emerald Group Publishing Limited, pp. 203-231. Tersedia di: [https://doi.org/10.1108/S1058-7497\(2012\)0000020011](https://doi.org/10.1108/S1058-7497(2012)0000020011)
- [20] Badan Pusat Statistik (2023a). *Produk domestik regional bruto provinsi-provinsi di Indonesia 2023*. BPS-Statistics Indonesia.
- [21] Badan Pusat Statistik (2023b). *Keadaan angkatan kerja di Indonesia Februari 2023*. BPS-Statistics Indonesia.
- [22] Badan Pusat Statistik (2023c). *Pengeluaran untuk konsumsi penduduk Indonesia 2023*. BPS-Statistics Indonesia.
- [23] Direktorat Jenderal Pajak (2023). *Laporan tahunan Direktorat Jenderal Pajak 2023*. Kementerian Keuangan Republik Indonesia.
- [24] International Monetary Fund (2019). *Measuring the digital economy*. IMF Policy Paper.
- [25] Kementerian Keuangan (2024). *APBN KITA Edisi Januari 2024*. Kementerian Keuangan Republik Indonesia.
- [26] OECD (2021). *Tax administration 2021: Comparative information on OECD and other advanced and emerging economies*. OECD Publishing.
- [27] OECD (2023a). *Revenue statistics 2023*. OECD Publishing.
- [28] OECD (2023b). *Revenue statistics in Asia and the Pacific 2023*. OECD Publishing.
- [29] Pesaresi, M., Ehrlich, D., Florczyk, A. J., Freire, S., Julea, A., Kemper, T., Soille, P. & Syrris, V. (2016). *GHS built-up grid, derived from Landsat, multitemporal (1975, 1990, 2000, 2014) [Dataset]*. European Commission, Joint Research Centre (JRC).
- [30] Internal Revenue Service (n.d.). *The tax gap*. Tersedia di: <https://www.irs.gov/newsroom/the-tax-gap> [Diakses 25 Oktober 2025].