

**LAPORAN ANALISIS DATA SPASIAL**  
**“ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI**  
**TINGKAT KEMISKINAN DI PROVINSI JAWA TIMUR TAHUN**  
**2024: PENDEKATAN PERBANDINGAN MODEL”**



Kelompok F

Ammar Hanafi	2206051582
Bryan Jonathan	2206052780
Renata Shaula Alfino Ritonga	2206815812

**Departemen Matematika**  
**Fakultas Matematika dan Ilmu Pengetahuan Alam**  
**Universitas Indonesia**  
**2025**

# DAFTAR ISI

<b>BAB 1 Pendahuluan.....</b>	<b>3</b>
1.1 Latar Belakang.....	3
1.2 Rumusan Masalah.....	4
1.3 Tujuan Penelitian.....	4
1.4 Batasan Masalah.....	4
<b>BAB 2 Landasan Teori.....</b>	<b>5</b>
2.1 Definisi Variabel dan alasan pemilihan variabel.....	5
2.1.1 Variabel Dependen: Persentase Penduduk Miskin.....	5
2.1.2 Variabel Independen.....	5
2.2 Teori Pendukung: Regresi Linier Berganda.....	6
2.2.1 Uji Signifikansi Simultan.....	7
2.2.2 Uji Signifikansi Parsial.....	8
2.2.3 Uji Multikolinieritas.....	8
2.2.4 Uji Normalitas.....	8
2.2.5 Uji Autokorelasi.....	9
<b>BAB 3 Data dan Metode.....</b>	<b>10</b>
3.1 Penjelasan Data.....	10
3.2 Teori Uji Kesesuaian Kondisi Data.....	10
3.2.1 Uji Moran's I.....	11
3.3 Pengujian Efek Spasial.....	11
3.3.1 Keragaman Spasial.....	11
3.3.2 Lagrange Multiplier Lag (LM Lag).....	12
3.3.3 Lagrange Multiplier Error (LM Error).....	13
3.3 Spatial Lag Model (SLM).....	13
3.4 Spatial Error Model (SEM).....	14
3.5 Random Forest.....	14
3.6 Penentuan Model Terbaik.....	15
3.6.1 Akaike Information Criterion (AIC).....	15
3.6.2 Koefisien Determinasi (R <sup>2</sup> ).....	15
3.6.3 Bayesian Information Criterion (BIC).....	16
<b>BAB 4 Hasil dan Pembahasan.....</b>	<b>17</b>
4.1 Statistik Deskriptif.....	17
4.2 Hasil Uji Dependensi Spasial.....	20
Implikasi Hasil Uji.....	23
4.3 Hasil Analisis dari Ketiga Model.....	25
4.4 Perbandingan Model.....	32
<b>Bab 5 Kesimpulan dan Saran.....</b>	<b>36</b>
5.1 Kesimpulan.....	36
5.2 Saran.....	36
<b>Lampiran.....</b>	<b>38</b>
<b>Daftar Pustaka.....</b>	<b>48</b>

# **BAB 1**

## **Pendahuluan**

### **1.1 Latar Belakang**

Kemiskinan merupakan isu sosial-ekonomi yang kompleks dan masih menjadi tantangan utama di Indonesia, termasuk di Provinsi Jawa Timur. Berdasarkan data Badan Pusat Statistik (BPS), persentase penduduk miskin di Jawa Timur pada Maret 2024 tercatat sebesar 9.79%, dengan jumlah penduduk miskin mencapai 3.893 juta orang. Meskipun terdapat penurunan dibandingkan tahun sebelumnya, disparitas kemiskinan antar wilayah masih signifikan. Sebagai contoh, Kabupaten Sampang mencatatkan angka persentase penduduk miskin sebesar 20.83% pada tahun 2024.

Fenomena kemiskinan di Jawa Timur dapat dilihat dari kesenjangan yang ada antara wilayah-wilayah yang lebih berkembang dengan yang kurang berkembang. Di wilayah yang lebih maju, umumnya tingkat kemiskinan lebih rendah dan terdapat lebih banyak akses terhadap berbagai layanan dasar seperti pendidikan, kesehatan, dan infrastruktur. Sebaliknya, di wilayah yang lebih terisolasi, tingkat kemiskinan cenderung lebih tinggi, yang berhubungan dengan kurangnya akses terhadap layanan-layanan tersebut.

Kemiskinan tidak hanya dipengaruhi oleh faktor ekonomi, tetapi juga dipengaruhi oleh faktor sosial dan geografi. Ketimpangan dalam distribusi pendapatan, tingkat pendidikan, serta keterbatasan dalam akses terhadap layanan kesehatan menjadi beberapa aspek penting yang perlu diperhatikan dalam upaya mengurangi tingkat kemiskinan. Selain itu, mobilitas ekonomi yang lebih rendah di wilayah tertentu dapat memperburuk ketimpangan yang ada.

Fenomena ini mendorong perlunya analisis yang lebih mendalam mengenai penyebab dan distribusi kemiskinan di tingkat kabupaten/kota di Provinsi Jawa Timur. Untuk menganalisis pengaruh faktor-faktor yang mempengaruhi tingkat kemiskinan, diperlukan pendekatan analisis data spasial yang mempertimbangkan keterkaitan antar wilayah. Model regresi spasial dan *Random Forest* merupakan dua metode yang dapat digunakan untuk tujuan ini. Regresi spasial memungkinkan identifikasi hubungan antar wilayah, sementara *Random Forest* dapat menangani kompleksitas data dengan baik.

Analisis yang memperhitungkan keterkaitan antarwilayah dan mempertimbangkan faktor-faktor yang mempengaruhi kemiskinan akan memberikan gambaran yang lebih jelas mengenai kondisi sosial ekonomi di provinsi ini, serta memungkinkan pembuatan kebijakan yang lebih tepat sasaran untuk mengurangi kesenjangan antar wilayah dan meningkatkan kesejahteraan masyarakat.

## **1.2 Rumusan Masalah**

Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana pengaruh variabel-variabel independen terhadap tingkat kemiskinan di Provinsi Jawa Timur tahun 2024?
2. Bagaimana perbandingan kinerja model regresi spasial dan *Random Forest* dalam memprediksi tingkat kemiskinan berdasarkan faktor-faktor tersebut?

## **1.3 Tujuan Penelitian**

Tujuan dari penelitian ini adalah:

1. Menganalisis pengaruh variabel-variabel independen terhadap tingkat kemiskinan di Provinsi Jawa Timur tahun 2024.
2. Membandingkan kinerja model regresi spasial dan *Random Forest* dalam memprediksi tingkat kemiskinan berdasarkan faktor-faktor tersebut.

## **1.4 Batasan Masalah**

Untuk membuat analisis lebih terfokus, penelitian ini memiliki batasan sebagai berikut:

1. Penelitian ini hanya mencakup 38 kabupaten/kota di Provinsi Jawa Timur.
2. Data yang digunakan adalah data tahun 2024.
3. Tingkat kemiskinan diukur berdasarkan persentase penduduk miskin menurut BPS.
4. Ketimpangan pendapatan (Gini Ratio), rata-rata lama sekolah, jumlah tenaga medis, dan jumlah kendaraan bermotor sepeda motor adalah variabel independen yang akan digunakan dalam analisis.
5. Metode analisis yang dilakukan adalah melakukan perbandingan antara model regresi spasial dan *Random Forest*.

## **BAB 2**

### **Landasan Teori**

#### **2.1 Definisi Variabel dan alasan pemilihan variabel**

##### **2.1.1 Variabel Dependen: Persentase Penduduk Miskin**

Persentase penduduk miskin adalah proporsi jumlah penduduk yang hidup di bawah garis kemiskinan, diukur berdasarkan konsumsi atau pendapatan per kapita yang tidak mencukupi kebutuhan dasar.

Variabel ini dipilih karena merupakan indikator utama dalam mengukur tingkat kemiskinan di suatu wilayah. Data persentase penduduk miskin tersedia secara rutin dan dapat dibandingkan antar kabupaten/kota di Provinsi Jawa Timur.

##### **2.1.2 Variabel Independen**

###### **1. Gini Ratio (Koefisien Gini)**

Koefisien Gini adalah ukuran ketimpangan distribusi pendapatan atau konsumsi dalam suatu populasi. Nilai Gini berkisar antara 0 (kesetaraan sempurna) hingga 1 (ketimpangan sempurna). Ketimpangan pendapatan yang tinggi dapat memperburuk distribusi kesejahteraan dan meningkatkan tingkat kemiskinan. Oleh karena itu, Gini Ratio dipilih sebagai indikator ketimpangan ekonomi.

###### **2. Rata-rata Lama Sekolah (RLS)**

Rata-rata Lama Sekolah adalah jumlah rata-rata tahun yang dihabiskan oleh penduduk usia 15 tahun ke atas dalam menempuh pendidikan formal. Tingkat pendidikan yang rendah sering kali berhubungan dengan rendahnya kualitas hidup dan peningkatan kemiskinan. RLS digunakan untuk menggambarkan tingkat pendidikan masyarakat di suatu wilayah.

###### **3. Jumlah Tenaga Medis**

Jumlah tenaga medis mencakup dokter, dokter gigi, dokter spesialis, dan tenaga medis lainnya yang memberikan pelayanan kesehatan di suatu

wilayah. Akses terhadap layanan kesehatan yang memadai dapat meningkatkan kualitas hidup dan mengurangi angka kemiskinan. Jumlah tenaga medis digunakan sebagai indikator aksesibilitas layanan kesehatan di wilayah tersebut.

#### 4. Jumlah Kendaraan Bermotor

Jumlah kendaraan bermotor (sepeda motor) mencakup jumlah sepeda motor yang terdaftar di suatu wilayah administratif. Jumlah sepeda motor digunakan sebagai indikator mobilitas ekonomi dan akses transportasi yang dapat memengaruhi tingkat kemiskinan. Sepeda motor banyak digunakan di wilayah perkotaan maupun perdesaan karena biaya yang lebih terjangkau dibandingkan kendaraan roda empat. Variabel ini dipilih karena sepeda motor dapat mencerminkan potensi mobilitas dan keterhubungan wilayah, yang berperan dalam memperbaiki kondisi ekonomi masyarakat.

## 2.2 Teori Pendukung: Regresi Linier Berganda

Regresi linier berganda adalah metode statistik yang digunakan untuk menganalisis hubungan antara satu variabel dependen dengan dua atau lebih variabel independen secara simultan. Model ini mengasumsikan adanya hubungan linier antara variabel-variabel tersebut.

### Model Regresi Linier Berganda

Model regresi linier berganda untuk  $n$  observasi dan  $p$  variabel independen dapat digambarkan sebagai berikut:

$$\begin{aligned}y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_p X_{1p} + \epsilon_1 \\y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_p X_{2p} + \epsilon_2 \\y_3 &= \beta_0 + \beta_1 x_{31} + \cdots + \beta_p X_{3p} + \epsilon_3 \\&\vdots \\y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p X_{np} + \epsilon_n\end{aligned}$$

Secara matriks, model ini dapat dituliskan sebagai:

$$Y = X\beta + \epsilon$$

Dimana:

- Y adalah vektor variabel dependen
- X adalah matriks variabel independen
- $\beta$  adalah vektor koefisien regresi yang menunjukkan pengaruh masing-masing variabel independen terhadap variabel dependen
- $\epsilon$  adalah vektor galat (*error terms*)

Koefisien regresi  $\beta$  dapat diestimasi menggunakan metode *Ordinary Least Squares* (OLS), yang meminimalkan jumlah kuadrat galat (JKG). Estimasi koefisien regresi diperoleh dengan rumus:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

#### Asumsi Dasar Regresi Linier Berganda

Model regresi linier berganda mengasumsikan beberapa hal dasar dalam analisisnya:

1. Linearitas: Hubungan antara variabel dependen dan variabel independen bersifat linier.
2. Independensi: Setiap observasi harus independen satu sama lain.
3. Homoskedastisitas: Varians galat harus konstan di seluruh nilai variabel independen.
4. Normalitas: Galat harus terdistribusi normal.

#### 2.2.1 Uji Signifikansi Simultan

Uji signifikansi simultan berguna untuk menguji apakah semua variabel independen secara bersamaan berpengaruh signifikan terhadap variabel dependen menggunakan uji F. Hipotesis dari uji signifikansi simultan adalah:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_1 : \text{minimal terdapat satu } \beta_k \neq 0; k = 1, 2, \dots, p.$$

Statistik uji yang digunakan adalah

$$F = \frac{\frac{(SS_{yy} - SSR)}{p}}{\frac{SSR}{[n-(p+1)]}}$$

dimana

$$SS_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^n \left( y_i - \widehat{y}_i \right)^2$$

$H_0$  ditolak pada taraf signifikansi  $\alpha$  jika  $F > F_{\alpha, p, n-(p+1)}$

### 2.2.2 Uji Signifikansi Parsial

Uji signifikansi parsial berguna untuk menguji pengaruh masing-masing variabel independen terhadap variabel dependen menggunakan uji t. Hipotesis dari uji signifikansi parsial adalah:

$$H_0 : \beta_j = 0, j = 1, 2, \dots, p$$

$$H_1 : \beta_j \neq 0, j = 1, 2, \dots, p$$

Statistik uji yang digunakan adalah

$$t = \frac{\widehat{\beta}_j}{s_{\widehat{\beta}_j}}$$

$H_0$  ditolak pada taraf signifikansi  $\alpha$  jika  $|t| > t_{\alpha/2}$ .

### 2.2.3 Uji Multikolinieritas

Uji Multikolinieritas berguna untuk menilai adanya hubungan kuat antar variabel independen menggunakan *Variance Inflation Factor* (VIF).

$$VIF = \frac{1}{(1 - R_k^2)}$$

dengan  $R_k^2$  adalah nilai koefisien determinasi dari model regresi variabel independen  $x_k$  yang dijelaskan oleh variabel independen lainnya. Nilai  $VIF > 10$  menandakan bahwa terdapat multikolinieritas pada variabel independen.

### 2.2.4 Uji Normalitas

Uji Normalitas atau uji Anderson-Darling digunakan untuk menguji apakah galat terdistribusi normal. Hipotesis dari uji normalitas adalah:

$$H_0 : \text{Residual berdistribusi normal}$$

$$H_1 : \text{Residual tidak berdistribusi normal}$$

Statistik uji yang digunakan adalah



$$A^2 = -n - \sum_{i=1}^n \frac{2i-1}{n} \left[ \ln F(Y_i) + \ln(1 - F(Y_{n+1-i})) \right]$$

$H_0$  ditolak pada taraf signifikansi  $\alpha$  jika  $p\text{-value} < \alpha$ .

### 2.2.5 Uji Autokorelasi

Uji autokorelasi atau uji Durbin-Watson dilakukan untuk memeriksa adanya korelasi antar galat pada waktu yang berbeda. Hipotesis dari uji autokorelasi adalah:

$H_0 : \rho = 0$  (tidak terdapat autokorelasi)

$H_1 : \rho \neq 0$  (terdapat autokorelasi)

Statistik uji yang digunakan adalah

$$d = \frac{\sum_{t=2}^{t=n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^n \hat{u}_t^2}$$

$H_0$  ditolak pada taraf signifikansi  $\alpha$  jika  $p\text{-value} < \alpha$ .

Dalam penelitian ini, regresi linier berganda digunakan untuk menganalisis pengaruh simultan dari variabel-variabel independen terhadap tingkat kemiskinan di Provinsi Jawa Timur. Namun, karena data yang digunakan bersifat spasial, yaitu data yang memiliki keterkaitan antar wilayah, model regresi linier berganda standar mungkin tidak dapat menangani efek spasial tersebut dengan baik. Oleh karena itu, diperlukan pendekatan regresi spasial yang dapat memperhitungkan keterkaitan antar wilayah dalam analisis.

## BAB 3

### Data dan Metode

#### 3.1 Penjelasan Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari Badan Pusat Statistik (BPS) Provinsi Jawa Timur untuk tahun 2024. Data ini mencakup tingkat kemiskinan (persentase penduduk miskin) dan beberapa variabel independen yang dipilih berdasarkan teori dan relevansi terhadap tingkat kemiskinan. Variabel-variabel yang digunakan adalah sebagai berikut.

Variabel	Keterangan	Jenis Variabel
Y	Persentase Penduduk Miskin	Dependen
X <sub>1</sub>	Gini Ratio	Independen
X <sub>2</sub>	Rata-rata Lama Sekolah (RLS)	Independen
X <sub>3</sub>	Jumlah Tenaga Medis	Independen
X <sub>4</sub>	Jumlah Kendaraan Bermotor (Sepeda Motor)	Independen

Data ini akan dianalisis dengan menggunakan tiga model analisis, yaitu regresi spasial SLM (*Spatial Lag Model*), regresi spasial SEM (*Spatial Error Model*), dan metode *Random Forest*, untuk memprediksi dan membandingkan faktor-faktor yang mempengaruhi tingkat kemiskinan di Provinsi Jawa Timur.

#### 3.2 Teori Uji Kesesuaian Kondisi Data

Sebelum melakukan analisis menggunakan model regresi spasial, perlu dilakukan uji kesesuaian terhadap kondisi data untuk memastikan bahwa data memenuhi asumsi-asumsi yang diperlukan. Salah satu uji yang digunakan dalam analisis data spasial adalah *Moran's I*.

### 3.2.1 Uji Moran's I

Moran's I digunakan untuk menguji adanya ketergantungan spasial dalam data. *Moran's I* mengukur apakah nilai variabel di suatu wilayah berhubungan dengan nilai variabel yang ada di wilayah tetangganya. *Moran's I* didefinisikan sebagai berikut:

$$I = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

dimana

- $x_i$  : nilai variabel pada lokasi ke- $i$
- $w_{ij}$  : elemen matriks pembobot spasial pada baris ke- $i$  kolom ke- $j$ .

Nilai *Moran's I* berkisar antara -1 hingga +1, dengan interpretasi sebagai berikut:

- Nilai *Moran's I* mendekati +1 menunjukkan adanya ketergantungan spasial positif (wilayah dengan kemiskinan tinggi cenderung berdekatan dengan wilayah lain yang juga memiliki tingkat kemiskinan tinggi).
- Nilai *Moran's I* mendekati -1 menunjukkan adanya ketergantungan spasial negatif (wilayah dengan kemiskinan tinggi berdekatan dengan wilayah yang memiliki kemiskinan rendah).
- Nilai *Moran's I* mendekati 0 menunjukkan tidak ada ketergantungan spasial.

Hipotesis dari uji *Moran's I* adalah:

$H_0 : I = 0$  (tidak ada autokorelasi antar lokasi)

$H_1 : I \neq 0$  (terdapat autokorelasi antar lokasi)

Statistik uji yang digunakan adalah

$$Z(I) = \frac{I - E(I)}{\sqrt{Var(I)}}$$

$H_0$  ditolak pada taraf signifikansi  $\alpha$  jika  $|Z(I)| > Z_{\alpha/2}$ .

Jika hasil uji Moran's I menunjukkan ketergantungan spasial yang signifikan, maka model regresi spasial seperti SLM atau SEM lebih tepat digunakan dibandingkan regresi linier biasa.

## 3.3 Pengujian Efek Spasial

### 3.3.1 Keragaman Spasial

Keragaman spasial mengacu pada adanya variasi atau perbedaan yang signifikan antar lokasi atau wilayah pengamatan. Untuk mendeteksi adanya keragaman spasial dalam data,

salah satu metode yang digunakan adalah uji Breusch-Pagan. Uji ini menguji apakah varians antar wilayah bersifat homogen atau terdapat perbedaan yang signifikan. Hipotesis yang digunakan dalam uji ini adalah:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_i^2 = \sigma^2 \text{ (variens antar lokasi adalah sama)}$$

$$H_1 : \sigma_i^2 \neq \sigma^2 \text{ (terdapat minimal satu variens yang berbeda)}$$

Statistik uji untuk Breusch-Pagan dapat dihitung dengan rumus:

$$BP = \frac{1}{2} f^T Z (Z^T Z)^{-1} Z^T f$$

dimana  $f_i$  adalah nilai residual kuadrat dari pengamatan dikurangi nilai rata-rata residual.

Uji ini akan menolak hipotesis nol jika nilai statistik BP lebih besar dari nilai kritis  $\chi_p^2$ , yang merupakan distribusi chi-square dengan derajat kebebasan p.

### 3.3.2 Lagrange Multiplier Lag (LM Lag)

Untuk mendeteksi adanya ketergantungan spasial pada variabel dependen (respon), uji *Lagrange Multiplier Lag* dapat digunakan. Uji ini bertujuan untuk mengidentifikasi apakah variabel dependen pada satu wilayah dipengaruhi oleh nilai variabel dependen pada wilayah tetangga. Hipotesis yang digunakan adalah:

$$H_0 : \rho = 0 \text{ (tidak ada ketergantungan spasial pada variabel dependen)}$$

$$H_1 : \rho \neq 0 \text{ (terdapat ketergantungan spasial pada variabel dependen)}$$

Statistik uji untuk *Lagrange Multiplier Lag* dapat dihitung dengan rumus:

$$LM_{\rho} = \frac{\left[ \frac{e' W Y}{\left( \frac{e e'}{n} \right)} \right]^2}{D}$$

dengan

$$D = \left[ \frac{(W X \beta)' (1 - X (X' X)^{-1} X') (W X \beta)}{\sigma^2} \right] + tr(W' W + W W)$$

dimana W adalah matriks bobot spasial, Y adalah vektor variabel dependen, dan e adalah residual. Uji ini akan menolak hipotesis nol jika nilai  $LM_{\rho}$  lebih besar dari nilai kritis  $\chi_{\alpha(p)}^2$ . Jika hipotesis nol ditolak, model yang lebih cocok untuk data adalah *Spatial Autoregressive Model* (SAR). Jika hipotesis nol diterima, ini menunjukkan bahwa tidak

ada ketergantungan spasial signifikan dalam data, dan model yang lebih cocok adalah *Spatial Lag Model* (SLM).

### 3.2.3 Lagrange Multiplier Error (LM Error)

Selain ketergantungan spasial pada variabel dependen, pengujian juga perlu dilakukan untuk mengidentifikasi ketergantungan spasial pada galat (residual). Uji *Lagrange Multiplier Error* digunakan untuk tujuan ini. Hipotesis yang digunakan adalah:

$H_0 : \lambda = 0$  (tidak ada ketergantungan spasial pada galat)

$H_1 : \lambda \neq 0$  (terdapat ketergantungan spasial pada galat)

Statistik uji untuk *Lagrange Multiplier Error* dapat dihitung dengan rumus:

$$LM_{\lambda} = \frac{\left[ \frac{e'W e}{\left( \frac{e'e}{n} \right)} \right]^2}{tr(W'W + WW)}$$

Uji ini akan menolak hipotesis nol jika nilai statistik  $LM_{\lambda}$  lebih besar dari nilai kritis  $\chi^2(p)$ . Jika  $H_0$  ditolak, maka ada ketergantungan spasial pada galat, dan model yang lebih cocok adalah *Spatial Error Model* (SEM).

### 3.3 Spatial Lag Model (SLM)

Model *Spatial Lag Model* (SLM) digunakan untuk mengatasi ketergantungan spasial antar wilayah. Dalam model ini, variabel dependen pada suatu wilayah dipengaruhi oleh variabel dependen di wilayah tetangga, serta variabel-variabel independen di wilayah tersebut. Model SLM dapat dinyatakan sebagai berikut:

$$y = \rho WY + X\beta + \varepsilon$$

$$\varepsilon \sim NIID(0, \sigma^2 I_n)$$

dimana:

- Y adalah vektor variabel dependen
- $\rho$  adalah parameter lag spasial
- W adalah matriks bobot spasial (mewakili keterkaitan antar wilayah)
- X adalah matriks variabel independen
- $\beta$  adalah vektor koefisien untuk variabel independen
- $\varepsilon$  adalah galat

Pendugaan parameter  $\rho$  dan  $\beta$  dilakukan dengan menggunakan metode *Maximum Likelihood* (ML) atau metode klasik OLS setelah penyesuaian untuk ketergantungan spasial. Uji hipotesis dilakukan untuk mengetahui apakah koefisien  $\beta$  signifikan.

### 3.4 *Spatial Error Model* (SEM)

Model *Spatial Error Model* (SEM) digunakan untuk menangani ketergantungan spasial yang muncul pada galat. Dalam model ini, galat di suatu wilayah dipengaruhi oleh galat di wilayah tetangga. Model SEM dapat dinyatakan sebagai:

$$\begin{aligned}y &= X\beta + u \\u &= \lambda Wu + \varepsilon \\ \varepsilon &\sim NIID(0, \sigma^2 I_n)\end{aligned}$$

dimana:

- $\lambda$  adalah parameter yang menunjukkan pengaruh galat dari wilayah tetangga
- $Wu$  adalah matriks galat spasial
- $\varepsilon$  adalah galat yang tidak berkorelasi

Pendugaan parameter dilakukan dengan menggunakan metode *Maximum Likelihood* (ML) atau metode klasik OLS setelah penyesuaian untuk galat spasial. Uji hipotesis untuk model ini bertujuan untuk menentukan apakah variabel independen  $\beta$  berpengaruh signifikan terhadap tingkat kemiskinan.

### 3.5 *Random Forest*

Model *Random Forest* digunakan untuk menangani kompleksitas dan non-linieritas dalam data spasial. *Random Forest* adalah algoritma berbasis *ensemble* yang membangun banyak pohon keputusan dan menggabungkan hasilnya untuk menghasilkan prediksi yang lebih stabil dan akurat. Model ini tidak memerlukan asumsi distribusi tertentu dan dapat menangani data dengan banyak variabel independen.

*Random Forest* bekerja dengan membagi data menjadi beberapa subset dan membangun pohon keputusan untuk setiap subset. Hasil dari pohon-pohon tersebut

digabungkan untuk memberikan hasil prediksi yang lebih baik. Model ini sangat berguna untuk menangani data yang tidak linier dan memiliki hubungan kompleks antara variabel.

Pengujian hipotesis dalam *Random Forest* dilakukan dengan melihat fitur-fitur yang paling berpengaruh dalam membentuk keputusan pohon seperti misalnya menggunakan *feature importance*.

### 3.6 Penentuan Model Terbaik

Pemilihan model yang terbaik dalam analisis regresi dilakukan dengan mempertimbangkan beberapa kriteria statistik yang digunakan untuk menilai sejauh mana model dapat menggambarkan data dengan baik. Beberapa indikator utama yang digunakan dalam evaluasi ini adalah *Akaike Information Criterion* (AIC), Koefisien Determinasi ( $R^2$ ), dan *Bayesian Information Criterion* (BIC).

#### 3.6.1 Akaike Information Criterion (AIC)

*Akaike Information Criterion* (AIC) adalah suatu ukuran yang digunakan untuk membandingkan kualitas relatif dari beberapa model statistik. AIC mempertimbangkan keseimbangan antara bagaimana baiknya model tersebut mencocokkan data (fit) dengan kompleksitas model, yang dihitung berdasarkan jumlah parameter yang digunakan. Semakin rendah nilai AIC, semakin baik model tersebut dalam hal *trade-off* antara kesesuaian dan kompleksitas. Persamaan AIC juga dapat dihitung dengan rumus berikut:

$$AIC = -2 \ln(\hat{L}) + 2k$$

dimana:

- $\hat{L}$  adalah nilai *likelihood* maksimum dari model
- $k$  adalah jumlah parameter dalam model, termasuk *intercept*

#### 3.6.2 Koefisien Determinasi ( $R^2$ )

Koefisien Determinasi ( $R^2$ ) adalah ukuran yang menggambarkan sejauh mana variabilitas dalam variabel dependen ( $Y$ ) dapat dijelaskan oleh variabel independen ( $X$ ) dalam model. Nilai  $R^2$  berkisar antara 0 hingga 1, dimana semakin mendekati 1, semakin baik model dalam menjelaskan variasi data.  $R^2$  dapat dihitung dengan menggunakan rumus:

$$R = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

dimana:

- $\hat{Y}_i$  adalah nilai prediksi dari model regresi
- $\bar{Y}$  adalah nilai rata-rata dari Y dalam data

Nilai  $R^2$  yang lebih tinggi menunjukkan bahwa model lebih mampu menjelaskan variabilitas data.

### 3.6.3 Bayesian Information Criterion (BIC)

Selain AIC, Bayesian Information Criterion (BIC) juga digunakan sebagai salah satu kriteria untuk memilih model terbaik. BIC mengukur keseimbangan antara kebaikan model dalam mencocokkan data dengan jumlah parameter yang digunakan dalam model, dengan lebih menekankan penalti yang lebih kuat terhadap kompleksitas model dibandingkan dengan AIC. Semakin rendah nilai BIC, semakin baik model tersebut.

Persamaan untuk menghitung BIC adalah sebagai berikut:

$$BIC = -2 \ln(\hat{L}) + k \ln(n)$$

dimana:

- $\hat{L}$  adalah nilai *likelihood* maksimum dari model
- $k$  adalah jumlah parameter dalam model
- $n$  adalah jumlah observasi dalam data

Nilai BIC yang lebih rendah menunjukkan model yang lebih baik dengan penalti yang lebih sedikit terhadap kompleksitas model. BIC seringkali digunakan ketika jumlah data sangat besar, karena penalti terhadap parameter model lebih kuat dari AIC, sehingga cenderung memilih model yang lebih sederhana.



## BAB 4

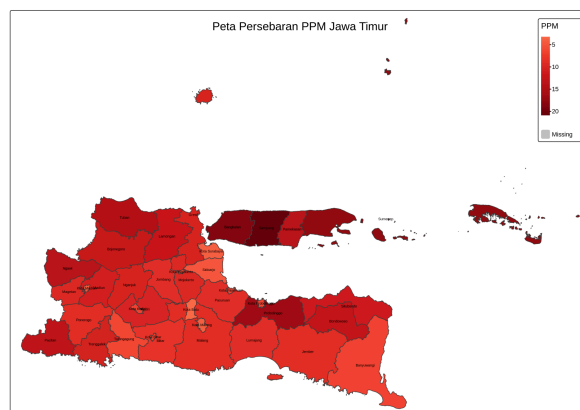
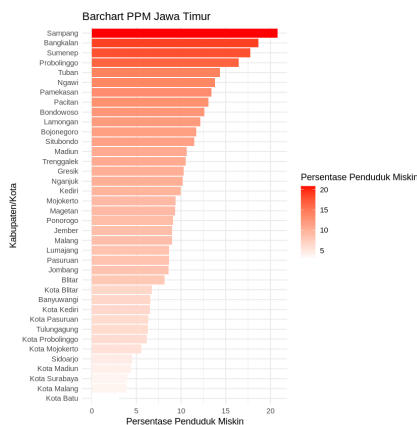
### Hasil dan Pembahasan

Bab ini menyajikan hasil analisis data yang telah dilakukan, dimulai dari statistik deskriptif variabel penelitian, hasil uji dependensi spasial yang diperoleh dari diagnostik model OLS, hasil analisis dari ketiga model regresi (OLS, SLM, dan SEM), hingga perbandingan antar model untuk menentukan model terbaik beserta analisis residualnya.

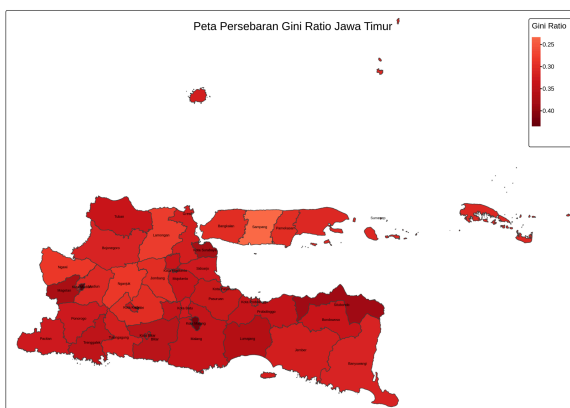
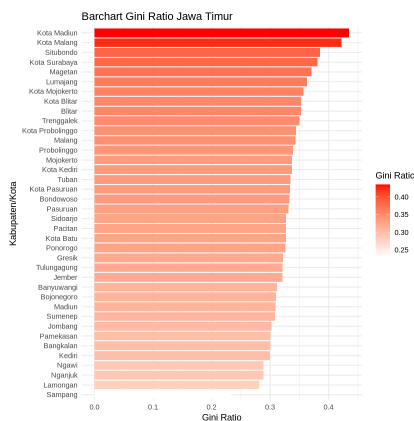
#### 4.1 Statistik Deskriptif

Penelitian ini menggunakan data dari 38 observasi (misalnya, kabupaten/kota) setelah proses pembersihan data dimana tidak ada data yang perlu dihapus karena nilai kosong atau tak terhingga (NaN/Inf). (Ini berdasarkan output "Data awal dimuat: 38 baris", dan "Data setelah menghapus NaN: 38 baris").

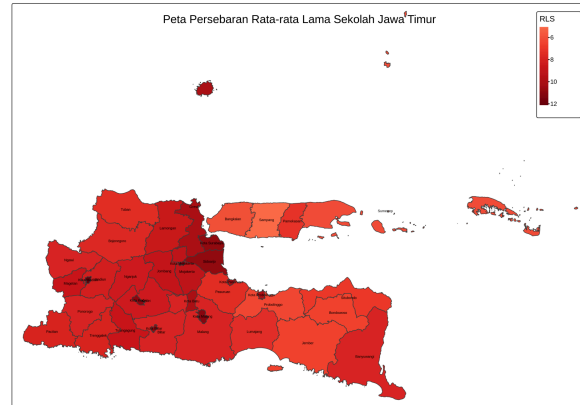
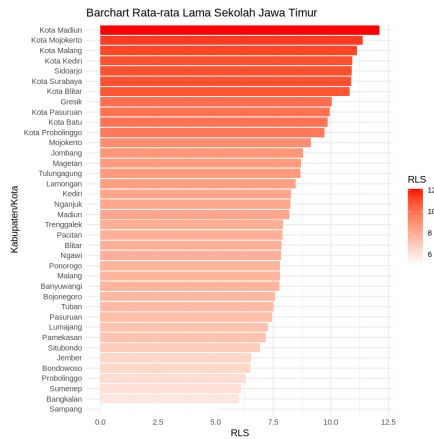
Untuk Variabel Y



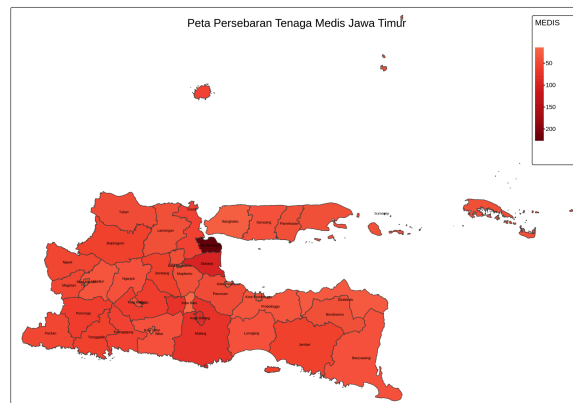
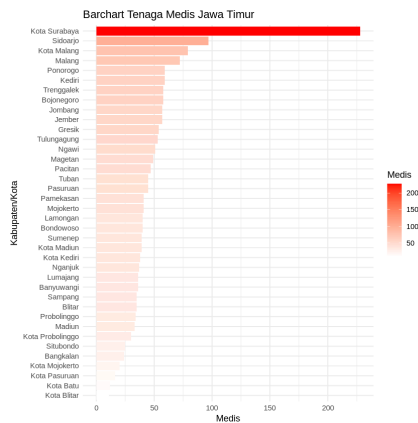
Untuk Variabel X1



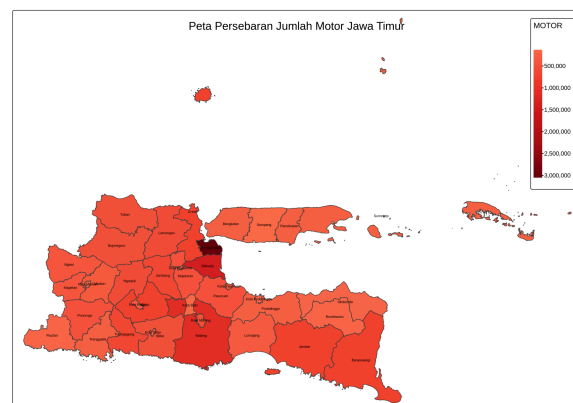
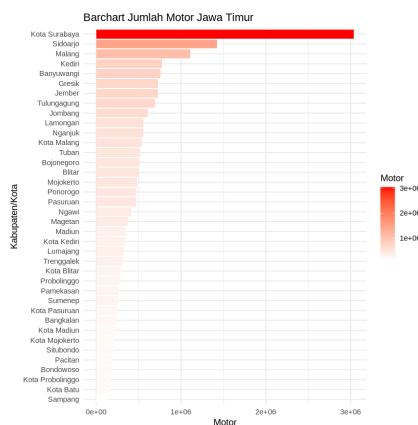
## Untuk Variabel X2



## Untuk Variabel X3



## Untuk Variabel X4



Analisis deskriptif dilakukan untuk memperoleh gambaran awal mengenai sebaran data variabel yang digunakan, yaitu persentase kemiskinan (Y), Gini Ratio (X1), Rata Lama Sekolah (X2), Jumlah Tenaga Medis (X3), dan Jumlah Kendaraan Bermotor (X4).

	Y	GiniRatio	RLS	MEDIS	MOTOR
count	38.000000	38.000000	38.000000	38.000000	3.800000e+01
mean	9.782368	0.331816	8.463947	48.157895	5.161917e+05
std	4.213447	0.037126	1.703252	34.714688	4.988880e+05
min	3.060000	0.233000	5.080000	11.000000	1.403630e+05
25%	6.517500	0.309250	7.477500	35.000000	2.515750e+05
50%	9.215000	0.329000	8.060000	40.500000	3.949030e+05
75%	12.042500	0.348500	9.832500	56.250000	5.550472e+05
max	20.830000	0.435000	12.110000	228.000000	3.034754e+06

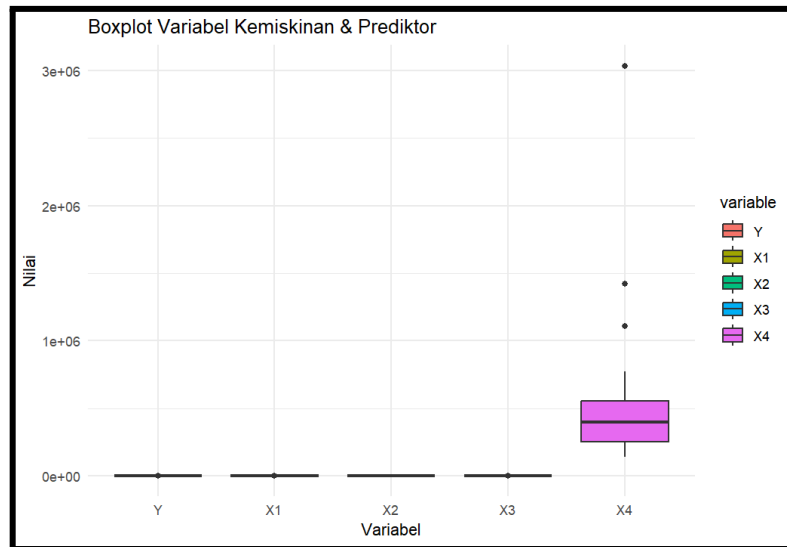
**Gambar 4.1. Statistika Deskriptif**

Berdasarkan **Gambar 4.1**, penelitian ini menggunakan data dari 38 observasi yang telah melalui proses pembersihan data. Variabel dependen Y, yang merepresentasikan tingkat kemiskinan, memiliki rata-rata 9.78% dengan standar deviasi 4.21, menunjukkan variasi yang cukup antar wilayah. Nilai tingkat kemiskinan ini berkisar antara 3.06% hingga 20.83%. Variabel GiniRatio memiliki rata-rata 0.33, mengindikasikan tingkat ketimpangan pendapatan rata-rata pada periode observasi di seluruh wilayah studi, dengan sebaran nilai dari 0.233 hingga 0.435.

Selanjutnya, variabel Rata-rata Lama Sekolah (RLS) menunjukkan nilai rata-rata 8.46 tahun, yang secara umum setara dengan tingkat pendidikan Sekolah Menengah Pertama (SMP). Sebaran nilai RLS tidak terlalu lebar, dengan nilai terendah 5.08 tahun (setara Sekolah Dasar) dan tertinggi mencapai 12.11 tahun (setara Sekolah Menengah Atas).

Untuk variabel MEDIS, yang kemungkinan merepresentasikan ketersediaan tenaga medis, ditemukan rata-rata sebesar 48.16. Namun, standar deviasi yang sangat besar (34.71) serta rentang nilai yang sangat lebar dari 11 hingga 228 menunjukkan adanya ketimpangan yang signifikan dalam ketersediaan fasilitas atau tenaga kesehatan antar wilayah observasi.

Terakhir, variabel MOTOR (jumlah kendaraan bermotor), yang dapat menjadi proksi untuk tingkat kesejahteraan dan aktivitas ekonomi, rata-rata berjumlah sekitar 516.191 unit per wilayah. Variabel ini juga memiliki standar deviasi yang sangat tinggi (498.888), hampir sama dengan nilai rata-ratanya. Hal ini mengindikasikan variasi yang ekstrem antar wilayah, mulai dari yang hanya memiliki sekitar 140 ribu unit hingga wilayah yang memiliki lebih dari 3 juta unit kendaraan bermotor."



**Gambar 4.2. Box Plot**

**Gambar 4.2** menunjukkan boxplot dari seluruh variabel. Terlihat bahwa variabel X4 (Jumlah Kendaraan Bermotor) memiliki skala nilai yang jauh lebih besar dibandingkan variabel lain, dengan nilai maksimum mencapai lebih dari 3 juta unit. Sementara itu, variabel Y, X1, X2, dan X3 memiliki sebaran yang relatif sempit, yang tercermin dari boxplot yang lebih rapat. Selain itu, terdapat beberapa outlier pada variabel X3 dan X4, yang menunjukkan adanya kabupaten dengan nilai yang jauh lebih tinggi daripada kabupaten lainnya.

## 4.2 Hasil Uji Dependensi Spasial

Untuk analisis spasial, matriks pembobot spasial (W) dikonstruksi menggunakan metode Queen. Jumlah observasi dalam matriks W adalah 38, sesuai dengan jumlah data yang dianalisis.

Analisis ini bertujuan untuk menentukan apakah ada efek spasial (ketergantungan antar lokasi) dalam data Anda. Jika ada, model OLS (Ordinary Least Squares) biasa tidak lagi menjadi model terbaik, dan kita perlu menggunakan model regresi spasial seperti Spatial Lag Model (SLM), Spatial Error Model (SEM).

### 1. Uji Asumsi Normalitas Error

TEST ON NORMALITY OF ERRORS			
TEST	DF	VALUE	PROB
Jarque-Bera	2	0.713	0.7002

Karena  $p\text{-value} = 0.8954 > 0.05$ , maka tidak ada cukup bukti untuk menolak  $H_0$ . Artinya, residual model berdistribusi normal, yang merupakan salah satu asumsi penting regresi OLS.

## 2. Uji Asumsi Heteroskedastisitas

### DIAGNOSTICS FOR HETEROSKEDASTICITY

#### RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	4	1.745	0.7825
Koenker-Bassett test	4	2.586	0.6293

Karena  $p\text{-value} = 0.7825 > 0.05$ , maka tidak ada cukup bukti untuk menolak  $H_0$ . Artinya, residual model homoskedastis (varian residual konstan), sesuai dengan asumsi regresi klasik.

## 3. Uji Asumsi Multikolinearitas

```
[ ] 1 vif(lm(formula, data = df))
```

↗ GiniRatio: 1.64422800376622 RLS: 1.65922928119909 MEDIS: 8.6168568374322 MOTOR: 8.75864003463505

Nilai VIF untuk  $X_3$  dan  $X_4 > 5$ , bahkan mendekati 9, namun masih  $< 10$ , menunjukkan indikasi multikolinearitas masih bisa ditoleransi antara  $X_3$  dan  $X_4$  dengan variabel lain. Sedangkan  $X_1$  dan  $X_2$  memiliki  $VIF < 5$ , artinya tidak bermasalah. Multikolinearitas dapat menyebabkan koefisien regresi tidak stabil dan interpretasi menjadi sulit.

## 4. Uji Dependensi Spasial Umum (Untuk Model *Lag* vs. *Error*)

Uji ini membantu kita memutuskan antara model OLS, model *Spatial Lag* (SLM), atau model *Spatial Error* (SEM).

### a. Moran's I (variable)

Sebelum melakukan analisis regresi, perlu dilakukan pengujian untuk mendeteksi ada atau tidaknya dependensi spasial pada data. Uji autokorelasi spasial bertujuan untuk mengetahui apakah nilai suatu variabel di satu lokasi dipengaruhi oleh nilai variabel yang sama di lokasi-lokasi tetangganya. Dalam penelitian ini, uji yang digunakan adalah Indeks Moran (Moran's I) terhadap setiap variabel penelitian di 38 kabupaten/kota di Jawa Timur.

Hipotesis yang digunakan dalam uji Indeks Moran adalah sebagai berikut:

- **$H_0$ : Tidak terdapat autokorelasi spasial** (nilai Moran's I mendekati 0, pola sebaran acak).

- **H<sub>1</sub>: Terdapat autokorelasi spasial** (nilai Moran's I signifikan, pola sebaran mengelompok atau menyebar).

Kriteria keputusan didasarkan pada nilai probabilitas (p-value). Jika  $p\text{-value} < 0.05$ , maka  $H_0$  ditolak, yang berarti terdapat bukti statistik adanya autokorelasi spasial.

Hasil pengujian Indeks Moran untuk setiap variabel disajikan pada Tabel 4.1 di bawah ini.

**Tabel 4.1 Hasil Uji Indeks Moran untuk Setiap Variabel**

Variabel Penelitian	Nama Variabel	Moran's I Statistic	p-value	Interpretasi
Variabel Dependen	Tingkat Kemiskinan (Y)	0.1998	0.03846	Signifikan
Variabel Independen	Gini Ratio (X1)	0.1817	0.05200	Hampir Signifikan
Variabel Independen	Rata-rata Lama Sekolah (X2)	0.2857	0.00474	Signifikan
Variabel Independen	Tenaga Medis (X3)	0.1337	0.04277	Signifikan
Variabel Independen	Pengguna Kendaraan Bermotor (X4)	0.1984	0.00664	Signifikan

Berdasarkan hasil pada Tabel 4.1, dapat diinterpretasikan sebagai berikut:

1. **Tingkat Kemiskinan (Y)** : Diperoleh nilai p-value sebesar 0.03846, yang lebih kecil dari tingkat signifikansi  $\alpha = 0.05$ . Hal ini menunjukkan bahwa  $H_0$  ditolak, sehingga terdapat autokorelasi spasial yang signifikan pada variabel tingkat kemiskinan. Nilai Indeks Moran yang positif (0.1998) mengindikasikan adanya **pola mengelompok (cluster)**, di mana kabupaten/kota dengan persentase tingkat kemiskinan tinggi cenderung berdekatan secara geografis dengan kabupaten/kota lain yang juga memiliki tingkat kemiskinan tinggi, dan sebaliknya.
2. **Gini Ratio (X1)** : Diperoleh nilai p-value sebesar 0.052, yang lebih besar dari tingkat signifikansi  $\alpha = 0.05$ . Dengan demikian,  $H_0$  gagal ditolak, yang berarti **tidak terdapat bukti statistik yang cukup** untuk menyatakan adanya autokorelasi spasial

pada variabel Gini Ratio di Jawa Timur. Pola sebaran Gini Ratio antar wilayah cenderung acak.

3. **Rata-rata Lama Sekolah (X2)** : Dengan p-value sebesar 0.00474 ( $p < 0.05$ ), variabel ini menunjukkan adanya autokorelasi spasial positif yang sangat signifikan. Artinya, wilayah dengan rata-rata lama sekolah yang tinggi cenderung bertetangga dengan wilayah lain yang juga memiliki rata-rata lama sekolah yang tinggi.
4. **Tenaga Medis (X3)** : Diperoleh p-value sebesar 0.04277 ( $p < 0.05$ ), yang mengindikasikan adanya autokorelasi spasial positif yang signifikan. Pola sebaran jumlah tenaga medis di Jawa Timur bersifat mengelompok.
5. **Pengguna Kendaraan Bermotor (X4)** : Dengan p-value sebesar 0.00664 ( $p < 0.05$ ), variabel ini juga menunjukkan autokorelasi spasial positif yang signifikan, yang berarti ada pola pengelompokan spasial pada jumlah pengguna kendaraan bermotor.

#### Implikasi Hasil Uji

Hasil uji Indeks Moran menunjukkan bahwa variabel dependen (Tingkat Kemiskinan) dan sebagian besar variabel independen (Rata-rata Lama Sekolah, Tenaga Medis, dan Pengguna Kendaraan Bermotor) memiliki dependensi spasial. Adanya autokorelasi spasial ini melanggar salah satu asumsi penting dari model regresi linear klasik (OLS), yaitu independensi antar observasi. Penggunaan model OLS dalam kondisi ini dapat menghasilkan estimasi parameter yang tidak efisien dan kesimpulan yang keliru. Oleh karena itu, untuk mengakomodasi adanya efek spasial dalam data, analisis selanjutnya dalam penelitian ini perlu menggunakan **model regresi spasial**.

#### b. Moran's I (error)

##### 1. OLS Model

Moran I test under randomisation

```
data: model_ols$residuals
weights: ww
```

```
Moran I statistic standard deviate = 1.1151, p-value = 0.1324
alternative hypothesis: greater
sample estimates:
```

Moran I statistic	Expectation	Variance
0.09647716	-0.02702703	0.01226684

## 2. SLM Model

```
Moran I test under randomisation

data: model_slm$residuals
weights: ww

Moran I statistic standard deviate = 0.67997, p-value = 0.2483
alternative hypothesis: greater
sample estimates:
Moran I statistic      Expectation      Variance
      0.04823760      -0.02702703      0.01225178
```

## 3. SEM Model

```
Moran I test under randomisation

data: model_sem$residuals
weights: ww

Moran I statistic standard deviate = 0.19616, p-value = 0.4222
alternative hypothesis: greater
sample estimates:
Moran I statistic      Expectation      Variance
      -0.005281305      -0.027027027      0.012289860
```

Nilai probabilitas (p-value) ketiganya jauh lebih besar dari 0.05 (atau 0.10). Ini berarti kita **gagal menolak hipotesis nol**, yang menyatakan tidak ada autokorelasi spasial pada sisaan (error) ketiga model tersebut. Dengan kata lain, tidak ada bukti bahwa error di satu lokasi berhubungan dengan error di lokasi tetangganya.

```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
- SARERR -
TEST      MI/DF      VALUE      PROB
Moran's I (error)      -0.0140      0.297      0.7665
Lagrange Multiplier (lag)      1      0.246      0.6199
Robust LM (lag)      1      0.279      0.5977
Lagrange Multiplier (error)      1      0.013      0.9086
Robust LM (error)      1      0.046      0.8308
Lagrange Multiplier (SARMA)      2      0.292      0.8643
```

### c. Lagrange Multiplier (lag) & Robust LM (lag):

Prob LM (lag) : 0.6199

Prob Robust LM (lag) : 0.5977

Interpretasi : Kedua p-value ini sangat besar ( $> 0.05$ ). Ini menunjukkan **tidak ada bukti statistik yang signifikan** untuk menggunakan *Spatial Lag Model* (SLM). Model SLM



mengasumsikan bahwa nilai variabel dependen (Y) di satu lokasi dipengaruhi oleh nilai Y di lokasi tetangganya.

d. Lagrange Multiplier (error) & Robust LM (error):

Prob LM (error) : 0.9086

Prob Robust LM (error) : 0.8308

Interpretasi : Sama seperti uji lag, kedua p-value ini juga sangat besar ( $> 0.05$ ). Ini menunjukkan **tidak ada bukti statistik yang signifikan** untuk menggunakan *Spatial Error Model* (SEM). Model SEM mengasumsikan adanya korelasi spasial pada galat/error yang tidak teramati.

Kesimpulan Sementara dari Bagian 1: Berdasarkan uji LM klasik dan robust, tidak ada dependensi spasial yang terdeteksi. Aturan standar akan menyarankan bahwa **model OLS sudah cukup** dan tidak perlu beralih ke model SLM atau SEM.

### 4.3 Hasil Analisis dari Ketiga Model

1. Model OLS (Ordinary Least Squares):

```
Call:
lm(formula = formula, data = df)

Residuals:
    Min       1Q   Median       3Q      Max
-4.2188 -1.3762 -0.0764  1.1948  4.3407

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.263e+01  3.316e+00   9.840 2.42e-11 ***
GiniRatio    -3.163e+01  1.265e+01  -2.500  0.0176 *
RLS          -1.508e+00  2.770e-01  -5.445 4.96e-06 ***
MEDIS         6.931e-02  3.097e-02   2.238  0.0321 *
MOTOR        -5.665e-06  2.173e-06  -2.607  0.0136 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.228 on 33 degrees of freedom
Multiple R-squared:  0.7506,    Adjusted R-squared:  0.7203
F-statistic: 24.83 on 4 and 33 DF,  p-value: 1.5e-09
```

**Gambar 4.3.b. Tabel Hasil OLS**

Model ini bertujuan untuk menjelaskan bagaimana variabel independen (**GiniRatio**, **RLS**, **MEDIS**, **MOTOR**) secara bersama-sama mempengaruhi sebuah variabel dependen.

1. Evaluasi Keseluruhan Model

### F-statistic & p-value:

- Nilai : Nilai **F-statistik** sebesar **24.83** dengan **p-value** yang sangat kecil (**1.5e-09**, atau 0.0000000015).
- **Interpretasi:** P-value ini jauh lebih kecil dari tingkat signifikansi manapun (misalnya, 0.05 atau 0.01). Ini berarti model ini secara keseluruhan sangat signifikan secara statistik. Dengan kata lain, setidaknya ada satu variabel independen yang memiliki pengaruh signifikan terhadap variabel dependen. Variabel-variabel independen secara kolektif mampu menjelaskan variasi pada variabel dependen.

### R-squared & Adjusted R-squared:

- **Multiple R-squared:** Nilainya **0.7506**. Ini berarti sekitar **75.06%** dari total variasi pada variabel dependen dapat dijelaskan oleh variabel-variabel independen (**GiniRatio**, **RLS**, **MEDIS**, **MOTOR**) dalam model ini. Ini menunjukkan daya penjelas model yang kuat.
- **Adjusted R-squared:** Nilainya **0.7203**. Nilai ini sedikit lebih rendah karena telah disesuaikan dengan jumlah variabel dalam model. Tetap saja, kemampuan model untuk menjelaskan **72.03%** variasi adalah sangat baik.

## 2. Interpretasi Koefisien Masing-Masing Variabel

Tabel **Coefficients** menunjukkan pengaruh individu dari setiap variabel, dengan asumsi variabel lain konstan.

- **(Intercept):** Nilai estimasinya **32.63** dan sangat signifikan (p-value = 2.42e-11). Ini adalah nilai prediksi dari variabel dependen ketika semua variabel independen bernilai nol.
- **GiniRatio:**
  - **Koefisien:** -31.63
  - **Signifikansi:** Signifikan (p-value = 0.0176, ditandai \*).
  - **Interpretasi:** Setiap kenaikan satu unit pada **GiniRatio** akan **menurunkan** nilai variabel dependen sebesar **31.63 unit**, dengan asumsi variabel lain tidak berubah.
- **RLS (Rata-rata Lama Sekolah):**
  - **Koefisien:** -1.508
  - **Signifikansi:** Sangat Signifikan (p-value = 4.96e-06, ditandai \*\*\*).
  - **Interpretasi:** Setiap kenaikan satu unit pada **RLS** akan **menurunkan** nilai variabel dependen sebesar **1.508 unit**, dengan asumsi variabel lain tidak berubah.
- **MEDIS:**

- **Koefisien:** 0.06931
- **Signifikansi:** Signifikan (p-value = 0.0321, ditandai \*).
- **Interpretasi:** Setiap kenaikan satu unit pada **MEDIS** akan **menaikkan** nilai variabel dependen sebesar **0.069 unit**, dengan asumsi variabel lain tidak berubah.
- **MOTOR:**
  - **Koefisien:** -5.665e-06 (atau -0.000005665)
  - **Signifikansi:** Signifikan (p-value = 0.0136, ditandai \*).
  - **Interpretasi:** Setiap kenaikan satu unit pada **MOTOR** akan **menurunkan** nilai variabel dependen sebesar **0.000005665 unit**. Meskipun signifikan secara statistik, pengaruhnya sangat kecil dan mendekati nol.

### 3. Persamaan Regresi

Berdasarkan nilai koefisien, persamaan regresi yang terbentuk adalah:

$$Y = 32.63 - 31.63(\text{GiniRatio}) - 1.508(\text{RLS}) + 0.0693(\text{MEDIS}) - 0.000005665(\text{MOTOR})$$

### Kesimpulan Akhir

Model OLS ini adalah model yang **baik dan signifikan secara statistik**. Model ini mampu menjelaskan sekitar 72-75% variasi dalam variabel dependen. Semua variabel independen yang dimasukkan (**GiniRatio**, **RLS**, **MEDIS**, **MOTOR**) terbukti memiliki pengaruh yang signifikan terhadap variabel dependen. **GiniRatio** dan **RLS** memiliki pengaruh negatif yang kuat, **MEDIS** memiliki pengaruh positif, sedangkan **MOTOR** memiliki pengaruh negatif yang signifikan secara statistik namun sangat kecil secara praktis.

### 2. Model Spatial Lag (SLM):

```

Call:lagsarlm(formula = formula, data = df, listw = ww)

Residuals:
    Min       1Q   Median       3Q      Max
-4.29567 -1.30049  0.11694  1.16984  3.93045

Type: lag
Coefficients: (asymptotic standard errors)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.5588e+01  3.4005e+00 10.4655 < 2.2e-16
GiniRatio    -3.3308e+01  1.1377e+01 -2.9277  0.003414
RLS          -1.6613e+00  2.7014e-01 -6.1498  7.757e-10
MEDIS         6.3486e-02  2.8242e-02  2.2480  0.024579
MOTOR        -5.0608e-06  2.0013e-06 -2.5288  0.011446

Rho: -0.031652, LR test value: 2.6049, p-value: 0.10653
Asymptotic standard error: 0.018098
      z-value: -1.7489, p-value: 0.080313
Wald statistic: 3.0586, p-value: 0.080313

Log likelihood: -80.38159 for lag model
ML residual variance (sigma squared): 4.0116, (sigma: 2.0029)
Number of observations: 38
Number of parameters estimated: 7
AIC: 174.76, (AIC for lm: 175.37)
LM test for residual autocorrelation
test value: 0.17527, p-value: 0.67547

```

#### Gambar 4.4. Hasil Model SLM (R)

Model ini menganalisis hubungan antara beberapa variabel independen (**GiniRatio**, **RLS**, **MEDIS**, **MOTOR**) dengan suatu variabel dependen, dengan mempertimbangkan adanya pengaruh spasial atau "efek tetangga".

Secara keseluruhan, **model SLM ini sedikit lebih baik daripada model regresi linear standar (OLS)**. Hal ini ditunjukkan oleh nilai AIC yang lebih rendah untuk model SLM (174.76) dibandingkan dengan AIC untuk model linear (175.37).

#### Interpretasi Koefisien Spasial (Rho)

- **Rho ( $\rho$ ):** Nilai estimasi Rho adalah **-0.031652** dengan *p-value* sebesar **0.080313**.
- **Makna:** Koefisien ini mengukur kekuatan dependensi spasial atau efek limpahan (*spillover*). Karena *p-value*-nya (0.08) lebih kecil dari tingkat signifikansi 10% ( $\alpha = 0.10$ ), kita dapat menyimpulkan bahwa **terdapat efek limpahan spasial yang negatif dan signifikan secara statistik**.
- **Interpretasi Praktis:** Adanya nilai Rho yang negatif menunjukkan bahwa nilai variabel dependen di suatu lokasi cenderung sedikit dipengaruhi secara negatif oleh nilai variabel dependen di lokasi-lokasi tetangganya. Dengan kata lain, jika nilai variabel dependen di wilayah tetangga tinggi, maka nilai di wilayah yang diamati cenderung sedikit lebih rendah, dan sebaliknya.

### Interpretasi Variabel Independen

Berikut adalah pengaruh masing-masing variabel independen terhadap variabel dependen, dengan mengasumsikan faktor lain konstan:

- **GiniRatio:** Memiliki koefisien **-3.3308e+01** (atau -33.31) dan sangat signifikan ( $p\text{-value} = 0.0034$ ). Ini berarti setiap kenaikan satu unit pada **GiniRatio** akan menurunkan nilai variabel dependen sekitar 33.31 unit.
- **RLS:** Memiliki koefisien **-1.6613e+00** (atau -1.66) dan sangat signifikan ( $p\text{-value} = 7.757e-10$ ). Ini berarti setiap kenaikan satu unit pada **RLS** (Rata-rata Lama Sekolah) akan menurunkan nilai variabel dependen sekitar 1.66 unit.
- **MEDIS:** Memiliki koefisien **6.3486e-02** (atau 0.063) dan signifikan ( $p\text{-value} = 0.0245$ ). Ini berarti setiap kenaikan satu unit pada **MEDIS** akan menaikkan nilai variabel dependen sekitar 0.063 unit.
- **MOTOR:** Memiliki koefisien **-5.0608e-06** dan signifikan ( $p\text{-value} = 0.0114$ ). Walaupun signifikan, nilai koefisiennya sangat kecil dan mendekati nol, menunjukkan pengaruh negatif yang sangat lemah.

### Diagnostik Model

- **LM test for residual autocorrelation:** Uji ini digunakan untuk mendeteksi apakah masih ada autokorelasi spasial yang tersisa di dalam sisaan (residuals) model.
- **Hasil:**  $p\text{-value}$  dari uji ini sangat tinggi, yaitu **0.67547**.
- **Kesimpulan:** Karena  $p\text{-value} > 0.10$ , kita tidak menolak hipotesis nol. Ini berarti **tidak ada lagi autokorelasi spasial yang signifikan dalam sisaan**. Hal ini adalah pertanda baik, yang menunjukkan bahwa model SLM telah berhasil menangani dependensi spasial yang ada di dalam data.

### 3. Model Spatial Error (SEM):

```

Call:errorsarlm(formula = formula, data = df, listw = ww)

Residuals:
    Min       1Q   Median       3Q      Max
-4.191796 -1.501315 -0.035765  1.132801  3.706161

Type: error
Coefficients: (asymptotic standard errors)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.2422e+01  3.2303e+00 10.0370 < 2.2e-16
GiniRatio    -2.5515e+01  1.2571e+01 -2.0296  0.04239
RLS          -1.6707e+00  2.7546e-01 -6.0652  1.318e-09
MEDIS        5.3759e-02  2.9571e-02  1.8179  0.06907
MOTOR        -4.6618e-06  2.0835e-06 -2.2375  0.02525

Lambda: 0.075242, LR test value: 1.0575, p-value: 0.30379
Asymptotic standard error: 0.049041
      z-value: 1.5342, p-value: 0.12497
Wald statistic: 2.3539, p-value: 0.12497

Log likelihood: -81.15528 for error model
ML residual variance (sigma squared): 4.098, (sigma: 2.0244)
Number of observations: 38
Number of parameters estimated: 7
AIC: 176.31, (AIC for lm: 175.37)

```

**Gambar 4.5. Hasil Model SEM (R)**

Model ini adalah *Spatial Error Model* (SEM), yang digunakan untuk mengatasi masalah autokorelasi spasial pada galat/error (*error term*) dari model regresi. Model ini mengasumsikan bahwa faktor-faktor tak teramati (yang masuk dalam galat) di satu lokasi saling berhubungan dengan faktor-faktor tak teramati di lokasi tetangganya.

#### Interpretasi Koefisien Spasial (Lambda)

- **Lambda ( $\lambda$ ):** Nilai estimasi Lambda adalah **0.075242** dengan *p-value* sebesar **0.12497** (dari z-value dan Wald statistic).
- **Makna:** Koefisien ini mengukur kekuatan autokorelasi spasial pada galat. Karena *p-value* (0.125) lebih besar dari tingkat signifikansi umum ( $\alpha = 0.10$  atau 0.05), kita dapat menyimpulkan bahwa **koefisien Lambda tidak signifikan secara statistik**.
- **Interpretasi Praktis:** Tidak ditemukan bukti yang cukup kuat untuk menyatakan adanya dependensi spasial pada galat model. Ini berarti bahwa galat atau variasi yang tidak dapat dijelaskan oleh variabel independen di satu lokasi tidak secara signifikan berhubungan dengan galat di lokasi-lokasi tetangganya.

#### Interpretasi Variabel Independen

Koefisien dalam model SEM diinterpretasikan seperti pada model regresi OLS standar, yaitu menunjukkan pengaruh langsung variabel independen terhadap variabel dependen di lokasi yang sama.

- **(Intercept):** Konstanta model signifikan secara statistik.
- **GiniRatio:** Memiliki koefisien **-2.5515e+01** (atau -25.52) dan signifikan pada tingkat 5% ( $p\text{-value} = 0.04239$ ). Artinya, setiap kenaikan satu unit pada **GiniRatio** berhubungan dengan penurunan nilai variabel dependen sebesar 25.52 unit.
- **RLS:** Memiliki koefisien **-1.6707e+00** (atau -1.67) dan sangat signifikan ( $p\text{-value} = 1.318\text{e-}09$ ). Artinya, setiap kenaikan satu unit pada **RLS** (Rata-rata Lama Sekolah) berhubungan dengan penurunan nilai variabel dependen sebesar 1.67 unit.
- **MEDIS:** Memiliki koefisien **5.3759e-02** (atau 0.054) dan signifikan pada tingkat 10% ( $p\text{-value} = 0.06907$ ). Artinya, setiap kenaikan satu unit pada **MEDIS** berhubungan dengan kenaikan nilai variabel dependen sebesar 0.054 unit.
- **MOTOR:** Memiliki koefisien **-4.6618e-06** dan signifikan pada tingkat 5% ( $p\text{-value} = 0.02525$ ). Nilai koefisien ini sangat kecil dan mendekati nol, menunjukkan pengaruh negatif yang sangat lemah dari variabel **MOTOR** terhadap variabel dependen.

#### Evaluasi Model

- **Akaike Information Criterion (AIC):** Nilai AIC untuk model SEM ini adalah **176.31**.
- **Perbandingan Model:** Nilai ini dibandingkan dengan AIC untuk model linear standar (ditulis sebagai **AIC for lm: 175.37**).
- **Kesimpulan:** Nilai AIC untuk model SEM (**176.31**) sedikit lebih tinggi daripada model linear (**175.37**). Dalam pemilihan model, nilai AIC yang lebih rendah menunjukkan model yang lebih baik. Oleh karena itu, berdasarkan kriteria AIC, model regresi linear standar (OLS) sedikit lebih unggul daripada model SEM untuk data ini.

#### Kesimpulan Akhir

Meskipun beberapa variabel independen menunjukkan pengaruh yang signifikan terhadap variabel dependen, **model SEM ini tampaknya tidak memberikan perbaikan yang berarti dibandingkan model regresi linear biasa**. Hal ini didukung oleh dua temuan utama:

1. Koefisien spasial (Lambda) tidak signifikan.
2. Nilai AIC model SEM lebih tinggi daripada AIC model linear.

Ini menunjukkan bahwa dependensi spasial dalam galat bukanlah masalah yang signifikan dalam dataset ini, dan model yang lebih sederhana (regresi OLS) mungkin sudah cukup untuk menjelaskan hubungan antar variabel.

#### 4. Model Spatial Random Forest:

Model Random Forest spasial diestimasi menggunakan fungsi `rf_spatial()` pada data sebanyak 38 kabupaten/kota di Provinsi Jawa Timur. Model ini dilatih dengan 500 decision trees dan parameter `mtry` sebesar 2.

**Table 4.1**

Ukuran Kebaikan Model	Nilai
R-squared (OOB)	0.6534121
R-squared (Prediksi)	0.9525424
Pseudo R-squared	0.9759828
RMSE (OOB)	2.480529
RMSE	1.116
Normalized RMSE	0.201991

Source: RStudio

Hasil analisis menunjukkan bahwa nilai pseudo R-squared mencapai 0.9759. Angka ini menunjukkan bahwa model Random Forest mampu menjelaskan hampir 98% variasi persentase kemiskinan di wilayah ini. Nilai Root Mean Squared Error (RMSE) yang diperoleh adalah sebesar 1.116, yang tergolong rendah dibandingkan dengan nilai rata-rata kemiskinan yang berada pada kisaran 9.78.

```
- Normality:
  - Shapiro-wilks W: 0.974
  - p-value       : 0.5099
  - Interpretation : Residuals are normal
```

**Gambar 4.6**



Analisis residual menunjukkan bahwa residual model berdistribusi normal, dengan nilai uji Shapiro-Wilk W sebesar 0.974 dan p-value sebesar 0.5099. Hal ini mengindikasikan tidak adanya penyimpangan signifikan dari normalitas pada residual. Selain itu, uji autokorelasi spasial residual menggunakan Moran's I juga menunjukkan tidak adanya autokorelasi spasial yang signifikan, dengan nilai p-value yang seluruhnya lebih besar dari 0.26 pada berbagai threshold jarak. Dengan demikian, tidak ditemukan pola spasial yang konsisten pada residual model Random Forest ini.

**Table 4.2**

<b>Variable</b>	<b>Importance</b>
X2	3.762
X1	1.448
X4	1.370
X3	0.794

Source: RStudio

Berdasarkan analisis variable importance (Gini index), variabel yang paling berpengaruh terhadap kemiskinan adalah rata-rata lama sekolah (X2), dengan nilai importance sebesar 3.762. Variabel berikutnya yang juga berpengaruh adalah Gini Ratio (X1), Jumlah Kendaraan Bermotor (X4), dan Jumlah Tenaga Medis (X3). Hasil ini menunjukkan bahwa rata-rata lama sekolah menjadi faktor utama yang menentukan variasi tingkat kemiskinan di Provinsi Jawa Timur.

Secara keseluruhan, model Random Forest menunjukkan kinerja prediksi yang sangat baik dan unggul dibandingkan dengan model regresi spasial (OLS, SLM, dan SEM). Hasil ini juga menunjukkan bahwa Random Forest dapat menangkap hubungan kompleks antarvariabel tanpa menyisakan pola spasial yang signifikan pada residual, sekaligus memberikan informasi mengenai pentingnya masing-masing variabel prediktor.

#### 4.4 Perbandingan Model

[1] "--- Perbandingan Kebaikan Model ---"

	Model	AIC	BIC	LogLikelihood	R_squared	p-value vs OLS
1	OLS	175.3681	185.1936	-81.68403	0.7505786	NA
2	Spatial Lag (SLM)	174.7632	186.2263	-80.38159	0.7505786	0.1065349
3	Spatial Error (SEM)	176.3106	187.7737	-81.15528	0.7505786	0.3037852

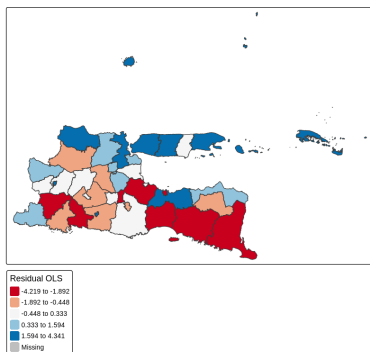
**Gambar 4.7. Tabel Perbandingan Model**

Berdasarkan tabel perbandingan, model OLS memiliki nilai AIC terendah (175.3681) dan BIC terendah (185.1936). Model SLM memiliki nilai AIC (174.7632) dan BIC (186.2263) yang cukup dekat dengan OLS dan juga SEM. Model SLM memiliki nilai AIC dan BIC yang lebih tinggi. Akan dipilih OLS sebagai model terbaik berdasarkan BIC terendah.

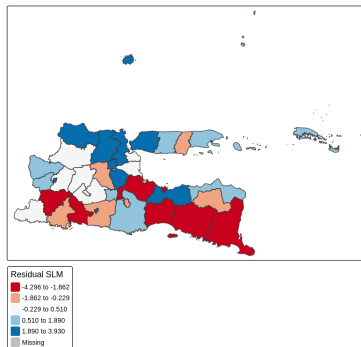
Mengingat:

1. Uji dependensi spasial pada residual OLS (Moran's I dan LM tests) tidak menunjukkan adanya autokorelasi spasial yang signifikan,
  2. Koefisien spasial  $\rho$  pada model SLM tidak signifikan, dan
  3. Koefisien spasial  $\lambda$  pada model SEM juga tidak signifikan,
- maka dapat disimpulkan bahwa efek spasial tidak memainkan peran penting dalam model ini. Cukup bisa dikatakan wajar saja jika OLS memiliki AIC/BIC sedikit lebih baik. Model OLS lebih parsimonious (sederhana) dan memberikan interpretasi yang serupa untuk koefisien variabel independen. Oleh karena itu, model OLS dapat dianggap sebagai model yang paling tepat dan memadai untuk data ini.

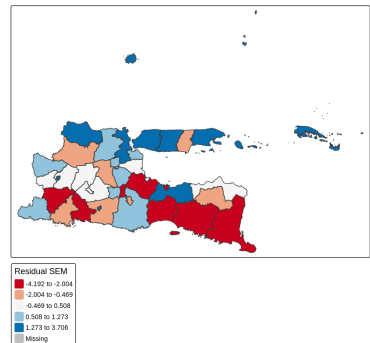
Peta Residual Model OLS



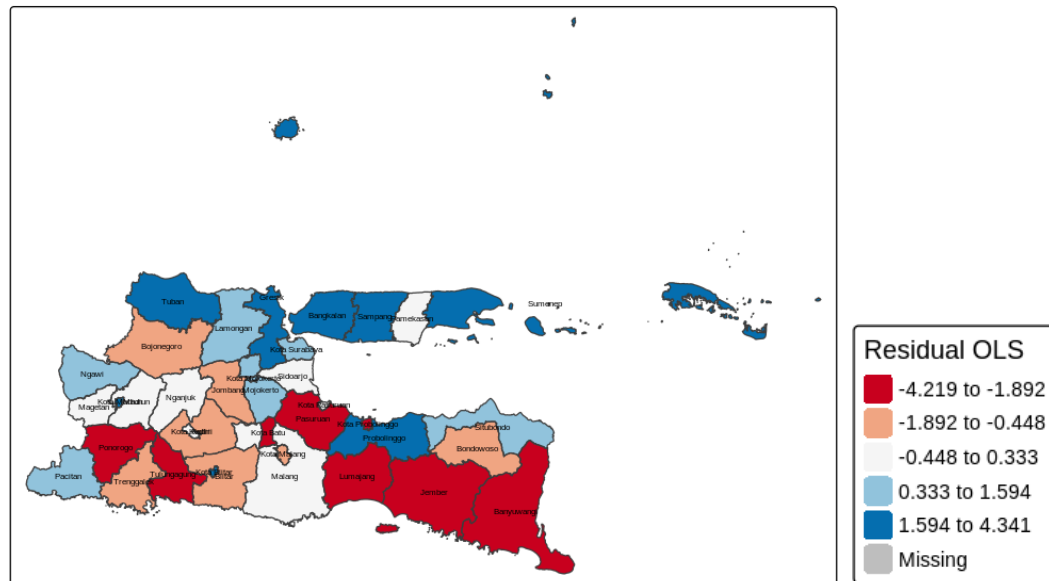
Peta Residual Model SLM



Peta Residual Model SEM



Berikut ini akan diinterpretasikan Peta Residual dari Model OLS



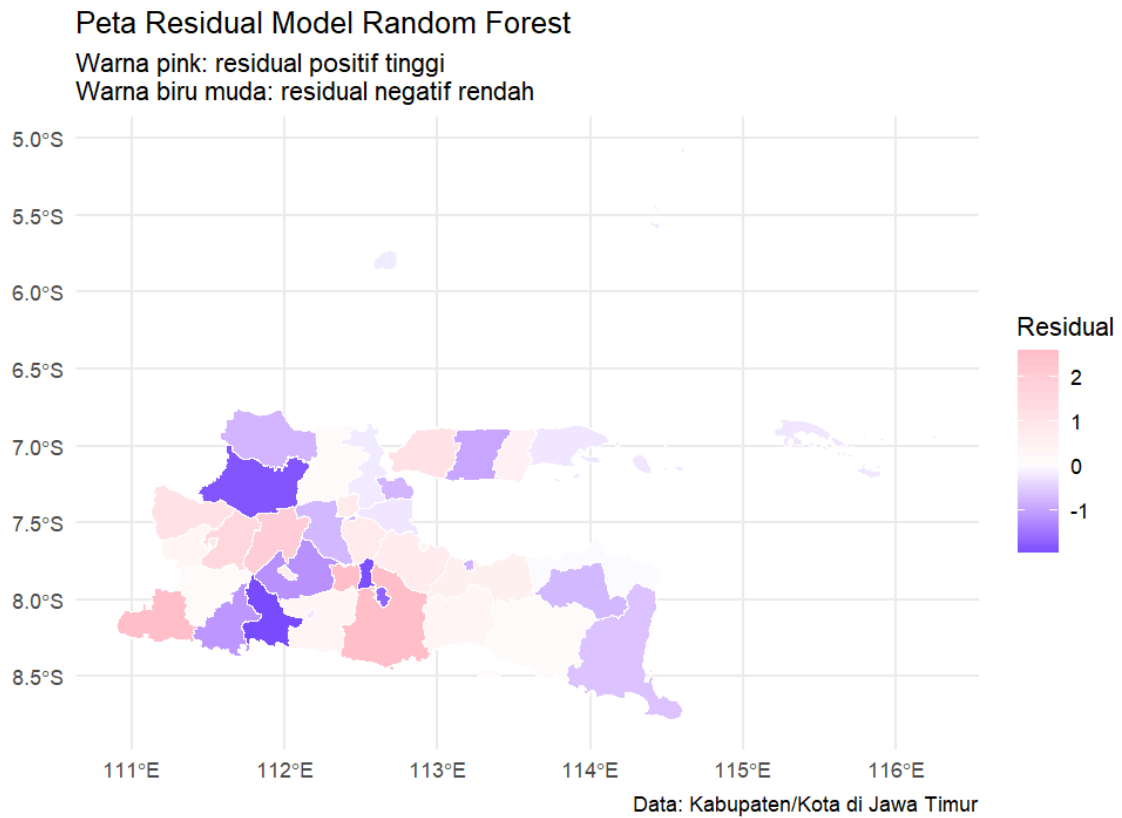
**Gambar 4.8. Peta Residual OLS (Heatmap)**

Analisis residual dilakukan pada model OLS yang dipilih script berdasarkan BIC terendah. Peta residual (**Gambar 4.8**) dibuat untuk mengamati pola sebaran residual secara visual. Observasi pada peta residual model OLS menunjukkan bahwa **pola residual tidak tersebar secara acak**. Terdapat pengelompokan spasial error (klaster) yang jelas, di mana wilayah dengan residual positif tinggi (merah) memiliki pola yang acak dan kecenderungan wilayahnya terkonsentrasi di sepanjang pesisir utara Jawa Timur. Begitupun, wilayah dengan residual negatif rendah (biru) yang juga memiliki pola yang acak dan kecenderungan wilayahnya berada di bagian selatan Jawa Timur.

Asumsinya pola residual pada peta OLS akan cenderung acak tanpa pengelompokan spasial yang jelas, namun pada kenyataannya masih terlihat adanya pola pengelompokan residual pada peta, ini bisa mengindikasikan adanya **faktor spasial lain atau heterogenitas spasial yang belum tertangkap oleh model**. Meskipun model statistik (OLS) tidak menemukan adanya autokorelasi spasial pada *error* secara global, visualisasi peta menunjukkan bahwa mungkin ada variabel-variabel penting yang hilang dari model yang memiliki pola spasial Utara-Selatan yang kuat, atau hubungan antar variabel mungkin berbeda antar wilayah (heterogenitas spasial).

Selain ketiga model regresi spasial (OLS, SAR, dan SEM), model Random Forest juga diikutsertakan dalam perbandingan untuk mengevaluasi kinerja prediksi. Hasil analisis menunjukkan bahwa Random Forest memiliki pseudo R-squared sebesar 0.9759 dan RMSE

sebesar 1.116, yang jauh lebih baik dibandingkan model OLS (R-squared 0.7506) maupun model spasial lainnya (SAR dan SEM).



**Gambar 4.9. Peta Residual Spatial Random Forest**

Analisis residual pada Random Forest menunjukkan bahwa residual model terdistribusi normal (Shapiro Wilk  $p=0.5099$ ) dan tidak menunjukkan adanya autokorelasi spasial (Moran's I negatif kecil dan  $p\text{-value} > 0.26$ ). Peta residual Random Forest (Gambar 4.9) menunjukkan sebaran spasial yang acak, di mana area residual positif tinggi (warna pink) dan negatif rendah (warna biru muda) tidak membentuk pola kluster spasial yang signifikan. Hal ini konsisten dengan hasil uji Moran's I, yang menunjukkan tidak adanya pola spasial pada residual model Random Forest.

Dengan demikian, meskipun model SEM memiliki AIC/BIC sedikit lebih baik, Random Forest menunjukkan kinerja prediksi yang lebih unggul dalam menjelaskan variasi kemiskinan antar wilayah. Namun, karena Random Forest tidak memberikan interpretasi koefisien secara eksplisit (black box), maka untuk tujuan interpretasi hubungan antar variabel, model OLS atau SEM tetap relevan. Sebaliknya, untuk tujuan prediksi yang lebih akurat, model Random Forest menjadi pilihan utama.

## **Bab 5**

### **Kesimpulan dan Saran**

#### **5.1 Kesimpulan**

Berdasarkan hasil analisis yang telah dilakukan pada data persentase kemiskinan di Provinsi Jawa Timur, beberapa kesimpulan utama dapat diambil sebagai berikut:

1. Model OLS, SAR, dan SEM memiliki nilai R-squared yang serupa (sekitar 0.75), dengan hasil uji dependensi spasial pada residual menunjukkan tidak adanya autokorelasi spasial signifikan. Hal ini menegaskan bahwa efek spasial tidak dominan dalam mempengaruhi variasi kemiskinan di wilayah ini.
2. Model Random Forest menunjukkan kinerja prediksi yang lebih unggul, dengan nilai pseudo R-squared sebesar 0.9759 dan RMSE 1.116. Analisis residual juga menunjukkan residual yang berdistribusi normal dan tidak terdapat autokorelasi spasial signifikan, sejalan dengan hasil visualisasi peta residual yang acak.
3. Variabel yang paling berpengaruh dalam menjelaskan variasi kemiskinan adalah rata-rata lama sekolah (X2), diikuti oleh Gini Ratio (X1), Jumlah Kendaraan Bermotor (X4), dan Jumlah Tenaga Medis (X3). Hal ini menunjukkan pentingnya faktor pendidikan dalam mengurangi tingkat kemiskinan.
4. Secara keseluruhan, meskipun model spasial (SAR dan SEM) tidak menunjukkan keunggulan signifikan dibandingkan OLS, Random Forest menjadi model terbaik untuk tujuan prediksi karena mampu menangkap hubungan yang kompleks dan non-linear antar variabel.

#### **5.2 Saran**

Berdasarkan hasil analisis dan kesimpulan yang diperoleh, beberapa saran yang dapat diberikan adalah sebagai berikut:

1. Untuk kepentingan interpretasi hubungan antar variabel (arah dan besar pengaruh), model OLS tetap relevan dan dapat digunakan. Namun, untuk tujuan prediksi yang lebih akurat, disarankan menggunakan model Random Forest.
2. Pemerintah daerah dapat mempertimbangkan faktor pendidikan (rata-rata lama sekolah) sebagai prioritas utama dalam program pengentasan kemiskinan, karena terbukti menjadi faktor paling berpengaruh dalam model prediksi ini.

3. Penelitian lanjutan dapat dilakukan dengan mempertimbangkan data yang lebih detail dan variabel tambahan (misalnya infrastruktur, akses kesehatan, dan program bantuan sosial) untuk meningkatkan akurasi prediksi dan pemahaman hubungan spasial yang lebih komprehensif.

## Lampiran

- **Link Presentasi:**  
<https://youtu.be/ODEhQMngCi8>
- **Data**

Kabupaten	latitude	longitude	Y	X1	X2	X3	X4
Pacitan	-8,20461 4	111,08769	13,08	0,327	7,9	47	187184
Ponorogo	-7,86782 7	111,466003	9,11	0,326	7,8	59	472695
Trenggalek	-8,05 7	111,716666	10,5	0,350	7,92	58	311129
Tulungagung	-8,06666 67	111,9	6,28	0,321	8,68	53	689732
Blitar	-8,10144 19	112,162762	8,16	0,353	7,87	35	493701
Kediri	-7,80935 6	112,032356	9,95	0,300	8,26	59	771741
Malang	-8,04956 43	112,688454 9	8,98	0,343	7,8	72	110863 6
Lumajang	-8,13702 2	113,226601	8,65	0,363	7,27	36	325250
Jember	-8,17235 7	113,700302	9,01	0,321	6,54	57	725747
Banyuwangi	-8,21861 11	114,366944 4	6,54	0,312	7,78	36	756029

<b>Bondowoso</b>	<b>-7,91770 4</b>	<b>113,813483</b>	<b>12,6</b>	<b>0,333</b>	<b>6,53</b>	<b>40</b>	<b>184105</b>
<b>Situbondo</b>	<b>-7,70253 4</b>	<b>113,955605</b>	<b>11,51</b>	<b>0,385</b>	<b>6,93</b>	<b>25</b>	<b>199995</b>
<b>Probolinggo</b>	<b>-7,75396 5</b>	<b>113,210675</b>	<b>16,45</b>	<b>0,339</b>	<b>6,31</b>	<b>34</b>	<b>272224</b>
<b>Pasuruan</b>	<b>-6,86230 98</b>	<b>108,800193 6</b>	<b>8,63</b>	<b>0,331</b>	<b>7,46</b>	<b>45</b>	<b>467988</b>
<b>Sidoarjo</b>	<b>-7,45302 78</b>	<b>112,717338 9</b>	<b>4,53</b>	<b>0,327</b>	<b>10,91</b>	<b>97</b>	<b>142315 7</b>
<b>Mojokerto</b>	<b>-7,48807 5</b>	<b>112,427027</b>	<b>9,37</b>	<b>0,337</b>	<b>9,13</b>	<b>41</b>	<b>483993</b>
<b>Jombang</b>	<b>-7,54683 95</b>	<b>112,226479 4</b>	<b>8,6</b>	<b>0,302</b>	<b>8,78</b>	<b>57</b>	<b>608569</b>
<b>Nganjuk</b>	<b>-7,60293 2</b>	<b>111,901808</b>	<b>10,17</b>	<b>0,289</b>	<b>8,25</b>	<b>37</b>	<b>552537</b>
<b>Madiun</b>	<b>-7,62775 3</b>	<b>111,505483</b>	<b>10,63</b>	<b>0,309</b>	<b>8,2</b>	<b>33</b>	<b>345211</b>
<b>Magetan</b>	<b>-7,64934 13</b>	<b>111,338159 3</b>	<b>9,32</b>	<b>0,371</b>	<b>8,69</b>	<b>49</b>	<b>375980</b>
<b>Ngawi</b>	<b>-7,38993</b>	<b>111,46193</b>	<b>13,81</b>	<b>0,289</b>	<b>7,84</b>	<b>51</b>	<b>413826</b>
<b>Bojonegoro</b>	<b>-7,22783 7</b>	<b>111,794685 58</b>	<b>11,69</b>	<b>0,310</b>	<b>7,59</b>	<b>58</b>	<b>512334</b>
<b>Tuban</b>	<b>-8,74931 46</b>	<b>115,171129 8</b>	<b>14,36</b>	<b>0,335</b>	<b>7,53</b>	<b>45</b>	<b>519999</b>



<b>Lamongan</b>	<b>-7,40615 3</b>	<b>109,394679 4</b>	<b>12,16</b>	<b>0,281</b>	<b>8,48</b>	<b>40</b>	<b>555884</b>
<b>Gresik</b>	<b>-7,15665</b>	<b>112,6555</b>	<b>10,32</b>	<b>0,322</b>	<b>10,03</b>	<b>54</b>	<b>727397</b>
<b>Bangkalan</b>	<b>-7,03069 12</b>	<b>112,745006 8</b>	<b>18,66</b>	<b>0,301</b>	<b>6,01</b>	<b>24</b>	<b>245175</b>
<b>Sampang</b>	<b>-7,57825 56</b>	<b>109,205843 6</b>	<b>20,83</b>	<b>0,233</b>	<b>5,08</b>	<b>35</b>	<b>140363</b>
<b>Pamekasan</b>	<b>-7,16666 67</b>	<b>113,466666 7</b>	<b>13,41</b>	<b>0,301</b>	<b>7,17</b>	<b>41</b>	<b>269486</b>
<b>Sumenep</b>	<b>-6,92539 99</b>	<b>113,906062 4</b>	<b>17,78</b>	<b>0,308</b>	<b>6,1</b>	<b>39</b>	<b>256780</b>
<b>Kota Kediri</b>	<b>-7,81689 5</b>	<b>112,011398</b>	<b>6,51</b>	<b>0,337</b>	<b>10,92</b>	<b>38</b>	<b>337697</b>
<b>Kota Blitar</b>	<b>-8,1</b>	<b>112,15</b>	<b>6,75</b>	<b>0,353</b>	<b>10,82</b>	<b>11</b>	<b>290994</b>
<b>Kota Malang</b>	<b>-7,98189 4</b>	<b>112,626503</b>	<b>3,91</b>	<b>0,422</b>	<b>11,14</b>	<b>79</b>	<b>535090</b>
<b>Kota Probolinggo</b>	<b>-7,75692 8</b>	<b>113,211502</b>	<b>6,18</b>	<b>0,344</b>	<b>9,72</b>	<b>30</b>	<b>175288</b>
<b>Kota Pasuruan</b>	<b>-7,64487 2</b>	<b>112,903297</b>	<b>6,32</b>	<b>0,334</b>	<b>9,94</b>	<b>16</b>	<b>249840</b>
<b>Kota Mojokerto</b>	<b>-7,47222 22</b>	<b>112,433611 1</b>	<b>5,57</b>	<b>0,357</b>	<b>11,38</b>	<b>20</b>	<b>208970</b>
<b>Kota Madiun</b>	<b>-7,62971 4</b>	<b>111,513702</b>	<b>4,38</b>	<b>0,435</b>	<b>12,11</b>	<b>39</b>	<b>217555</b>
<b>Kota Surabaya</b>	<b>-7,28916 6</b>	<b>112,734398</b>	<b>3,96</b>	<b>0,381</b>	<b>10,89</b>	<b>228</b>	<b>303475 4</b>

<b>Kota Batu</b>	<b>-7,8671</b>	<b>112,5239</b>	<b>3,06</b>	<b>0,327</b>	<b>9,87</b>	<b>12</b>	<b>168249</b>
------------------	----------------	-----------------	-------------	--------------	-------------	-----------	---------------

<https://colab.research.google.com/drive/1hD3SH8R3mqIiAyAeMATvajwnGZ4ysNzH>

- **R CODE**

- (Regresi Spasial)

## # Library

```
library(dplyr)
```

```
library(sf)
```

library(sp)

```
library(car)
```

```
library(lmtest)
```

```
library(ggplot2)
```

```
library(corrplot)
```

library(GGally)

```
library(RColorBrewer)
```

```
library(spdep)
```

```
Data <- read_excel("C:\\Users\\Bryan\\Documents\\UI Stuff\\Spasial Stuff\\T4 RF DATASET\\T.xlsx")
```

View(Data)

```
maps <- st_read("C:\\Users\\Bryan\\Documents\\UI Stuff\\Spatial Stuff\\[geosai.my.id]Jawa Timur Kab\\Jawa Timur ADMIN BPS.shp")
```

```
Data <- Data[order(Data$`Kabupaten`), ]
```

```
maps <- maps[order(maps$Kabupaten), ]
```

```
# Create a list of neighbours using the Queen criteria
```

```
w <- poly2nb(maps)
```

```
ww <- nb2listw(w, style = 'B')
```

```
# Uji Indeks Moran
# Variabel Dependen
moran.test(Data$Y, ww, alternative="two.sided")
# Variabel Independen
moran.test(Data$X1, ww, alternative="two.sided")
# Variabel Independen
moran.test(Data$X2, ww, alternative="two.sided")
# Variabel Independen
moran.test(Data$X3, ww, alternative="two.sided")
# Variabel Independen
moran.test(Data$X4, ww, alternative="two.sided")
```

Output:

Moran I test under randomisation

data: Data\$Y

weights: ww

Moran I statistic standard deviate = 2.0699, p-value = 0.03846

alternative hypothesis: two.sided

sample estimates:

Moran I statistic	Expectation	Variance
0.19982194	-0.02702703	0.01201116

```
> # Variabel Independen
```

```
> moran.test(Data$X1, ww, alternative="two.sided")
```

Moran I test under randomisation

data: Data\$X1

weights: ww

Moran I statistic standard deviate = 1.9432, p-value = 0.052

alternative hypothesis: two.sided

sample estimates:

Moran I statistic	Expectation	Variance
0.18166856	-0.02702703	0.01153479

> # Variabel Independen

> moran.test(Data\$X2, ww, alternative="two.sided")

Moran I test under randomisation

data: Data\$X2

weights: ww

Moran I statistic standard deviate = 2.824, p-value = 0.004744

alternative hypothesis: two.sided

sample estimates:

Moran I statistic	Expectation	Variance
0.28567789	-0.02702703	0.01226183

> # Variabel Independen

> moran.test(Data\$X3, ww, alternative="two.sided")

Moran I test under randomisation

data: Data\$X3

weights: ww

Moran I statistic standard deviate = 2.026, p-value = 0.04277

alternative hypothesis: two.sided

sample estimates:

Moran I statistic	Expectation	Variance
0.133724696	-0.027027027	0.006295556

> # Variabel Independen

> moran.test(Data\$X4, ww, alternative="two.sided")

## Moran I test under randomisation

data: Data\$X4

weights: ww

Moran I statistic standard deviate = 2.7146, p-value = 0.006635

alternative hypothesis: two.sided

sample estimates:

Moran I statistic	Expectation	Variance
0.198383461	-0.027027027	0.006894885

### - Script untuk Uji Asumsi Regresi Linear

```
# -----  
# Load library  
# -----  
library(readxl)  
library(car)  
library(lmtest)  
library(ggplot2)  
  
# -----  
# Load dataset  
# -----  
data <- read_excel("C:\\Users\\Bryan\\Documents\\UI  
Stuff\\Spasial Stuff\\T4 RF Dataset.xlsx")  
data_lm <- data.frame(  
  Y = data$Y,  
  X1 = data$X1,  
  X2 = data$X2,  
  X3 = data$X3,  
  X4 = data$X4  
)  
  
# -----  
# Fit model OLS
```

```

# -----
model <- lm(Y ~ X1 + X2 + X3 + X4, data = data_lm)

# -----
# 1. Normalitas residual (Shapiro-Wilk)
# -----
residuals <- resid(model)
shapiro.test(residuals)

# -----
# 2. Multikolinearitas (VIF)
# -----
vif(model)

# -----
# 3. Heteroskedastisitas (Breusch-Pagan)
# -----
bptest(model)

# -----
# 4. Autokorelasi residual (Durbin-Watson)
# -----
dwtest(model)

# -----
# 5. Plot residual vs fitted
# -----
ggplot(data = data.frame(fitted = fitted(model), residuals =
residuals),
       aes(x = fitted, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Plot Residual vs Fitted", x = "Fitted Values",
y = "Residuals") +

```

```
theme_minimal()
```

## Output

```
> # Load library
> # -----
> library(readxl)
> library(car)
> library(lmtest)
> library(ggplot2)
>
> # -----
> # Load dataset
> # -----
> data <- read_excel("C:\\Users\\Bryan\\Documents\\UI Stuff\\Spasial
Stuff\\T4 RF Dataset.xlsx")
> data_lm <- data.frame(
+   Y = data$Y,
+   X1 = data$X1,
+   X2 = data$X2,
+   X3 = data$X3,
+   X4 = data$X4
+ )
>
> # -----
> # Fit model OLS
> # -----
> model <- lm(Y ~ X1 + X2 + X3 + X4, data = data_lm)
>
> # -----
> # 1. Normalitas residual (Shapiro-Wilk)
> # -----
> residuals <- resid(model)
> shapiro.test(residuals)
```

Shapiro-Wilk normality test

```
data: residuals
W = 0.98554, p-value = 0.8954
```

```
>
```



```

> # -----
> # 2. Multikolinearitas (VIF)
> # -----
> vif(model)
      X1      X2      X3      X4
1.644228 1.659229 8.616857 8.758640
>
> # -----
> # 3. Heteroskedastisitas (Breusch-Pagan)
> # -----
> bptest(model)

```

studentized Breusch-Pagan test

```

data: model
BP = 4.9949, df = 4, p-value = 0.2878

```

```

>
> # -----
> # 4. Autokorelasi residual (Durbin-Watson)
> # -----
> dwtest(model)

```

Durbin-Watson test

```

data: model
DW = 1.2626, p-value = 0.005238
alternative hypothesis: true autocorrelation is greater than 0

```

```

>
> # -----
> # 5. Plot residual vs fitted
> # -----
> ggplot(data = data.frame(fitted = fitted(model), residuals = residuals),
+       aes(x = fitted, y = residuals)) +
+   geom_point() +
+   geom_hline(yintercept = 0, color = "red") +
+   labs(title = "Plot Residual vs Fitted", x = "Fitted Values", y =
"Residuals") +
+   theme_minimal()

```

### - Script untuk Spatial Random Forest

```
if (!requireNamespace("devtools", quietly = TRUE)) {  
  install.packages("devtools")  
}  
  
if (!requireNamespace("sf", quietly = TRUE)) {  
  install.packages("sf")  
}  
  
if (!requireNamespace("readxl", quietly = TRUE)) {  
  install.packages("readxl")  
}  
  
if (!requireNamespace("spatialRF", quietly = TRUE)) {  
  devtools::install_github("BlasBenito/spatialRF")  
}  
  
library(sf)  
library(readxl)  
library(spatialRF)  
  
data      <-      read_excel("C:\\Users\\Bryan\\Documents\\UI  
Stuff\\Spasial Stuff\\T4 RF Dataset.xlsx")  
  
data_rf <- data.frame(  
  Y = data$Y,  
  X1 = data$X1,  
  X2 = data$X2,  
  X3 = data$X3,  
  X4 = data$X4,  
  latitude = data$latitude,  
  longitude = data$longitude  
)  
  
coords <- data_rf[, c("longitude", "latitude")]  
dist_matrix <- as.matrix(dist(coords))
```

```

rf_model <- rf_spatial(
  data = data_rf,
  dependent.variable.name = "Y",
  predictor.variable.names = c("X1", "X2", "X3", "X4"),
  distance.matrix = dist_matrix,
  n.cores = 2
)

print(rf_model)
plot_importance(rf_model)

# Ambil residual
data_rf$residual <- rf_model$residuals$values

# Load shapefile untuk plot spasial
shp <- st_read("C:\\Users\\Bryan\\Documents\\UI Stuff\\Spasial
Stuff\\[geosai.my.id]Jawa_Timur_Kab\\Jawa_Timur_ADMIN_BPS.shp"
) # pastikan path-nya sama

shp$residual <- data_rf$residual

# Plot peta residual
library(ggplot2)
ggplot(shp) +
  geom_sf(aes(fill = residual), color = "white") +
  scale_fill_gradient2(
    low = "blue", mid = "white", high = "pink",
    midpoint = 0, name = "Residual"
  ) +
  labs(
    title = "Peta Residual Model Random Forest",

```

```

        subtitle = "Warna pink: residual positif tinggi\nWarna
biru muda: residual negatif rendah",
        caption = "Data: Kabupaten/Kota di Jawa Timur"
    ) +
    theme_minimal()

```

### **Output**

```

> if (!requireNamespace("devtools", quietly = TRUE)) {
+   install.packages("devtools")
+ }
> if (!requireNamespace("sf", quietly = TRUE)) {
+   install.packages("sf")
+ }
> if (!requireNamespace("readxl", quietly = TRUE)) {
+   install.packages("readxl")
+ }
> if (!requireNamespace("spatialRF", quietly = TRUE)) {
+   devtools::install_github("BlasBenito/spatialRF")
+ }
>
> library(sf)
> library(readxl)
> library(spatialRF)
>
> data <- read_excel("C:\\Users\\Bryan\\Documents\\UI Stuff\\Spasial
Stuff\\T4 RF Dataset.xlsx")
>
> data_rf <- data.frame(
+   Y = data$Y,
+   X1 = data$X1,
+   X2 = data$X2,
+   X3 = data$X3,
+   X4 = data$X4,
+   latitude = data$latitude,
+   longitude = data$longitude
+ )
>
> coords <- data_rf[, c("longitude", "latitude")]
> dist_matrix <- as.matrix(dist(coords))
>
> rf_model <- rf_spatial(

```

```

+ data = data_rf,
+ dependent.variable.name = "Y",
+ predictor.variable.names = c("X1", "X2", "X3", "X4"),
+ distance.matrix = dist_matrix,
+ n.cores = 2
+ )

```

The model residuals are not spatially correlated, there is no need to fit a spatial model

```

>
>
> print(rf_model)

```

Model type

```

- Fitted with: ranger()
- Response variable: Y

```

Random forest parameters

```

- Type: Regression
- Number of trees: 500
- Sample size: 38
- Number of predictors: 4
- Mtry: 2
- Minimum node size: 5

```

Model performance

```

- R squared (oob): 0.6534121
- R squared (cor(obs, pred)^2): 0.9525424
- Pseudo R squared (cor(obs, pred)): 0.9759828
- RMSE (oob): 2.480529
- RMSE: 1.116
- Normalized RMSE: 0.201991

```

Model residuals

```

- Stats:

```

Min.	1st Q.	Median	Mean	3rd Max.
-1.96	-0.78	0.01	0.00	0.65

- Normality:

- Shapiro-Wilks W: 0.974

```
- p-value      : 0.5099
```

- Interpretation : Residuals are normal

- Spatial autocorrelation:

Distance	Moran's I	P value	Interpretation
----------	-----------	---------	----------------

0.0	-0.037	0.900	No spatial correlation
-----	--------	-------	------------------------

1.0      -0.039      0.265      No spatial correlation

2.0      -0.033      0.592      No spatial correlation

3.0	-0.033	0.532	No spatial correlation
-----	--------	-------	------------------------

A horizontal number line with tick marks at 0, 1, 2, and 3.

Variable importance:

Variable	Importance
x2	3.762
x1	1.448
x4	1.370
x3	0.794

```
> plot_importance(rf_model)
```

>

>

```

> # Ambil residual
> data_rf$residual <- rf_model$residuals$values
>
> # Load shapefile untuk plot spasial
> shp <- st_read("C:\\Users\\Bryan\\Documents\\UI Stuff\\Spasial
Stuff\\[geosai.my.id]Jawa_Timur_Kab\\Jawa_Timur_ADMIN_BPS.shp") # pastikan
path-nya sama
Reading layer `Jawa_Timur_ADMIN_BPS' from data source
          `C:\\Users\\Bryan\\Documents\\UI Stuff\\Spasial
Stuff\\[geosai.my.id]Jawa_Timur_Kab\\Jawa_Timur_ADMIN_BPS.shp'
  using driver `ESRI Shapefile'
Simple feature collection with 38 features and 6 fields
Geometry type: MULTIPOLYGON
Dimension:      XY
Bounding box:   xmin: 110.8987 ymin: -8.78036 xmax: 116.2702 ymax: -5.048857
Geodetic CRS:   WGS 84
>
> shp$residual <- data_rf$residual
>
> # Plot peta residual
> library(ggplot2)
> ggplot(shp) +
+   geom_sf(aes(fill = residual), color = "white") +
+   scale_fill_gradient2(
+     low = "blue", mid = "white", high = "pink",
+     midpoint = 0, name = "Residual"
+   ) +
+   labs(
+     title = "Peta Residual Model Random Forest",
+     subtitle = "Warna pink: residual positif tinggi\nWarna biru muda:
residual negatif rendah",
+     caption = "Data: Kabupaten/Kota di Jawa Timur"
+   ) +
+   theme_minimal()

```

## Daftar Pustaka

- A.A. Maulana, A. Fauzan. "Spatially informed insights: Modeling percentage poverty in East Java Province using SEM with spatial weight variations," *BAREKENG: Journal of Mathematics and Its Applications*, vol. 18, no. 6, pp. 1317-1332, June 2024.
- Badan Pusat Statistik (BPS), [Online]. Available: <https://www.bps.go.id/id> [Accessed: 26 May 2025].
- Badan Pusat Statistik (BPS), "Persentase Penduduk Miskin di Jawa Timur September 2024 turun menjadi 9,56 persen", [Online]. Available: <https://jatim.bps.go.id/id/pressrelease/2025/01/15/1474/persentase-penduduk-miskin-di-jawa-timur-september-2024-turun-menjadi-9-56-persen.html> [Accessed: 5 June 2025].
- E.W. Fox, J.M.V. Hoef. "Comparing spatial regression to random forests for large environmental data sets," *PLoS ONE*, vol. 15, no. 3, March 2020.
- IBM, "Apa itu random forest?", [Online]. Available: <https://www.ibm.com/id-id/think/topics/random-forest> [Accessed: 7 June 2025].
- L. Anselin, *Spatial econometrics: methods and models*. New York: McGraw Hill, 1953.
- N.U. Ayudia, D.E. Putri. "Pemodelan Spatial Autoregressive Model (SAR) pada kasus kemiskinan di Jawa Timur," *Jurnal Gaussian*, vol. 13, no. 2, pp. 308-1318, November 2024.
- R. Hasanah, Syaparuddin, Rosmeli. "Pengaruh angka harapan hidup, rata-rata lama sekolah dan pengeluaran perkapita terhadap tingkat kemiskinan pada Kabupaten /Kota di Provinsi Jambi," *E-Jurnal Perspektif Ekonomi Dan Pembangunan Daerah*, vol. 10, no. 3, pp. 223-232, September 2021.



**Tabel Kontribusi Kelompok**

<b>No.</b>	<b>Nama</b>	<b>NPM</b>	<b>Kontribusi</b>	<b>Tingkat Kontribusi</b>
1.	Ammar Hanafi	2206051582	Aktif berdiskusi, menyusun makalah, paper, dan ppt bagian hasil dan pembahasan analisis regresi spasial.	100%
2.	Bryan Jonathan	2206052780	Aktif berdiskusi, menyusun makalah, paper, dan ppt bagian hasil dan pembahasan random forest & kesimpulan.	100%
3.	Renata Shaula Alfino Ritonga	2206815812	Aktif berdiskusi, menyusun makalah, paper, dan ppt bagian pendahuluan, landasan teori, & data dan metode.	100%