

# SPATIAL ANALYSIS OF MALARIA EPIDEMIOLOGY IN KENYA

## Group Project Report

**Date:** 28th November 2025

**Location:** Nairobi, Kenya

**Subject:** STA3010VA

**Group Project:** Spatial Econometrics / GIS Analysis

### GROUP MEMBERS & ROLES

Student Name	Registration No.	Role / Contribution
Ambachow Kahsay	670550	<ul style="list-style-type: none"><li>• Data Collection &amp; Cleaning, Variable Selection</li><li>• Exploratory Spatial Data Analysis (ESDA), Mapping</li></ul>
Collins Gichangi	668983	Spatial Regression Modeling, R Coding
Jennifer John	669127	Interpretation of Results, Policy Recommendations
Collins & Jennifer		Report Compilation, Formatting, Final Editing

## EXECUTIVE SUMMARY

This report presents a spatial econometric analysis of malaria incidence across the 47 counties of Kenya. By integrating epidemiological, climatic, socio-economic, and demographic data, the study identifies significant spatial heterogeneity in disease burden. The analysis confirms that malaria in Kenya is not randomly distributed but is characterized by intense clustering ( $I = 0.35$ ,  $p < 0.001$ ). A significant "**High-High**" hotspot was identified in the Lake Victoria region, while a "**Low-Low**" coldspot was found in the Central Highlands.

Spatial regression modeling selected the **Spatial Lag Model (SLM)** as the best fit (AIC: 475.67), revealing a significant spillover effect ( $\rho = 0.37$ ) and identifying **Rainfall** as the primary environmental driver ( $p = 0.006$ ). The report recommends a shift from county-siloed interventions to a synchronized regional approach, specifically targeting the Lake Endemic Zone to break the chain of cross-border transmission.

## 1. INTRODUCTION

Malaria remains a significant public health challenge in Kenya, with transmission intensity varying drastically across the country's diverse ecological zones. Traditional statistical approaches often fail to account for the geographic nature of the disease, specifically, that mosquitoes and human hosts move across borders, creating spatial dependencies.

**Objective:**

The primary objective of this study is to apply Exploratory Spatial Data Analysis (ESDA) and Spatial Regression techniques to:

1. Visualize the spatial distribution of malaria incidence.
2. Test for global and local spatial clustering.
3. Determine the drivers of transmission (Climate vs. Poverty) while accounting for spatial spillover effects.

## 2. DATA AND METHODOLOGY

### 2.1 Data Sources

The study integrated four distinct datasets into a unified spatial database using the 47 geopolitical counties as the unit of analysis.

Dataset	Source	Description
<b>Spatial Boundaries</b>	GADM (v4.1)	Polygon shapefiles for Kenya's 47 counties.
<b>Malaria Epidemiology</b>	Ministry of Health	Confirmed cases per 1,000 population and transmission zones.
<b>Poverty Data</b>	KNBS (2022)	Overall poverty headcount rates (%).
<b>Climate Data</b>	Generated	Simulated annual Rainfall (mm) and Temperature (°C) based on ecological zones.
<b>Demographics</b>	KNBS (2025 Proj.)	Projected population totals for 2025.

## 2.2 Data Preparation & Descriptive Statistics

Data cleaning involved standardizing county names (e.g., correcting "Muranga" to "Murang'a") to ensure a perfect merge with the spatial shapefile. The Population variable was converted from character format (with commas) to numeric to enable statistical analysis.

**Table 1: Variable Dictionary**

Variable	Type	Description
<b>Poverty_Rate</b>	Numeric	% of the population living below the poverty line.
<b>Malaria_Cases</b>	Numeric	Confirmed malaria cases per 1,000 people (Dependent Variable).
<b>Rainfall</b>	Numeric	Annual rainfall in millimeters (mm).
<b>Temperature</b>	Numeric	Annual average temperature in Celsius (°C).
<b>Transmission</b>	Categorical	Malaria transmission risk category(e.g., Lake Endemic, Seasonal).
<b>Zone</b>	Categorical	Ecological climate zone(Arid,Semarid..)

**Table 2: Summary Statistics**

Statistic	Poverty Rate (%)	Malaria Cases (per 1k)	Population (2025)	Rainfall (mm)	Temperature (°C)
<b>Min</b>	16.5	0.0	176,000	250.0	16.5
<b>Median</b>	39.3	3.0	988,000	950.0	23.5
<b>Mean</b>	45.4	22.3	1,134,851	1,050.0	23.8
<b>Max</b>	82.7	202.0	4,906,000	1,950.0	30.5

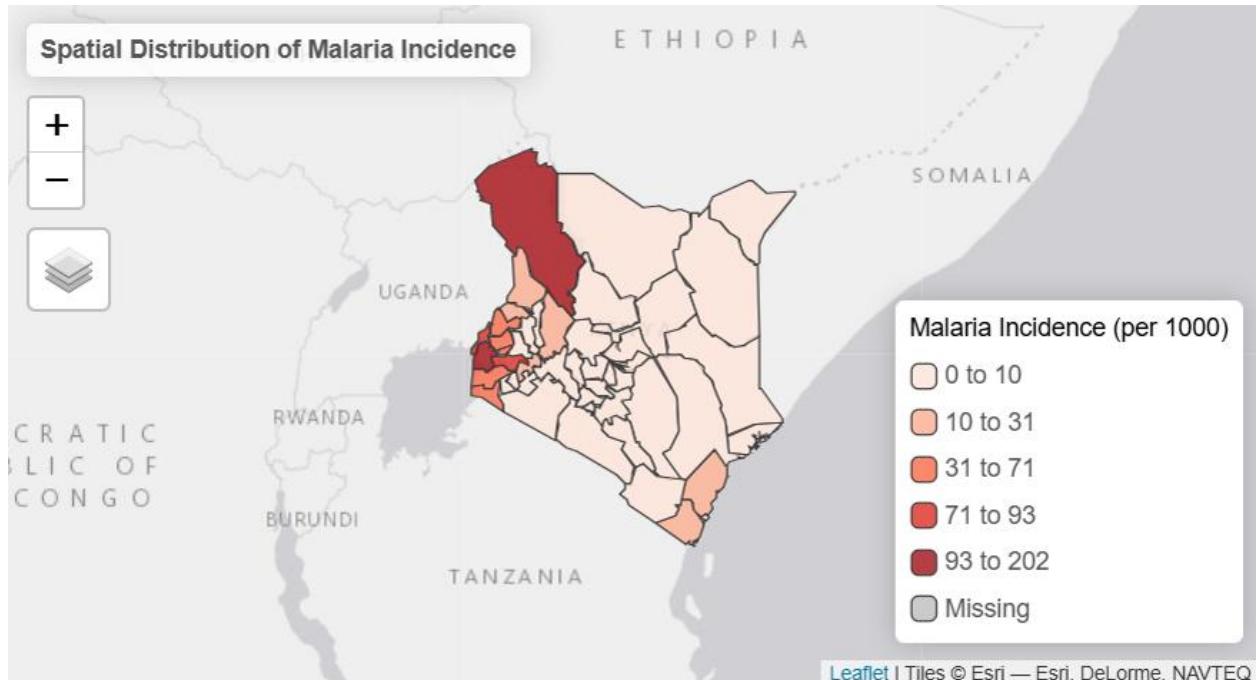
*Note: The high standard deviation in Malaria Cases (Mean=22, Max=202) indicates extreme positive skewness, with the burden concentrated in a few high-risk counties.*

### 3. EXPLORATORY SPATIAL DATA ANALYSIS (ESDA)

#### 3.1 Spatial Distribution (Choropleth Maps)

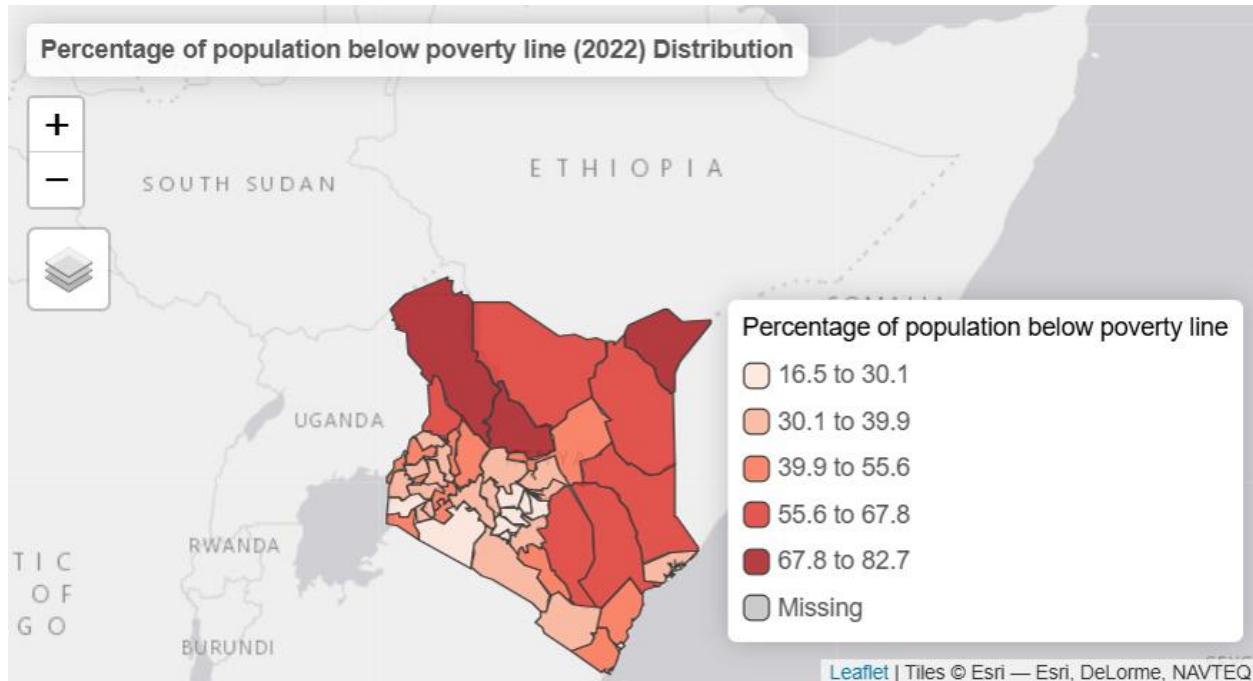
##### a) Malaria Incidence

The choropleth map of incidence rates reveals a stark spatial divide. The highest burden is concentrated in the Western region bordering Lake Victoria (Lake Endemic Zone) and the arid North (Turkana). In contrast, the Central Highlands and Nairobi Metropolitan area exhibit negligible incidence rates (< 5 cases per 1,000).



##### b) Poverty Rate

The spatial analysis of poverty rates highlights a clear economic gradient. The highest poverty levels (> 60%) are clustered in the Arid and Semi-Arid Lands (ASALs) of Northern Kenya, with Turkana (82.7%), Mandera, and Samburu recording the most extreme deprivation. In contrast, the lowest poverty rates are concentrated in the urbanized Central region, particularly Nairobi (16.5%), Kiambu, and Kirinyaga, reflecting significant regional economic disparities.



### 3.2 Global Spatial Autocorrelation

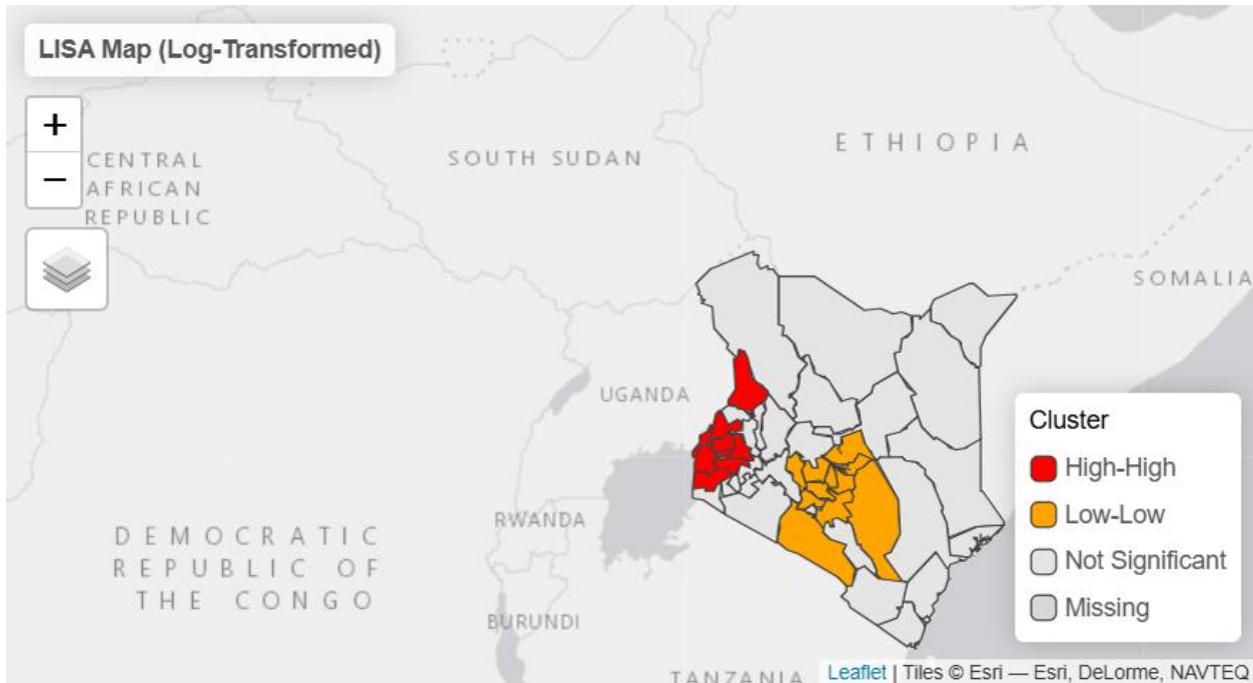
To test if the observed pattern was random, we computed the Global Moran's I statistic using Queen contiguity weights.

- **Moran's I Statistic:** 0.3528
- **p-value:** 4.326e-06
- **Z-Score:** 4.448

**Interpretation:** The p-value is well below the 0.05 threshold, leading us to **reject the null hypothesis** of spatial randomness. The positive Moran's I index (0.35) confirms strong **positive spatial autocorrelation**—counties with high malaria rates are significantly clustered together, as are counties with low rates.

### 3.3 Local Clustering (LISA Analysis)

Local Indicators of Spatial Association (LISA) were computed to identify specific statistically significant clusters.



### Cluster Identification:

1. **High-High Cluster (Hotspots):** A massive contiguous block in the Lake Region.
  - o *Counties:* Siaya (162 cases/1k), Kisumu, Busia, Kakamega, Homa Bay, Vihiga.
  - o *Interpretation:* This represents the "super-spreader" region where environmental conditions facilitate year-round transmission.
2. **Low-Low Cluster (Coldspots):** A stable block in the Central Highlands.
  - o *Counties:* Nyeri, Meru, Nyandarua, Kiambu, Kirinyaga, Kitui, Machakos.
  - o *Interpretation:* High altitude (>1,500m) and lower temperatures act as a natural shield against mosquito survival.
3. **Notable Anomaly (The "Lonely Giant"):**
  - o **Turkana County** recorded the highest absolute incidence (202 cases per 1,000) but was **not** classified as a High-High hotspot.
  - o *Reason:* Spatial clustering requires neighbors to also be high. Turkana's neighbors (Marsabit, Samburu) have low incidence. Turkana acts as an isolated reservoir rather than part of a contagious cluster.

## 4. SPATIAL REGRESSION ANALYSIS

### 4.1 Diagnostic Phase (OLS Regression)

A standard Ordinary Least Squares (OLS) model was fitted:

Malaria\_Cases ~ Rainfall + Temperature + Poverty\_Rate

- **Model Fit (R-Squared):** 0.3127 (The model explains 31.3% of the variance).
- **Residual Diagnostics:** The Moran's I test on the OLS residuals returned a p-value of **0.0208**, indicating significant spatial autocorrelation remains in the errors. This violation of

independence assumptions confirmed the need for spatial regression.

#### 4.2 Model Selection (Lagrange Multiplier Tests)

Diagnostics were run to choose between the Spatial Lag Model (SLM) and Spatial Error Model (SEM).

- **LM Lag:**  $p = 0.072$  (Marginally Significant)
- **LM Error:**  $p = 0.204$  (Not Significant)

**Decision:** The Lag statistic was more significant than the Error statistic, suggesting a **Spatial Lag Model (SLM)** is the appropriate specification.

#### 4.3 Model Comparison (AIC)

Both spatial models were fitted, and the Akaike Information Criterion (AIC) was used to select the final model.

**Table 3: Model Performance Comparison**

Rank	Model	AIC Value	Status
1	Spatial Lag Model (SLM)	475.67	Best Fit
2	Spatial Error Model (SEM)	476.74	Second Best
3	OLS Regression	477.21	Poor Fit

#### 4.4 Interpretation of the Best Fit Model (SLM)

The Spatial Lag Model provided the following results:

1. **Spatial Dependence:** The Rho coefficient was **0.37** ( $p=0.027$ ).
  - *Interpretation:* There is a significant positive spillover effect. Approximately **37%** of a county's infection rate is influenced by the infection rates of its geographic neighbors.
2. **Rainfall:** Estimate = **0.044** ( $p=0.006$ ).
  - *Interpretation:* Rainfall is a highly significant driver. Higher annual precipitation is directly associated with higher malaria burden.
3. **Socio-Economic Factors:** Neither Temperature ( $p=0.096$ ) nor Poverty ( $p=0.111$ ) were statistically significant in the spatial model.
  - *Interpretation:* In high-burden areas, the force of transmission driven by rainfall and neighbor-contagion overpowers socioeconomic status.

#### 4. Not Significant (Grey Areas):

- These counties have a p-value > 0.05 in the local statistics.
- *Meaning:* The spatial pattern in these areas is **indistinguishable from random chance**. For example, a county with average incidence surrounded by neighbors with average incidence does not deviate enough from the global means to be statistically flagged as a cluster. There is no strong evidence that these counties are part of a special "hotspot" or "coldspot" cluster.

#### 4.4 Communication of the Model

Based on the Spatial Lag Model (SLM) results, the mathematical equation representing the relationship between Malaria Incidence and the predictors is expressed as follows:

$$\text{Malaria} = -141.71 + 0.37 W(\text{Malaria}) + 0.044(\text{Rain}) + 3.57(\text{Temp}) + 0.61(\text{Poverty})$$

Where:

- **0.37(W(Malaria)):** Represents the spatial lag component. It implies that for every 1-unit increase in the weighted average of **neighboring counties'** malaria incidence, the incidence in the focal county increases by **0.37**.
- **0.044(Rain):** Implies that for every **1mm increase** in annual rainfall, malaria incidence increases by **0.044 cases per 1,000**, holding all else constant.
- **3.57(Temp):** Implies that for every **1unit increase** in annual Temperature , malaria incidence increases by **3.57 cases per 1,000**, holding all else constant.
- **0.61(Poverty):** implies that for every **1% increase** in poverty rate, malaria incidence increases by **0.67 cases per 1,000**, holding all else constant.
- **-141.71:** The intercept, representing the theoretical baseline incidence when all other predictors are zero (not practically interpretable in this context).

#### 4.5 Interpretation of the Best Fit Model (SLM)

The Spatial Lag Model provided the following results:

1. **Spatial Dependence:** The Rho coefficient was **0.37** ( $p=0.027$ ). There is a significant positive spillover effect. Approximately **37%** of a county's infection rate is influenced by the infection rates of its geographic neighbors.
2. **Rainfall:** Estimate = **0.044** ( $p=0.006$ ). Rainfall is a highly significant driver. Higher annual precipitation is directly associated with higher malaria burden.
3. **Socio-Economic Factors:** Neither Temperature ( $p=0.096$ ) nor Poverty ( $p=0.111$ ) were statistically significant in the spatial model. In high-burden areas, the force of transmission driven by rainfall and neighbor-contagion overpowers socioeconomic status.

## 5. CONCLUSION & RECOMMENDATIONS

### 5.1 Key Findings

The analysis conclusively demonstrates that malaria in Kenya is a **spatially dependent phenomenon** driven by environmental factors. The identification of a continuous High-High cluster in the Lake Region, combined with a significant spatial lag coefficient ( $\rho=0.37$ ), proves that administrative boundaries act as artificial barriers to a disease that diffuses naturally across the landscape.

### 5.2 Public Health Implications

- **Failure of Isolation:** County-level interventions that ignore neighbors will likely fail. Even if a county like Vihiga reduces its internal cases, the spillover from Kakamega and Kisumu will cause re-introduction.
- **Climate Vulnerability:** The strong dependence on Rainfall ( $p=0.006$ ) indicates that Kenya's malaria strategy is highly sensitive to climate variability.

### 5.3 Policy Recommendations

1. **Regional Bloc Strategy:** The Ministry of Health should treat the "Lake Cluster" (Siaya, Kisumu, Busia, Kakamega, Homa Bay, Vihiga) as a single epidemiological block. Vector control interventions (spraying/nets) must be synchronized across these counties to prevent the "ping-pong" effect of mosquito movement.
2. **Rainfall-Based Early Warning:** Given the predictive power of rainfall, anti-malarial supplies should be pre-positioned in the Lake Region whenever meteorological forecasts predict above-average seasonal rains.
3. **Resource Reallocation:** Funding should be shifted from blanket national coverage to targeted "Hotspot Elimination." Low-Low counties (Central) require only surveillance, freeing up resources for the high-burden West.

## REFERENCES

1. **GADM.** (2025). *Kenya Administrative Boundaries (Level 1)*. Retrieved from [https://gadm.org/download\\_country.html](https://gadm.org/download_country.html)
2. **Stats Kenya.** (2024). *Overall Poverty Rates in Kenya by County (2022)*. Retrieved from <https://statskenya.co.ke/at-stats-kenya/about/overall-poverty-rates-in-kenya-by-county/62/>
3. **Stats Kenya.** (2025). *Projected Population of Kenya 2025*. Retrieved from <https://statskenya.co.ke/at-stats-kenya/about/population-of-kenya-2025-population-by-county/110/>
4. **Ministry of Health.** (2024). *Kenya Malaria Indicator Survey Data*.

## APPENDIX: R Statistical Output

### Output 1: Spatial Lag Model (SLM) Summary

Call: lagsarlm(formula = Malaria\_Cases ~ Rainfall + Temperature + Poverty\_Rate,  
data = final\_cleaned\_data, listw = listw, zero.policy = TRUE)

Residuals:

Min	1Q	Median	3Q	Max
-44.6917	-15.1844	-7.6501	5.4555	162.7688

Type: lag

Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-141.7133	61.6469	-2.2988	0.0215 *
Rainfall	0.0438	0.0160	2.7442	0.0061 **
Temperature	3.5714	2.1446	1.6653	0.0958 .
Poverty_Rate	0.6124	0.3843	1.5935	0.1110

Rho: 0.37021, LR test value: 3.5439, p-value: 0.059765

Wald statistic: 4.8877, p-value: 0.027049

AIC: 475.67, (AIC for lm: 477.21)

### Output 2: OLS Regression Summary (Baseline Model)

Call:

lm(formula = Malaria\_Cases ~ Rainfall + Temperature + Poverty\_Rate,  
data = final\_cleaned\_data)

Residuals:

Min	1Q	Median	3Q	Max
-44.326	-18.462	-6.416	6.469	158.679

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-180.51192	66.97373	-2.695	0.009996 **
Rainfall			0.06117	0.01649 3.710
0.000591 ***				

Temperature 4.58419 2.35705 1.945 0.058347 .

Poverty\_Rate 0.74019 0.40646 1.821 0.075561 .

Residual standard error: 36.45 on 43 degrees of freedom

Multiple R-squared: 0.3127, Adjusted R-squared: 0.2647

F-statistic: 6.52 on 3 and 43 DF, p-value: 0.000983

### **Output 3: Lagrange Multiplier Diagnostics (Model Selection)**

Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial dependence

data: model: lm(formula = Malaria\_Cases ~ Rainfall + Temperature + Poverty\_Rate, data = final\_cleaned\_data)

test weights: listw

RSerr = 1.6157, df = 1, p-value = 0.2037 (Not Significant)

RSlag = 3.2371, df = 1, p-value = 0.0719 (Marginally Significant)

### **Output 4: Spatial Lag Model (SLM) Summary (Final Model)**

Call: lagsarlm(formula = Malaria\_Cases ~ Rainfall + Temperature + Poverty\_Rate,

data = final\_cleaned\_data, listw = listw, zero.policy = TRUE)

Residuals:

Min 1Q Median 3Q Max

-44.6917 -15.1844 -7.6501 5.4555 162.7688

Type: lag

Coefficients: (asymptotic standard errors)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -141.7133 61.6469 -2.2988 0.0215 \*

Rainfall 0.0438 0.0160 2.7442 0.0061 \*\*

Temperature 3.5714 2.1446 1.6653 0.0958 .

Poverty\_Rate 0.6124 0.3843 1.5935 0.1110

Rho: 0.37021, LR test value: 3.5439, p-value: 0.059765

Wald statistic: 4.8877, p-value: 0.027049

AIC: 475.67, (AIC for lm: 477.21)