Geographically Weighted Regression Model for Spatial Interpolation of PM2.5 Concentrations in Northern Thailand

AUTHOR: Thiraphat Oatthanith
SUPERVISOR: ASSOC. PROF. Dr.Sompop Moonchai

Abstract

This independent study analyzes the relationship between monthly and weekly PM2.5 concentration data and meteorological variables, including air pressure, temperature, and relative humidity, in Northern Thailand. The study utilizes data from 15 air quality monitoring stations of the Pollution Control Department and meteorological stations of the Thai Meteorological Department during March and April of 2023–2024. Pearson correlation analysis was employed to examine these relationships. Additionally, spatial PM2.5 concentration estimates at unknown locations were conducted using the Geographically Weighted Regression (GWR) model with Gaussian and exponential kernel functions. The results indicate that PM2.5 concentration strongly correlates with air pressure and relative humidity, while its correlation with maximum temperature is weaker. A model training dataset of 12 locations and a test dataset of 3 locations were used to evaluate the spatial interpolation performance of PM2.5 concentration estimation. The findings reveal that the GWR model with an exponential kernel function provided better predictive accuracy than the Gaussian one when using air pressure and relative humidity as independent variables. The mean absolute percentage errors (MAPE) for the weekly estimates in March and April of 2023 and 2024 range from approximately 28% to 70%, indicating a prediction accuracy that varies from low to moderate.

Introduction

Thailand faces ongoing challenges with PM2.5 pollution, especially in areas with high traffic, agricultural burning, or industrial activities. However, air quality monitoring is limited due to the high cost of equipment, maintenance, and personnel. This results in insufficient data coverage, particularly in remote or rural areas.

The lack of adequate monitoring stations leads to gaps in PM2.5 data, affecting policy decisions and timely responses to pollution events. Some high-risk areas may not receive proper warnings or mitigation measures.

This study aims to address these gaps by utilizing data from air quality monitoring stations across 15 provinces in northern Thailand. By incorporating advanced analytical methods, the research seeks to enhance PM2.5 data coverage, support more informed environmental policies, and contribute to better air quality management.

Main Result

Influencing factors

The monthly Pearson correlation coefficients showed a moderate to high correlation with air pressure and relative humidity, respectively, but a low correlation with maximum temperature. PM2.5 concentrations showed a high correlation with air pressure for March of both years, and a moderate to high correlation with relative humidity for March-April 2024.

The weekly Pearson correlation coefficients showed a moderate to high correlation with air pressure and relative humidity, respectively, but a low correlation with maximum temperature. PM2.5 concentrations showed a high correlation with air pressure for March of both years, and a moderate to high correlation with relative humidity for March-April 2024, as shown in the graph.

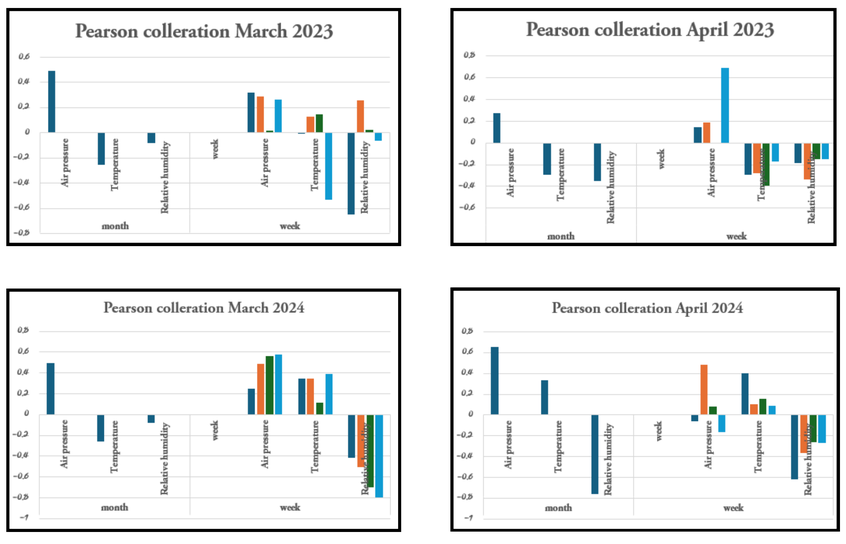

Figure 1. The graph shows the average Pearson correlation coefficient between PM2.5 concentration and independent variable data in 2023-2024.

Model Performance

The spatial interpolation of PM2.5 concentrations across 15 provinces in northern Thailand from March to April during the years 2023–2024 was estimated using the Geographically Weighted Regression (GWR) model with an exponential kernel function. Atmospheric pressure and relative humidity were used as explanatory variables. The study found that the mean absolute percentage error (MAPE) for each week in March and April 2023 was 56.4% and 141.1%, respectively. In March and April 2024, the MAPE was approximately 28.7% and 56.6%, respectively. Over the entire two-year period, the average MAPE was approximately 70.7%.

Year	Month	MAPE for Training Data	MAPE for Testing Data
2023	March	4.368	56.424
2023	April	5.782	141.106
2024	March	3.462	28.662
2024	April	3.135	56.567
	Average	4.187	70.689

Table 1. Mean of absolute percentage error of GWR with exponential kernel function for March and April 2023-2024 .

Objectives

- Study and analyze the relationship between meteorological variables, including temperature, relative humidity, and air pressure, and PM2.5 concentration in the upper northern region of Thailand over a two-year period from March to April 2023-2024.
- Create a geographic weighted regression model to estimate the spatial range of PM2.5 concentration in the northern region of Thailand over a two-year period from March to April 2023-2024.

Methodology

Data Collection

The study area encompassed fifteen provinces: Chiang Rai, Mae Hong Son, Chiang Mai, Lamphun, Lampang, Phayao, Nan, Phrae, Uttaradit, Sukhothai, Tak, Kamphaeng Phet, Phitsanulok, Phichit, and Phetchabun.

The daily data used in this study consisted of two primary components collected from March to April during the years 2023 to 2024:

- The meteorological variable data are daily data in March-April 2023-2024 from the weather station of the Meteorological Department in the northern region of Thailand.
- PM2.5 concentration measurements obtained from the Pollution Control Department.

Correlation Analysis

Let X and Y be random variables, where Y is the target variable and X is the chosen independent variable. From the observed values $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, we analyze the statistical relationship of the values as follows:

$$r = \frac{\sum((x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Accuracy Assessment

The mean absolute percentage error (MAPE) was employed to assess the accuracy of the method.

$$MAPE = \frac{\sum_{i=1}^{m}\left|\frac{\hat{Z}_i - Z_i}{Z_i}\right|}{n} \times 100$$

where $\hat{Z}_i$ represents the estimated data, Z_i represents the actual data, and m is the number of data points.

GWR Algorithm

Given n observed values, $\{X(s_i), Z(s_i)\}_{i=1}^{n}$, where $X(s_i) = [X_1(s_i), ..., X_p(s_i)]^T$ and a target point s_0

Step 1 : Distance Calculation

Calculate distance $s_0 = (u_0, v_0)$ and $s_0 = (u_k, v_k)$ for $k = 1, \ldots, n$ by

$$d_{0k} = \sqrt{(u_0 - u_k)^2 + (v_0 - v_k)^2}$$

Step 2 : Weight Calculation

Construct Weighted Matrix $W_0 = \begin{bmatrix} \alpha_{01} & 0 & \ldots & 0 \\ 0 & \alpha_{02} & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \alpha_{0n} \end{bmatrix}$

where a_{0k} is the weight between positions s_0 and s_k ,

, calculated using the formula $a_{0k} = K(d_{0k})$ for $k = 1, \ldots, n$.

Step 3 : Weight Least Squares Estimation

$$\boldsymbol{\beta} = (X_0^T W_0 X_0)^{-1} X_0^T W_0 Y \ ,$$

where $X_0 = \begin{bmatrix} 1 & x_1(s_0) & x_2(s_0) & \ldots & x_p(s_0) \\ 1 & x_1(s_1) & x_2(s_1) & \ldots & x_p(s_1) \\ 1 & x_1(s_2) & x_2(s_2) & \ldots & x_2(s_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1(s_n) & x_2(s_n) & \ldots & x_p(s_n) \end{bmatrix}$ and $Y = \begin{bmatrix} Z(s_0) \\ Z(s_1) \\ Z(s_2) \\ \vdots \\ Z_p(s_n) \end{bmatrix}$,

Step 4 : Model Estimation $Z(s_0) = \beta_0(s_0) + \sum_{k=1}^{p} \beta_k(s_0) x_k(s_0)$.

Conclusion

The results indicate that PM2.5 concentration exhibits a strong correlation with airpressure and relative humidity, while its correlation with maximum temperature is weaker.

The results of the average absolute percentage error range from 28%-70%. The results indicate a moderate to low level of accuracy. This is because some weeks have relatively high error values, which may be due to the small number of training data and testing data, and some weeks have relatively low correlations between PM2.5 concentration and air pressure and relative humidity. In addition, the choice of kernel function and bandwidth also affects the accuracy of the prediction