Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia

Vol. 10, 2022, page 29 - 41

Modeling COVID-19 Cases for Malaysia using Geographically Weighted Regression

Ayu Farhana Amir Hamzah, Shariffah Suhaila Syed Jamaludin*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: suhailasj@utm.my

Abstract

Coronavirus disease (COVID-19) has affected people in many ways. Since it has been a global health issue, it is crucial to identify the influence factors to curb the virus from spreading. This study aims to examine the influence of demographic and climate factors on the increase of COVID-19 cases in Malaysia and to model their relationship. Since the data are taken spatially, the issue of non-stationarity may exist. Thus, a Geographically Weighted Regression (GWR) model, an extension of the ordinary least square (OLS), is a valuable technique for exploring the degree to which the relationship between COVID-19 cases and the significant factors varies according to location and at different spatial scales. Thirty random districts over Peninsular Malaysia are considered in this study. Population demographic and climate factors such as temperature and humidity are used. Normality, multicollinearity, spatial autocorrelation, and stationarity tests were conducted as diagnostic checking. The findings demonstrated that GWR models provide a better fit and more geodata-oriented information than OLS models, based on the minimum Akaike Information Criterion (AIC). In addition, the total number of populations significantly influenced the number of COVID-19 cases reported in Malaysia while climate factors do not affect the number of COVID-19 cases.

Keywords: Geographically Weighted Regression; Ordinary Least Square; COVID-19; Stationarity

1. Introduction

COVID-19 is an ongoing global pandemic of Coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. The novel virus was first identified in the Chinese city of Wuhan in December 2019; a lockdown in Wuhan and other cities surrounding Hubei failed to contain the outbreak. It quickly spread to other parts of mainland China and worldwide [2].

The World Health Organization (WHO) declared a Public Health Emergency of International Concern on 30th January 2020 and a pandemic on 11th March 2020 [3]. Multiple virus variants have become dominant in many countries since 2021, with the Alpha, Beta, Gamma, and Delta variants being the most virulent. As of 20th November 2021, more than 256 million cases and 5.14 million deaths have been confirmed, making the pandemic one of the deadliest in history [4].

Malaysia began containment in early 2020. A movement control order (MCO) utilizing strong suppression measures was initiated on 18th March 2020. These measures were eased on 4th May 2020 and labeled a conditional movement control order (CMCO). On 9th June 2020, measures were further reduced into a recovery movement control order (RMCO) - a transitional phase before suppression measures were lifted entirely.

However, due to a rise in cases, measures to CMCO and MCO levels were intensified on 7th October 2020 and 13th January 2021, respectively. Restrictions were eased to CMCO on 5th March 2021 before another surge caused another intensification of measures to MCO on 11th May 2021. The intensity of these measures has varied at the state level after September 2020 [5].

Since the very beginning of the pandemic, many models have been proposed to understand the outbreak dynamics of COVID-19, especially using statistical modeling. Several statistical methods including the generalized linear model (GLM), spearman rank correlation, generalized additive model

(GAM), multivariate linear regression (MLR), Susceptible Infected Removed (SIR), Functional Data Analysis (FDA), and Geographically Weighted Regression (GWR) have been proposed by the researchers to model the COVID-19 cases.

This study will investigate the suitability of GWR model to examine the non-stationarity relationship between climate and demographic factors in COVID-19 cases. The GWR model will assist in capturing this variation by calibrating a multiple regression model that allows for different relationships at different points in space. In addition, a series of related statistical tests are considered.

2. Overview of COVID-19 and its analysis

2.1. Phenomena of COVID-19

COVID 19, the disease caused by the novel variant of the Sars-Cov-2 virus, still presents a clear and dynamic global threat. Despite the glimmers of hope offered by the rollout of several vaccines, the virus is still sweeping through many international communities [6]. In addition, multiple virus variants have become dominant in many countries since 2021, with the Alpha, Beta, Gamma, and Delta variants being the most virulent [7]. The global crisis following the outbreak of the Covid-19 pandemic has been the highest disruptive event in the world's history [8].

The world's economy and the daily life of all citizens have been impacted by Covid-19. Many businesses have endured drastic consequences with sudden disruptive changes. People had to adopt new patterns of behavior. In searching for solutions to this pandemic, the crisis has led to an unprecedented mobilization of the academic community to advance knowledge on the virus and its cure [9]. Furthermore, it is also reported that various activities could not be carried out normally when a COVID-19 pandemic crisis engulfed the world [10].

Activities may continue from home during a crisis using a smartphone through the internet because almost everyone has their smartphones without requiring additional hardware purchases. Teenagers or adults still studying in higher education institutions can undertake continuous learning for their work assignments. Although many details, the source of the virus and its ability to spread between individuals, remain unknown. An increasing number of cases have been confirmed to be caused by human-to-human transmission [11].

2.2. Symptoms of COVID-19

COVID-19 is a respiratory condition caused by a coronavirus, while some people are infected but do not notice any symptoms [12]. Furthermore, the signs and symptoms of coronavirus disease 2019 (COVID-19) may appear 2 to 14 days after exposure [13]. This time after exposure and before having symptoms is called the incubation period. Moreover, people with COVID-19 are said to have reported a wide range of symptoms, ranging from mild to severe illness [14].

In addition, the primary symptoms of COVID-19 include fever, dry cough, and fatigue [15]. Other than that, myalgia, arthralgia, and diarrhea can also classify as the general symptoms of COVID-19 [16]. Case reports and mainstream media articles from various countries indicate that several patients diagnosed with COVID-19 had developed anosmia, a loss of smell [17]. The loss of smell and taste is a potential predictor of COVID-19 in addition to the most established symptoms of a high temperature and a new, continuous cough [18]. There may also be a combination of at least two symptoms: chills, muscle pain, headache, and sore throat. Other than that, some people will experience severe symptoms such as shortness of breath, confusion, chest pain, and difficulty moving or talking in such cases [19].

2.3. Government precautions

The first case of COVID-19 in Malaysia was detected on 25th January 2020, involving three tourists from China. After that, the number of cases steadily increased before the nation's first two deaths were recorded on 17th March. As of 20th April 2020, Malaysia has recorded more than 5300 positive cases involving 89 deaths [20]. On the other hand, the first option against COVID-19 is the prevention method to prevent the spread of the virus, considering that there is currently no available vaccine or treatment at the first face of the pandemic [21].

The Malaysian Prime Minister enforced a Movement Control Order (MCO) on 18th March 2020 as a mitigation effort to reduce community spread and the overburdening of the country's health system. The MCO also restricted Malaysians from leaving the country, and all the foreigners from entry and non-essential sectors were ordered to close operations or allow employees to work from home [22]. Furthermore, the media actively spread the hashtag #stayathome, non-governmental organizations, as well as prison inmates, started to produce personal protective equipment for the frontline, and various organizations hosted fundraising events to provide essentials mainly to hospitals.

Moreover, the Government of Malaysia is committed to helping its citizens who are in need. Accordingly, numerous policies and programs have been introduced to cater to these needs, for example, 'Bantuan Sara Hidup' (Household Living Aid) or BSH, MySalam Scheme, 'Bantuan Awal Persekolahan' (Back to School), MyRapid Unlimited Pass, and Health Care Scheme for B40 [23].

2.4. Statistical analysis of COVID-19

2.4.1. Spearman Rank Correlation

Spearman rank correlation was used to analyze the association between COVID-19 and climate indicators in New York City [24]. The climate indicators included in their study are average temperature, minimum temperature, maximum temperature, rainfall, average humidity, wind speed, and air quality. It is found that the average temperature, minimum temperature, and air quality were significantly associated with the COVID-19 pandemic.

Moreover, the determination of online searches for COVID-19 related to international media announcements or national epidemiology [25]. Therefore, they studied the correlations using Spearman's rank correlation coefficient for Covid-19 epidemiology, such as incidence and mortality, with the national online searches. In their findings, the online searches for Covid-19 were not correlated with the actual incidence and mortality of Covid-19. In addition, online searches are not correlated with epidemiology but strongly correlated with global WHO announcements.

2.4.2. Linear Modeling

2.4.2.1. Generalized Linear Model (GLM)

The investigation on the effects of temperature, humidity, precipitation, wind speed, and the specific government policy intervention of partial lockdown on the new cases of COVID-19 infection in Ghana was conducted. The researchers used a time series generalized linear model, which allows for regressing past observations of the response variable and covariates for model fitting. The results indicate significant maximum temperature, relative humidity, and precipitation effects in predicting new disease cases [26].

Furthermore, the negative binomial regression model exhibited a best-fit model in building a death curve compared to the Poisson regression model obtained by the GLM method. Therefore, statistical methods such as chi-squared, ANOVA, logistic regression, Poisson, and negative binomial regression models were utilized to analyze Covid-19 data in Georgia. The results show that the difference among the mean ages of death for overall underlying conditions with p = 0.0248 and with 'no' and 'unknown' medical conditions with p = 0.0196 were found to be significant. Moreover, the covariates regions, minimum, maximum, and average age, were found to have a significant effect [27].

2.4.2.2. Generalized Additive Model (GLM)

The associations of daily average temperature (AT) and relative humidity (ARH) with the daily counts of COVID-19 cases in 30 Chinese Hubei provinces were tested. The generalized Additive Model (GAM) was fitted to quantify the province-specific associations between meteorological variables and the daily cases of COVID-19 during the study periods. In the model, the 14-day exponential moving averages (EMAs) of AT and ARH and their interaction were included with time trend and health-seeking behavior adjusted. In addition, their spatial distributions were visualized. AT and ARH showed significant negative associations with COVID-19 with a significant interaction between them in Hubei [28].

On the other hand, the study uses the log-linear GAM additive model to analyze the effects of temperature and relative humidity on daily new cases and daily new deaths of COVID-19 in 166 countries, excluding China. Their findings revealed that temperature and relative humidity negatively relate to new cases and deaths daily. For example, 1 °C increase in temperature was associated with a 3.08% reduction in daily new cases and a 1.19% reduction in daily new deaths. In contrast, a 1% increase in relative humidity was associated with a 0.85% reduction in daily new cases and a 0.51% reduction in daily new deaths. Furthermore, these findings provide preliminary evidence that the COVID-19 pandemic may be partially suppressed with temperature and humidity increases [29].

2.4.2.3. Multivariate Linear Regression (MLR)

The study on the impact of climate and urban factors on confirmed cases of COVID-19 uses multivariate linear regression (MLR) to propose a more accurate prediction model. Some important climate parameters, including daily average temperature, relative humidity, and wind speed, in addition to urban parameters such as population density, were considered, together impacts on confirmed cases of COVID-19 were analyzed. The analysis results show the proposed model's effectiveness and the impact of climate parameters on the trend of confirmed cases [30].

A valid global data set is collected from the WHO daily statistics, and the correlation among the total confirmed, active, deceased, and positive cases is stated in the research paper. Regression models such as Linear and Multiple Linear Regression techniques are applied to the data set to visualize the trend of the affected cases. A comparison of Linear Regression and Multiple Linear Regression model is performed where the score of the model R2 tends to be 0.99 and 1.0, which indicates a strong prediction model to forecast the coming days' active cases. Using the Multiple Linear Regression model as in July, the forecast value of 52,290 active cases is predicted for the next month of 15th June in India and 9,358 active cases in Odisha if the situation continues [31].

2.4.3. Susceptible Infected Removed (SIR)

The use of secondary and official data sources from Malaysia, Singapore, and Thailand to study the SIR model to predict potential cases and deaths from COVID-19 in the coming days was being investigated. In addition, the various aspects of COVID-19 and their relationships with various climatologic parameters and further forecasting were investigated. It was discovered that COVID- 19 spread and fatality rates are high globally, but when compared to tropical countries, it will be incredibly high in temperate countries due to lower temperatures (7-16°C) and humidity (80-90 percent) [32].

In addition, one of the research papers provided the estimations of the primary reproduction number (R0) and the per day infection mortality and recovery rates of COVID-19 using the SIR model to the reported data in China. According to the findings, the total number of infected people could reach 180,000 by 29th February, with a lower bound of 45,000. In terms of deaths, simulations predict that the death toll may exceed 2,700 as a lower bound by 29th February based on data reported up to 10th February. Furthermore, their analysis reveals a significant decrease in the case fatality ratio since 26th January, which could be attributed to various factors, including the strict control measures implemented in Hubei, China [33].

2.4.4. Functional Data Analysis (FDA)

A researcher modeled daily hospitalized, deceased, ICU cases, and return home patient numbers along the COVID-19 outbreak as functional data across different departments in France [34]. Their response variables are vaccinations, deaths, infection, recovery, and tests in France. Furthermore, these data sets were considered before and after France's vaccination began. Moreover, by comparing different countries with different vaccination rates and quantifying the phase of descent of the second component's curves, the study could undoubtedly demonstrate the predictive nature of this second component on the future success of a vaccination policy [34].

In addition, there is an investigation of the patterns of COVID-19 mortality across 20 Italian regions [35].They pinpoint significant trends by exploiting information in curves and shapes with Functional Data Analysis techniques. In addition, they could document strong associations of COVID-19 mortality with local mobility and positivity, which persist in models that control for other relevant covariates [35].

2.5. Geographically Weighted Regression (GWR)

GWR is used to model to create a linear regression model that generates local model parameter estimators for each observation location. The study's goal was to forecast the number of COVID-19 cases in Bandung from March to 4th June 2020. Variations in the variable population size, distance to the capital city, number of ODP, number of PDP, and several health facilities in the GWR model were found to explain variations in the number of COVID-19 positive cases in Bandung City [36].

One study uses COVID-19 cumulative cases as the dependent variable and the common factors as the independent variables in Texas. According to the virus prevalence hierarchy, the spatial-temporal disparity is categorized into four quarters in the GWR analysis model. The findings exhibited that GWR models provide higher fitness and more geodata-oriented information than OLS models [37].

In addition, exists an examination of how the relationships between structural inequalities and confirmed COVID-19 cases spatially vary across Arizona using a geographically weighted regression (GWR). They found that the structural inequality indicators and the presence of Native Americans are significantly associated with higher confirmed COVID-19 cases, and the relationships between structural inequalities and confirmed COVID-19 cases are significantly more potent in areas with a high concentration of Native Americans [38].

3. Methodology

3.1. Regression

Regression encompasses a wide range of methods for modeling the relationship between a dependent variable and a set of one or more independent variables. The dependent variable is sometimes known as the y-variable, the response variable or the regressand. A regression model is expressed as an equation. In its simplest form a linear regression model can take the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{for } i = 1 \dots n \tag{1}$$

In this equation y_i is the response variable, here measured at some location i, x_i is the independent variable, ε_i is the error term, and β_0 and β_1 are parameters which are to be estimated such that the value of $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is minimised over the n observations in the dataset. The $\hat{y}_i$ is the predicted or fitted value for the ith observation, given the ith value of $x$. The term $(y_i - \hat{y}_i)$ is known as the residual for the ith observation, and the residuals should be both independent and drawn identically from a Normal Distribution with a mean of zero. Such a model is usually fitted using a procedure known as Ordinary Least Squares (OLS).

More generally, a multiple linear regression model may be written as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} + \varepsilon_i \qquad \text{for } i = 1 \dots n \tag{2}$$

where the predictions of the dependent variable are obtained through a linear combination of the independent variables. The OLS estimator takes the form

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

where $\hat{\beta}$ is the vector of estimated parameters, X is the design matrix which contains the values of the independent variables and a column of 1s, y is the vector of observed values, and $(X^T X)^{-1}$ is the inverse of the variance-covariance matrix. The weights are placed in the leading diagonal of a square matrix W and the estimator is altered to include the weighting

$$\hat{\beta} = (X^T W X)^{-1} X^T W y \tag{4}$$

3.2. Geographically Weighted Regression (GWR)

In this technique, a particular variable, the dependent variable, is modelled as a linear function of a set of independent or predictor variables

$$y_i = a_o + \sum_{k=1,m} a_k x_{ik} + \varepsilon_i \tag{5}$$

where y_i is the ith observation of the dependent variable, x_{ik} is the ith observation of the kth independent variable, the ε_is are independent normally distributed error terms with zero means, and each a_k must be determined from a sample of n observations. Usually, the least squares method is used to estimate the a_ks. Using matrix notation this may be expressed as

$$\hat{a} = (x^t x)^{-1} x^t y \tag{6}$$

where the independent observations are the columns of x and the dependent observations are the single column vector y^1. The column vector $\hat{a}$ contains the coefficient estimates. Each of these estimates can be thought of as a "rate of change" between one of the independent variables and the dependent variable.

3.3. Criteria related to model performance

The bandwidth selection can be determined using the Akaike Information Criterion (AIC) finding from Hurvich et al., (1998). Minimizing the AIC provides a trade-off between goodness-of-fit and degrees of freedom. The AIC is defined for GWR as the following [39]

$$\mathrm{AIC}_c = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left\{\frac{n = tr(S)}{n - 2 - tr(S)}\right\} \tag{7}$$

where n is the sample size, $\hat{\sigma}$ is the estimated standard deviation of the error term, and $tr(S)$ refers to the trace of the hat matrix which is a function of the bandwidth. As the general rule, the lower the AIC, the closer the model approximation to the reality.

3.4. Diagnostic Checking

3.4.1. Spatial Autocorrelation

Spatial autocorrelation measures the similarity between samples for a given variable as a function of spatial distance. Furthermore, Moran's I coefficient is the most used in univariate autocorrelation analyses [40] and is given as

$$I = \left(\frac{n}{s}\right)\left[\frac{\sum_i \quad \sum_j (y_i - \bar{y})(y_j - \bar{y})w_{ij}}{\sum_i (y_i - \bar{y})^2}\right] \tag{8}$$

where n is the number of samples, y_i and y_j are the data values in quadrats i and j, $\bar{y}$ is the average of y and w_{ij} is an element of the spatial weight matrix W. Under the null hypothesis of no spatial autocorrelation, I has an expected value near zero for large n, with positive and negative values indicating positive and negative autocorrelation, respectively.

3.4.2. Spatial Non-stationarity

The Breusch-Pagan test is used to determine whether heteroscedasticity is present in a regression model [41]. The test uses the following null and alternative hypotheses.

- Null Hypothesis (H_0): Homoscedasticity is present (the residuals are distributed with equal variance)
- Alternative Hypothesis (H_1): Heteroscedasticity is present (the residuals are not distributed with equal variance)

If the *p*-value of the test is less than some significance level, then we reject the null hypothesis and conclude that heteroscedasticity is present in the regression model. In addition, these are the steps to perform this test.

1. Fit the regression model.
2. Calculate the squared residuals of the model.
3. Fit the new regression model, using the squared residuals as the response values,
4. Calculate the Chi-Square test statistics X^2 as

$$n * R^2{}_{new} \tag{9}$$

where n : The total number of observations.

$R^2{}_{new}$: The R-squared of the new regression model that used the squared residuals as the response values.

If the *p*-value that corresponds to this Chi-Square test statistic with p which is the number of predictors, degrees of freedom is less than some significant level then reject the null hypothesis and conclude that heteroscedasticity is present. Otherwise, fail to reject the null hypothesis. In this case, it is assumed that homoscedasticity is present.

3.5. Choice of Spatial weighting function

First, consider the implicit weighting scheme of the OLS framework in equation (3.1). Here

$$w_{ij} = 1, \qquad \forall i, j \tag{10}$$

where j represents a specific point in space at which data are observed and i represents any point in space for which parameters are estimated. That is, in the global model each observation has a weight of unity. An initial step towards weighting based on locality might be to exclude from the model calibration observations that are further than some distance d from the locality. This would be equivalent to setting their weights to zero, giving a weighting function of

$$w_{ij} = \begin{cases} 1, & if\ d_{ij} < d \\ 0, & otherwise \end{cases} \tag{11}$$

However, the spatial weighting function in equation (3.11) suffers the problem of discontinuity. As i varies around the study area, the regression coefficients could change drastically as one sample point moves into or out of the circular buffer around i, which defines the data to be included in the calibration for location i. Although sudden changes in the parameters over space might genuinely occur, in this case changes in their estimates would be artifacts of the arrangement of sample points rather than any underlying process in the phenomena under investigation. One way to combat this is to specify w_{ij} as a continuous function of d_{ij}, the distance between i and j. One obvious choice is

$$w_{ij} = exp\left\{-\frac{d_{ij}{}^2}{\beta^2}\right\} \tag{12}$$

where β is referred to as the bandwidth. If i and j coincide, the weighting of data at that point will be unity and the weighting of other data will decrease according to a Gaussian curve as the distance between i and j increases.

4. Results and discussion

4.1. Study Area and Data Collection

This study will analyze the pattern and possible relationships between the variables that may affect the Covid-19 cases in everyday life. The population from every district was obtained from the Department of Statistics Malaysia while temperature and humidity data are from the Malaysian Meteorological Department. In this study, we focus on 30 districts in Malaysia. The data was taken from 15th June 2021 to 28th June 2021, about two weeks of cumulative data, and can be observe as tabulated in Table 1.

Table 1: Cumulative cases data from 15th to 28th June 2021 over 30 selected districts.

Station	Cases	Station	Cases
Kangar	27	Klang	4853
Kota Setar	213	Sepang	1937
Kuala Muda	1614	Jerantut	185
Langkawi	63	Raub	19
Kuala Besut	51	Cameron Highlands	114

Station	Cases
Kemaman	6
Kuala Terengganu	174
Setiu	14
Hulu Terengganu	42
Hilir Perak	79
Manjung	639
Kuala Kangsar	8
Batang Padang	202
Petaling	8237
Gombak	3190
Kuantan	634
Seberang Perai Tengah	718
Timur Laut	368
Jempol	494
Kuala Pilah	609
Port Dickson	947
Batu Pahat	205
Kulai	268
Muar	589
Segamat	761

Table 1 explains the cumulative cases and tabulated from 15th to 28th June 2021 that are being tested in this study. Therefore, Petaling recorded the highest number of cases with 30,975 cases, followed by Klang with 14,802 cases.

4.2. Multicollinearity

Table 2: Results from Spearman correlation between temperature and humidity

	Humidity
Temperature	*p*-value = 0.0000 rho = -0.7361745

Table 3: Results from Spearman correlation between temperature and population

	Population
Temperature	*p*-value = 0.0520 rho = 0.00383609

Table 4: Results from Spearman correlation between humidity and population

	Population
Humidity	*p*-value = 0.0586 rho = -0.00373111

Based on Tables 2, 3, and 4, the *p*-value and rho (Spearman's correlation coefficient) were obtained between the climate and demographic factors. The *p*-value tabulated in Table 2, which is 0.0000 appear to be smaller than alpha, α, which is 0.05. Hence, there is enough evidence to reject the null hypothesis, H_0, which explains a relationship between temperature and humidity. Moreover, the negative value for rho shows that the variable temperature and humidity have a strong negative correlation since the value is approaching -1.

In addition, the *p*-value tabulated in Tables 3 and 4 appears to be bigger than the alpha. Therefore, there is not enough evidence to reject the null hypothesis, H0, which explains that no relationship exists between climate factors and the number of populations. Moreover, since there is no relationship or correlation between the variables in Tables 3 and 4, the rho value is approximately zero. In addition, since these climate factors are related, they will need to be separated in the next step of modeling.

4.3. Diagnostic Checking

Table 5: Diagnostic checking results for each model checking

Model	**Jacque-Bera**	**Breusch-Pagan**	**Moran's**
1	0.9428 (Accept H_0)	0.9428 (Accept H_0)	0.9428 (Accept H_0)
2	0.9428 (Accept H_0)	0.9428 (Accept H_0)	0.9428 (Accept H_0)

Model 1 – cases ~ temperature and population
Model 2 – cases ~ humidity and population

Table 5 shows the results from diagnostic checkings. Before proceeding with GWR analysis, the normality, stationarity, and spatial autocorrelation tests are implemented. The values tabulated are the *p*-values for both Model 1 and Model 2 of each test. Therefore, it can be proven that both models have normally distributed data based on the *p*-value, which is larger than 0.05.

Thus, there is strong evidence to accept Ho, and this result has already passed the normality test throughout the diagnostic checking for both models. In addition, the normality of both models can be observed by the graph plotted in Figure 1 and Figure 2 below. Hence, both graphs can be considered normal since all the points show a linear pattern.

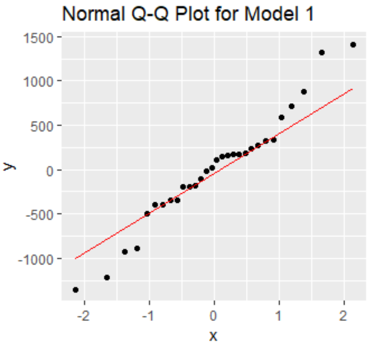

Figure 1: Normality graph for Model 1

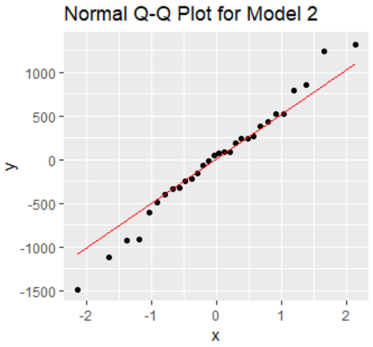

Figure 2: Normality graph for Model 2

Besides, the stationarity test based on Table 5 explained that there exists non-stationarity between the variables from both models since the p-value recorded was less than 0.05. This can also give the meaning of the residuals from the data distributed not equal variance. Thus, there is enough evidence to reject Ho, passing the second test in this study. Therefore, since GWR will be used if non-stationarity exists between the variables, the data from both models were suitable to be model.

In addition, a smaller p-value in both Moran's tests indicates that spatial autocorrelation exists between the variables, which means the data from both models were not randomly disbursed. Therefore, there is strong evidence to reject Ho since the *p*-value is smaller than 0.05. Thus, these results proved that it had already passed the spatial autocorrelation test throughout the diagnostic checking. Hence, the results showed that the diagnostic checking for both models was a success.

4.4. Ordinary Least Square

Table 6: Result from OLS for both Model 1 and Model 2

Model	**Variable**	**Estimate**	***p*-value**	**R-squared**
1	Intercept	**1474**	***0.435***	0.8563
	Temperature	-71.51	0.310	
	Population	0.003990	3.38e-13	
2	Intercept	1316	0.669	0.8524
	Humidity	-20.63	0.567	
	Population	0.003835	0.0000	

Based on the results in Table 6, this can be concluded that only the total population has a significant relationship to the number of COVID-19 cases since *p*-values for both Model 1 and Model 2 are smaller than 0.05 compared to other variables. In addition, since the value for *R*-square determine the fitness of the data, both Model 1 and Model 2 appear to show good fitness of the dataset because the results approaching 1. Thus, the formulated equation obtained for Model 1 can be written as

$$y_{(cases)} = 1474 - 71.51x_{(temperature)} + 0.00399x_{(population)} \tag{4.1}$$

While for the tabulated equation for Model 2 is written as

$$y_{(cases)} = 1316 - 20.63x_{(humidity)} + 0.003835x_{(population)} \tag{4.2}$$

4.5. Geographically Weighted Regression

Table 7: Results from GWR for both Model 1 and Model 2

Model	Regression	AIC$_c$	Adjusted R-squared	Bandwidth
1	OLS	480.925992	*0.850746*	1.056
	GWR	476.654408	0.894497	
2	OLS	481.723572	0.846725	1.000
	GWR	467.685908	0.928602	

The aim of this modeling is to compare the value of AIC and *R*-squared between both global regression and GWR method. The best model is chosen based on the minimum AIC value with larger R square. Based on Table 7, the value for AIC$_c$ and adjusted *R* square will be used. The results indicated that the number of AICc for Model 1 and Model 2 was smaller using GWR than global regression. Moreover, the value for adjusted R square was also larger in GWR than the global regression. In other word, since GWR is the extension from OLS, this can be concluded that GWR gives better and more accurate results compared to OLS. The parameter estimation can be obtained from the GWR output for both Model 1 and Model 2 and is given in Table 8.

Table 8: Estimated parameters of GWR models

Model	Variable	Estimate
1	Intercept	1474.1135
	Temperature	-71.5051
	Population	0.004
2	Intercept	1315.6284
	Humidity	-20.6323
	Population	0.0038

Hence, the formulated equation obtained from GWR can be written as

Model 1 $$y_{(Cases)} = 1474.1135\,(u_i\,, v_i) - 71.5051\,(u_i\,, v_i).\,x_{(Temperature)} + 0.004\,(u_i\,, v_i).\,x_{(Population)} + \varepsilon_i \tag{4.3}$$

Model 2

$$y_{(Cases)} = 1315.6284\,(u_i\,, v_i) - 20.6323(u_i\,, v_i).\,x_{(Humidity)} + 0.0038\,(u_i\,, v_i).\,x_{(Population)} + \varepsilon_i \tag{4.4}$$

To summarize, the number of Covid Cases at unknown location can be predicted based on the equations which is given in Model 1 and 2. In addition, the number of populations appear to be the only variables that related to the number of COVID-19 cases from 15th June 2021 to 28th June 2021.

Conclusion

Three diagnostic checking, namely, Jacque Bera, Breusch Pagan and Moran's tests are evaluated to check the existence of non-stationarity and autocorrelation between the variables before modeling the data into GWR. Since there exists multicollinearity between temperature and humidity, there are two models to be considered throughout the study. The study found that both models excel all the diagnostic checking and qualify to be model into GWR. In addition, the findings revealed that GWR is more efficient than the global regression, OLS based on the minimum AIC and the value of adjusted *R*-square. Total population is the significant factor that has a significant impact to the number of Covid cases.

Acknowledgement

The authors would like to express their gratitude to the Ministry of Higher Education (MOHE) for the funding under the Fundamental Research Grant Scheme (FRGS/1/2020/STG06/UTM/02/3) under vote 5F311. We are also grateful to Universiti Teknologi Malaysia for supporting this project with Research University Grant (QJ130000.3854.19J58).

References

[1] Sharma, A., Tiwari, S., Deb, M. K., &amp; Marty, J. L. 2020. Severe acute respiratory syndrome coronavirus-2 (SARS-COV-2): *A global pandemic and treatment strategies. International Journal of Antimicrobial Agents*, 56(2), 106054.

[2] Lai, S., Bogoch, I. I., Ruktanonchai, N. W., Watts, A., Lu, X., Yang, W., Yu, H., Khan, K., &amp; Tatem, A. J. 2020. *Assessing spread risk of Wuhan novel coronavirus within and beyond China, January-April 2020*: A travel network-based modelling study.

[3] *COVID-19 (Novel Coronavirus).* Dynamed.com. 2022.

[4] Islam, S., Islam, T., &amp; Islam, M. R. 2022. *New coronavirus variants are creating more challenges to global healthcare system: A brief report on the current knowledge*. Clinical Pathology, 15.

[5] Jayaraj, V. J., Rampal, S., Ng, C.-W., &amp; Chong, D. W. 2021. *The epidemiology of covid-19 in Malaysia*. The Lancet Regional Health - Western Pacific, 17, 100295.

[6] World Health Organisation. Weekly Epidemiological. 2021.

[7] Wikimedia Foundation. 2021, 27th November. *Covid-19 pandemic*.

[8] Fassin, Y. *Research on Covid-19: a disruptive phenomenon for bibliometrics*. Scientometrics 126, 5305–5319 2021.

[9] Haghani, M., Bliemer, M.C.J. *Covid-19 pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: Scientometric comparisons across SARS, MERS and 2019-nCoV literature*. Scientometrics 125, 2695–2726 2020.

[10] Susanto, H., Fang Yie, L., Mohiddin, F., Rahman Setiawan, A. A., Haghi, P. K., & Setiana, D. 2021. *Revealing social media phenomenon in time of covid-19 pandemic for boosting start-up businesses through Digital Ecosystem*. Applied System Innovation, 4(1), 6.

[11] Li, Q., et.al, 2020. *Early Transmission Dynamics in Wuhan, China, of novel coronavirus–infected pneumonia*. New England Journal of Medicine, 382(13), 1199–1207.

[12] Pathak, N. 2021, 2nd July. Symptoms of coronavirus: Early signs, serious symptoms and more. WebMD.

[13] Mayo Foundation for Medical Education and Research. 2021, 23rd November. *Coronavirus disease 2019 (covid-19)*. Mayo Clinic.

[14] Centers for Disease Control and Prevention. (n.d.). *Symptoms of COVID-19*. Centers for Disease Control and Prevention.

[15] Wang, H.-Y., Li, X.-L., Yan, Z.-R., Sun, X.-P., Han, J., & Zhang, B.-W. 2020. *Potential neurological symptoms of COVID-19*. Therapeutic Advances in Neurological Disorders, 13, 175628642091783.

[16] Elibol, E. *Otolaryngological symptoms in COVID-19*. Eur Arch Otorhinolaryngol 278, 1233–1236 2021.

[17] Gane, S. B., Kelly, C., & Hopkins, C. 2020. *Isolated sudden onset anosmia in covid-19 infection. A novel syndrome?* Rhinology Journal, 58(3), 299–301.

[18] Menni, C., et.al, 2020. *Real-time tracking of self-reported symptoms to predict potential COVID-19*. Nature Medicine, 26(7), 1037–1040.

[19] Pathak, N. 2021, 2nd July. *Symptoms of coronavirus: Early signs, serious symptoms and more*. WebMD.

[20] Thomson Reuters. 2020, 25th January. *Malaysia confirms first cases of coronavirus infection*. Reuters.

[21] Alshammari, T. M., Altebainawi, A. F., & Alenzi, K. A. 2020. *Importance of early precautionary actions in avoiding the spread of covid-19: Saudi Arabia as an example*. Saudi Pharmaceutical Journal, 28(7), 898–902.

[22] Azlan, A. A., Hamzah, M. R., Sern, T. J., Ayub, S. H., & Mohamad, E. 2020. *Public knowledge, attitudes and practices towards covid-19: A cross-sectional study in Malaysia*. PLOS ONE, 15(5).

[23] Prime minister's Office of Malaysia. Prime Minister's Office of Malaysia. (n.d.).

[24] Bashir, M. F., Ma, B., Bilal, Komal, B., Bashir, M. A., Tan, D., & Bashir, M. 2020. *Correlation between climate indicators and covid-19 pandemic in New York, USA*. Science of The Total Environment, 728, 138835.

[25] Kumar, G., & Kumar, R. R. 2020. *A correlation study between meteorological parameters and covid-19 pandemic in Mumbai, India*. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(6), 1735–1742.

[26] Iddrisu, W. A., Appiahene, P., & Kessie, J. A. 2020, April 28. *Effects of weather and policy intervention on covid-19 infection in Ghana*.

[27] Khan, H. 2020. *Covid-19 epidemic models: A study from Georgia State in the USA.* American Journal of Biomedical Science & Research, 10(3), 295–302.

[28] Qi, H., et.al, 2020. *Covid-19 transmission in mainland China is associated with temperature and humidity: A Time-series analysis*. Science of The Total Environment, 728, 138778.

[29] Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., Du, M., & Liu, M. 2020. *Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries*. Science of The Total Environment, 729, 139051.

[30] Pirouz, B., Shaffiee Haghshenas, S., Pirouz, B., Shaffiee Haghshenas, S., & Piro, P. 2020. *Development of an assessment method for investigating the impact of climate and urban parameters in confirmed cases of COVID-19: A new challenge in sustainable development.* International Journal of Environmental Research and Public Health, 17(8), 2801.

[31] Rath, S., Tripathy, A., & Tripathy, A. R. 2020. *Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model*. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(5), 1467–1474.

[32] Hasan, N. A., & Haque, M. M. 2020. *Predict the next moves of covid-19: Reveal the temperate and tropical countries scenario*.

[33] Anastassopoulou, C., Russo, L., Tsakris, A., & Siettos, C. 2020. *Data-based analysis, modelling and forecasting of the COVID-19 Outbreak*. PLOS ONE, 15(3).

[34] Oshinubi, K., Ibrahim, F., Rachdi, M., & Demongeot, J. 2021. *Functional Data Analysis: Transition from daily observation of covid-19 prevalence in France to functional curves.*

[35] Boschi, T., Di Iorio, J., Testa, L. et al. *Functional data analysis characterizes the shapes of the first COVID-19 epidemic wave in Italy*. Sci Rep 11, 17054 2021.

[36] Marhamah, E., & Mindra Jaya, G. N. (2020). *Modeling positive covid-19 cases in Bandung city by means geographically weighted regression.* Communications in Mathematical Biology and Neuroscience.

[37] Wu, X., Zhang, J. *Exploration of spatial-temporal varying impacts on COVID-19 cumulative case in Texas using geographically weighted regression (GWR).* Environ Sci Pollut Res 28, 43732–43746 2021.

[38] Yellow Horse, A.J., Yang, TC. & Huyser, K.R. *Structural Inequalities Established the Architecture for COVID-19 Pandemic Among Native Americans in Arizona: a Geographically Weighted Regression Perspective.* J. Racial and Ethnic Health Disparities 2021.

[39] Fotheringham, A. S., Charlton, M. E., & Brunsdon, C. 1998. *Geographically weighted regression: A natural evolution of the expansion method for Spatial Data Analysis*. Environment and Planning A: Economy and Space, 30(11), 1905–1927.

[40] Wang, Q., Ni, J., & Tenhunen, J. 2005. *Application of a geographically-weighted regression analysis to estimate net primary production of Chinese Forest Ecosystems*. Global Ecology and Biogeography, 14(4), 379–393.

[41] Zach. 2020, 31st December. *The breusch-pagan test: Definition & Example*. Statology.