Practice Research Paper on Geographically Weighted Regression on Jakarta COVID-19 cases (Regressors)

Kwek Yi Chen
Singapore Management University
yichen.kwek.2019@smu.edu.sg

Ngah Xin Yan
Singapore Management University
xinyan.ngah.2019@smu.edu.sg

Toh Jun Long
Singapore Management University
junlong.toh.2019@smu.edu.sg

ABSTRACT

COVID-19 has become an indisputable part of our daily life ever since the virus spread to the majority of the world. Some countries are able to keep the situation under control, while some suffer the devastating effects from it. Among Asia countries, Indonesia has the highest COVID related mortality and positive rates for COVID-19 cases (Worldometers, n.d.), mainly in Jakarta, the main capital. This is likely due to certain underlying common factors within the countries. Researchers claimed that Jakarta could have as many as 4.7 million people who are possibly infected by the virus in March 2021 (Sood, 2021). This is alarming as this number constitutes to "nearly half" of Jakarta's population. After our group was made aware about the seriousness of this matter, we decided to come up with Regressors, a Geographically Weighted Regression (GWR) application, to investigate the impacts of various variables (independent variable) on the mortality and positive rates (dependent variable).

This application aims to allow users to import a dataset of their preference and use it to identify the relationship between the selected independent variables, such as proximity to healthcare facilities and proximity to attraction, and the dependent variable, such as Number of positive COVID cases. Functions include Exploratory Data Analysis (EDA), GWR, GWR prediction model. EDA visualizes the different variables on spatial point map and Histogram. GWR builds a GWR model based on selected dependent and independent variables, provides analysis on their relationship, thus allowing users to select the best parameters for the GWR base model. Additionally, users are able to visualize the accuracy of the model geographically. GWR prediction model is built based on selected dependent and independent variables with the dataset provided. The output is the predicted values which will be analyzed and visualized geographically on the interactive map.

1. MOTIVATION OF THE APPLICATION

The insufficient amount of GWR applications to collate insights of the various variables effectively and having an user-friendly application to run a wide range of GWR models with different configurations is what drives our research and application developing process. The goal is to provide researchers and the government with an application consisting of all the GWR methods in one. To calibrate the model with the best parameter found through the analysis of the relationship In more details, it aims to achieve the following requirements:

- To be able to understand the data by visualizing the individual variables on a spatial point map and histogram.
- To calibrate a GWR model to choose a statistically significant independent variable to test the dependent variable.
- Prediction with selected statistically significant independent variables in relation to dependent variables.

2. REVIEW AND CRITIC ON PAST WORKS

2.1 Case Study 1: Geographically weighted regression (GWR) analysis on the death incidence by COVID-19 in São Paulo, Brazil

2.1.1 Objective

To gain an understanding of how socio-spatial behaviour causes COVID-19 transmission in the most impacted area in Brazil.

2.1.2 Methodology Used

- Spearman correlation test
- Adjusted R2
- Ordinary least squares (OLS)
- Spatial error model (SEM)
- Spatial lag model (SLM)
- Geographically weighted regression (GWR)
- Multiscale geographically weighted regression (MGWR)

2.1.3 Learning Point

The results showed that GWR model well represented the spatial distribution of COVID-19 cases in São Paulo, successfully highlighting the impact of geospatial factors in GWR model.

Additionally, as the study was conducted specifically in São Paulo, our project intends to proceed in a similar fashion by investigating the impact of geographical spatial factors that would affect the positive cases in our GWR model. Our project intend to adopt the study's use of performance measures such as Adjusted R2, while using OLS for the GWR model.

2.2 Case Study 2: Geographically varying relationships of COVID-19 mortality with different factors in India

2.2.1 Objective

To understand the relationship geographically for how different driving factors affect COVID-19 deaths.

2.2.2 Methodology Used

- Variance Inflation Factor (VIF): to get rid of unnecessary redundancy in explanatory variables.
- Ordinary least squares (OLS)
- Geographically weighted regression (GWR)

2.2.3 Learning Point

The use of map visualization for the model on the localized level helps to identify possible geographical spatial point patterns, such as clustering. This could indicate possible higher correlations of selected variables in certain regions than other regions. Our project will be using this as an inspiration for plotting the model's performance across the regions to highlight possible spatial point patterns.

3. DESIGN FRAMEWORK

Regressors are designed to be informative and user friendly without being over cluttered. The three main visualizations include: interactive maps, text outputs and graphs.

Figure 3.1: Tab navigation

In our application, we implemented the use of tabs for easy navigation between the app's 3 main functions: EDA, GWR and GWR Prediction as seen in Figure 3.1. Followed by sub tabs containing the different analysis for the selected function.

3.1 EDA

3.1.1 Spatial Point Map

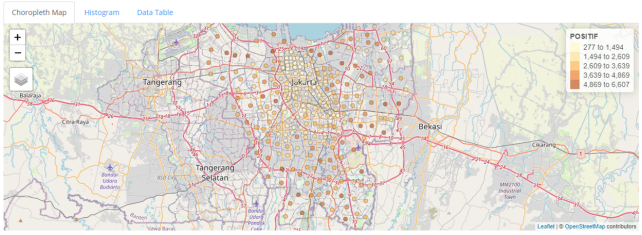

Figure 3.2: Spatial Point Map of variable selected

Figure 3.2 is a spatial point map that shows the selected variables geographically. Sub-districts with a spatial point of darker shades indicate that it has a higher number of positive cases whereas sub-districts with lighter shade spatial point indicate that it has a lower number of positive cases.

Users are able to select the variables they want to visualise, and classification method. Classification method options include jenks, equal, pretty, sd, fixed, bclust, fisher, hclust, kmeans and quantile.

3.1.2 Histogram

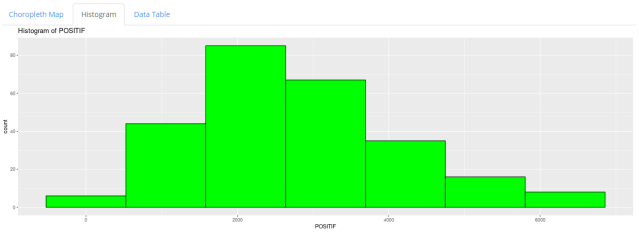

Figure 3.3: Histogram of selected variable

Figure 3.3 shows the distribution of the selected variable.

Users are able to choose the number of bins by adjusting the slider and change the color of the histogram to their preference using the Histogram Fill function.

3.2 GWR

3.2.1 Correlation Plot

Figure 3.4: Correlation plot of independent variables

Figure 3.4 is a correlation plot that explains the correlation between the independent variables the user has selected. A correlation value of greater than 0.75 is a good indication of

high correlation, as such one of the two variables should be removed from the analysis.

Users are able to select/deselect the independent variables in the list using the checkbox to see the correlation of the selected independent variables. Users can also choose the correlation order, method and type using the sidebar select option. Correlation order options include "AOE," "FPC," "hclust" and "alphabet." Correlation model options include "circle," "square," "ellipse," "number," "shade," "color" and "pie." Correlation type options include "full," "upper" and "lower."

3.2.2 *Summary*

```
Call:
stats::lm(formula = formula_reactive, data = uploaded_data())

Residuals:
    Min      1Q  Median      3Q     Max
-2782.6  -803.0  -108.1   672.7  3262.8

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1600.16     181.20   8.831  < 2e-16 ***
PROX_ATTRACTION   218.44      53.89   4.054 6.70e-05 ***
PROX_RESTAURANT   376.63      94.71   3.977 9.11e-05 ***
PROX_MALL         107.17      51.76   2.071  0.03940 *
PROX_HEALTHCARE  -273.80      82.56  -3.316  0.00104 **
PROX_RAILWAYS      59.40      23.91   2.484  0.01363 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1192 on 255 degrees of freedom
Multiple R-squared:  0.246,     Adjusted R-squared:  0.2312
F-statistic: 16.64 on 5 and 255 DF,  p-value: 3.237e-14
```

Figure 3.5: Summary statistics of independent variables

Figure 3.5 is a text output that provides key statistical information of the selected independent variables, such as the significance level, adjusted R-square and p-value. Based on the p-value provided for each variable, we are able to determine if the variable is statistically insignificant, thus should be removed from the analysis.

Users are able to select/deselect the independent variable(s) in the list using the checkbox.

3.2.3 *Multicollinearity*

	Variables	Tolerance	VIF
1	PROX_ATTRACTION	0.7814029	1.279750
2	PROX_RESTAURANT	0.5462111	1.830794
3	PROX_MALL	0.6916343	1.445851
4	PROX_HEALTHCARE	0.5780983	1.729810
5	PROX_RAILWAYS	0.7561808	1.322435

Figure 3.6: Multicollinearity of indpendent variables

Figure 3.6 shows the VIF value of the independent variables, which checks the multicollinearity. An independent variable with VIF value greater than 10 indicates multicollinearity and should be removed from the analysis.

Users are able to select/deselect the independent variable(s) in the list using the checkbox to see the VIF value of the selected independent variables.

3.2.4 *Linearity*

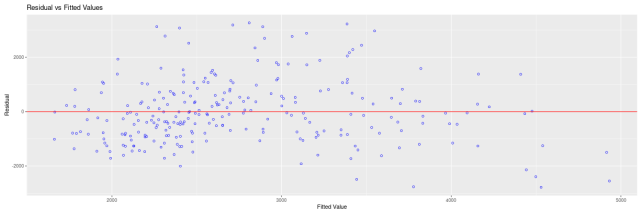

Figure 3.7: Residual plot for linearity assumption

Figure 3.7 checks for the linearity of the residuals of the calibrated model.

Users are able to select/deselect the independent variables in the list using the checkbox to see the scatterplot of the residual of the calibrated model.

If the points are scattered along the zero line (red), it shows a clustering distribution, which means the calibrated model is linear.

3.2.5 *Normality*

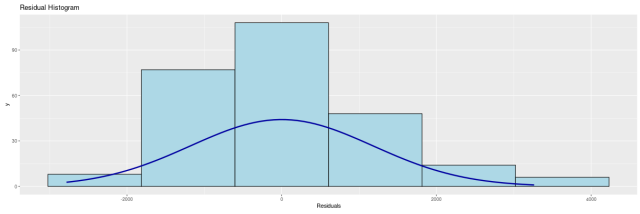

Figure 3.8: Residual binned distribution for Normality assumption

Figure 3.8 shows the linearity of the residuals of the calibrated model.

Users are able to select/deselect the independent variables in the list using the checkbox to see the histogram of the residual of the calibrated model.

3.2.6 *Base Model's Performance*

```
************************************************************************
*          Results of Geographically Weighted Regression               *
************************************************************************

*********************Model calibration information*********************
Kernel function: gaussian
Fixed bandwidth: 5790.332
Regression points: the same locations as observations are used.
Distance metric: Euclidean distance metric is used.

***************Summary of GWR coefficient estimates:******************
                     Min.   1st Qu.    Median   3rd Qu.    Max.
Intercept        237.1950  462.1771  837.8348 1569.9620 3075.04
PROX_ATTRACTION   -8.2455  273.3619  330.1201  383.5645  477.74
PROX_RESTAURANT  -94.2505  204.5687  367.4870  515.5377 1419.73
PROX_MALL       -176.9493  137.6075  209.4173  280.6602  367.48
PROX_HEALTHCARE -897.1231 -318.3257 -125.7860  -19.0274  126.31
PROX_RAILWAYS    -46.4377   14.6822   76.7124  133.0701  190.03
**********************Diagnostic information************************
Number of data points: 261
Effective number of parameters (2trace(S) - trace(S'S)): 37.26864
Effective degrees of freedom (n-2trace(S) + trace(S'S)): 223.7314
AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 4392.521
AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 4354.728
BIC (GWR book, Fotheringham, et al. 2002,GWR p. 61, eq. 2.34): 4222.305
Residual sum of squares: 241770147
R-square value:  0.4965624
Adjusted R-square value:  0.4123244

************************************************************************
Program stops at: 2021-11-18 18:13:44
```

Figure 3.9: Statistical information of base model

Figure 3.9 shows the GWR's model performance. The adjusted R-square value explains how much the calibrated model is able to explain the dependent variable selected. Even though the current model is only able to explain 41% of the dependent variable, it is still accurate because only significant independent variables were chosen. Other independent variables and other GWR configurations can be considered to calibrate a model with a higher adjusted R-square.

Users are able to select/deselect the independent variable(s) in the list using the checkbox. Users can also choose the bandwidth, approach, kernel and distance metric using the sidebar select option. Bandwidth options include "Fixed" and "Adaptive." Approach options include "Cross Validation (CV)" and "Akaike information Criterion (AIC)." Kernel options include " "Gaussian, "Exponential," "Bisquare," "Tricube," "Boxcar." Distance metric options include "Euclidean," "Great Circle." This configurations can be applied to the 3.2.7, 3.3.1 and 3.3.2

3.2.7 Visualization

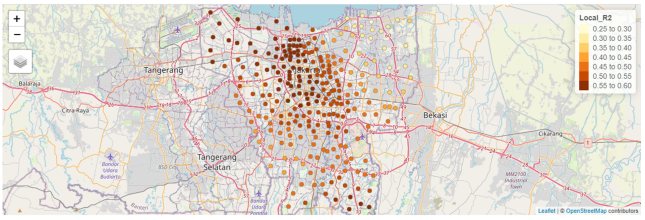

Figure 3.10: Map distribution of local R2

Figure 3.10 is a interactive map that shows the distribution of the local R2 score geographically, points with a darker shade indicates higher local R2 value, which then indicates higher explainability of the GWR model for the region, while points with a lighter shade indicates lower local R2 value, which then indicates poorer explainability of the model for the region.

3.3 Prediction

3.3.1 Prediction Model's Performance

```
************************************************************************
*     Results of Geographically Weighted Regression for prediction     *
************************************************************************

*********************Model calibration information*********************
Kernel function: gaussian
Fixed bandwidth: 5790.332
Distance metric: Euclidean distance metric is used.

***************Summary of GWR coefficient estimates:******************
                              Min.    1st Qu.     Median    3rd Qu.    Max.
Intercept_coef          222.9880   458.6119   829.3407  1536.0697 3027.39
PROX_ATTRACTION_coef      4.8857   266.6092   331.6605   382.7814  473.28
PROX_RESTAURANT_coef   -166.7706   197.8838   379.5288   511.7744 1494.88
PROX_MALL_coef         -158.8877   146.6247   209.4273   281.2723  388.18
PROX_HEALTHCARE_coef -1101.1131  -325.2831  -126.4679   -18.8950  127.70
PROX_RAILWAYS_coef     -58.1249    15.0024    75.9123   132.0846  231.29

***************       Results of GW prediction       ******************
                     Min.   1st Qu.    Median   3rd Qu.      Max.
prediction       -298.5    2027.4    2609.0    3298.8    5532.4
prediction_var 1089807.0 1106963.9 1123052.7 1178036.4 3443111.3

************************************************************************
Program stops at: 2021-11-18 18:37:42
```

Figure 3.11: Statistical information of prediction model

Figure 3.11 shows the prediction model's performance. Information such as min, max, median are provided to understand statistics of the predicted value.

3.3.2 Visualization

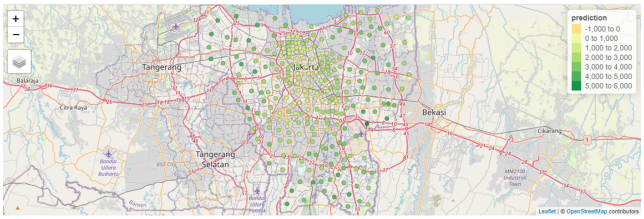

Figure 3.12: Map distribution of predicted values

Figure 3.12 is an interactive map that shows the distribution of the prediction values geographically, points with a darker shade indicates higher predicted value, while points with a lighter shade indicates lower predicted value.

4. USE CASE

4.1 POSITIF (Positive) COVID-19 cases against independent variables

To analyse the impacts of COVID-19 in Jakarta, the number of positive (POSITIF) cases has been identified as the dependent variable to be analysed with the independent variables. Before building a Geographically Weighted Regression (GWR) model, it is important to explore the variables using Exploratory Data Analysis (EDA).

Figure 4.1: Spatial Point distribution of positive cases in Jakarta

In Figure 4.1, the interactive spatial point map shows the number of positive cases geographically in each sub-district (kelurahan) of Jakarta. It is observed that there are generally more positive cases around the outer boundary of Jakarta, whereas there are fewer positive cases in the central area of Jakarta. This indicates that the spread of the virus could have came from neighbouring provinces near the boundary of Jakarta which resulted in the high number of positive cases near the outer boundaries in Jakarta.

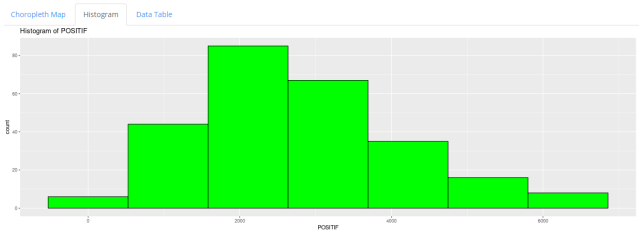

Figure 4.2: Histogram plot of number of positive cases

Figure 4.2 shows the distribution of the number of positive cases in Jakarta. Majority of the sub-district has around 2000 to 3000 positive cases. It is shown that there are sub-districts with zero positive cases which is the lowest among all the other sub-districts. The highest number of positive cases found in some sub-districts is above 6000.

After exploring the variables with EDA, the GWR model can be built to identify the relationship between the independent variables and the dependent variable, which in this case is the number of positive COVID cases. Before building the GWR model, correlation analysis should be performed to highlight the highly correlated variables.

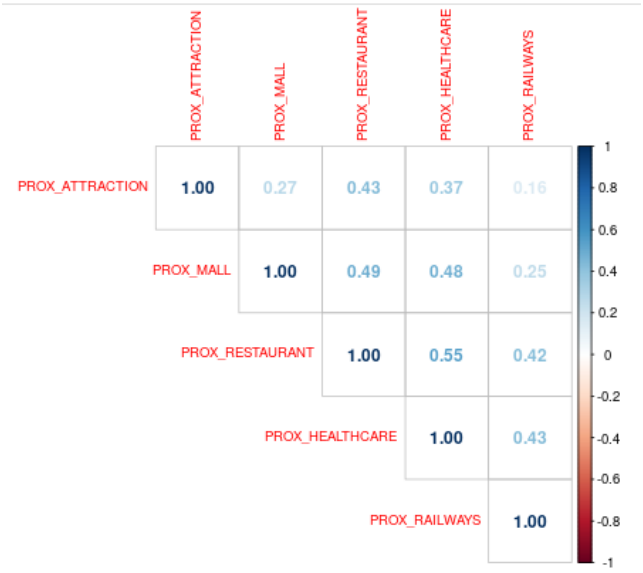

Figure 4.3: Correlation plot of proximities

In Figure 4.3, the relationship between the independent variables can be seen. Majority of them are not highly correlated with each other with correlation values lower than 0.75. Therefore, these independent variables may be considered for the model after calibration.

```
Call:
stats::lm(formula = formula_reactive, data = uploaded_data())

Residuals:
    Min      1Q  Median      3Q     Max
-2782.6  -803.0  -108.1   672.7  3262.8

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       1600.16     181.20   8.831  < 2e-16 ***
PROX_ATTRACTION    218.44      53.89   4.054 6.70e-05 ***
PROX_RESTAURANT    376.63      94.71   3.977 9.11e-05 ***
PROX_MALL          107.17      51.76   2.071  0.03940 *
PROX_HEALTHCARE   -273.80      82.56  -3.316  0.00104 **
PROX_RAILWAYS       59.40      23.91   2.484  0.01363 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1192 on 255 degrees of freedom
Multiple R-squared:  0.246,     Adjusted R-squared:  0.2312
F-statistic: 16.64 on 5 and 255 DF,  p-value: 3.237e-14
```

Figure 4.4: Highlight of proximities significance level

Figure 4.4 shows the independent variables that are statistically significant at 95% confidence level are PROX_ATTRACTION, PROX_RESTAURANT, PROX_MALL, PROX_HEALTHCARE and PROX_RAILWAYS. These variables have p-value below 0.05. The p-value of the model, 3.237e-14 is less than 0.05 which means the model is a good estimator of the number of positive covid cases in Jakarta.

To improve the accuracy of the model, it is important to

choose independent variables that are statistically significant. After a few iterations of calibrating the model, independent variables that are not statistically significant were removed.

	Variables	Tolerance	VIF
1	PROX_ATTRACTION	0.7814029	1.279750
2	PROX_RESTAURANT	0.5462111	1.830794
3	PROX_MALL	0.6916343	1.445851
4	PROX_HEALTHCARE	0.5780983	1.729810
5	PROX_RAILWAYS	0.7561808	1.322435

Figure 4.5: VIF values of proximity variables

Figure 4.5 shows the VIF value of the independent variables. The selected independent variables are all below 10, thus no there is no sign of multicollinearity and no further independent variables are needed to be removed.

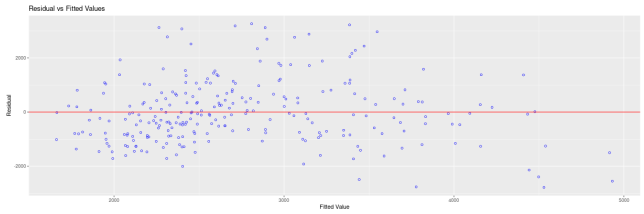

Figure 4.6: Linearity check of variables

Figure 4.6 checks the linearity of the residuals of the calibrated model. The above scatter plot shows that most points centers around the zero line with the exception of a batch of minority that are scattered further away from the line. Thus it is safe to assume model linearity.

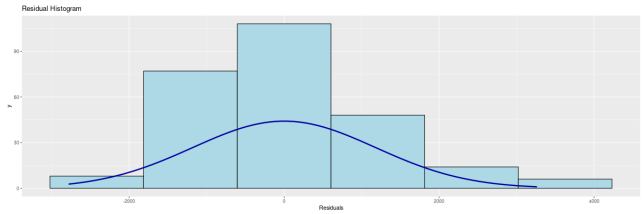

Figure 4.7: Normality check of variables

Figure 4.7 checks the normality of the residuals of the calibrated model. The above graph shows some resemblance to a normal distribution, thus the model can be assumed to fulfil the normality assumption.

```
***********************************************************************
*          Results of Geographically Weighted Regression              *
***********************************************************************

********************Model calibration information********************
Kernel function: gaussian
Fixed bandwidth: 5790.332
Regression points: the same locations as observations are used.
Distance metric: Euclidean distance metric is used.

***************Summary of GWR coefficient estimates:******************
                     Min.   1st Qu.    Median   3rd Qu.    Max.
Intercept        237.1950  462.1771  837.8348 1569.9620 3075.04
PROX_ATTRACTION   -8.2455  273.3619  330.1201  383.5645  477.74
PROX_RESTAURANT  -94.2505  204.5687  367.4870  515.5377 1419.73
PROX_MALL       -176.9493  137.6075  209.4173  280.6602  367.48
PROX_HEALTHCARE -897.1231 -318.3257 -125.7860  -19.0274  126.31
PROX_RAILWAYS    -46.4377   14.6822   76.7124  133.0701  190.03
**********************Diagnostic information************************
Number of data points: 261
Effective number of parameters (2trace(S) - trace(S'S)): 37.26864
Effective degrees of freedom (n-2trace(S) + trace(S'S)): 223.7314
AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 4392.521
AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 4354.728
BIC (GWR book, Fotheringham, et al. 2002,GWR p. 61, eq. 2.34): 4222.305
Residual sum of squares: 241770147
R-square value:  0.4965624
Adjusted R-square value:  0.4123244

***********************************************************************
Program stops at: 2021-11-18 18:13:44
```

Figure 4.8: R-squared value of the Jakarta's base GWR model

Figure 4.8 shows the GWR's Model performance. The model has an adjusted R-square value of 0.41 which means that the model is able to explain 41% of the number of positive cases in Jakarta. Even though the current model is only able to explain 41% of the dependent variable, it is still accurate because only significant independent variables were chosen. Other independent variables and other GWR configurations can be considered to calibrate a model with a higher adjusted R-square. The formula of the model is:

POSITIF = 237.195 - 8.2455 (PROX_ATTRACTION) - 94.2505 (PROX_RESTAURANT) - 176.9493 (PROX_MALL) - 897.1231 (PROX_HEALTHCARE) - 46.4377 (PROX_RAILWAYS)

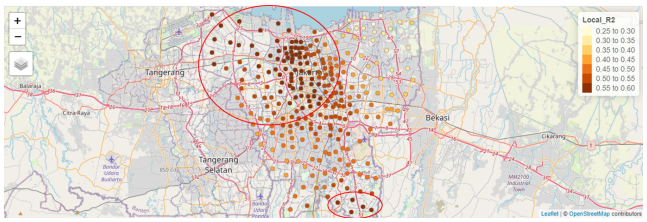

Figure 4.9: Map distribution of local R2 in Jakarta

Figure 4.9 shows that North and South-East regions of Jakarta have a higher Local_R2 value as compared to other regions. This indicates that the calibrated model better explains the number of positive cases in these regions, which also means that the independent variables, such as the proximity to attractions, restaurants, malls, healthcare facilities and railways, have a strong relation to the number of positive cases.

```
*********************************************************************
*     Results of Geographically Weighted Regression for prediction  *
*********************************************************************

********************Model calibration information********************
Kernel function: gaussian
Fixed bandwidth: 5790.332
Distance metric: Euclidean distance metric is used.

***************Summary of GWR coefficient estimates:*****************
                          Min.    1st Qu.     Median    3rd Qu.    Max.
Intercept_coef         222.9880   458.6119   829.3407  1536.0697 3027.39
PROX_ATTRACTION_coef     4.8857   266.6092   331.6605   382.7814  473.28
PROX_RESTAURANT_coef  -166.7706   197.8838   379.5288   511.7744 1494.88
PROX_MALL_coef        -158.8877   146.6247   209.4273   281.2723  388.18
PROX_HEALTHCARE_coef -1101.1131  -325.2831  -126.4679   -18.8950  127.70
PROX_RAILWAYS_coef     -58.1249    15.0024    75.9123   132.0846  231.29

***************       Results of GW prediction        ******************
                     Min.   1st Qu.    Median   3rd Qu.      Max.
prediction        -298.5    2027.4    2609.0    3298.8    5532.4
prediction_var 1089807.0 1106963.9 1123052.7 1178036.4 3443111.3

*********************************************************************
Program stops at: 2021-11-18 18:37:42
```

Figure 4.10: Predicted positive cases base on selected variables

Figure 4.10 shows the GWR prediction with the same variables as the base model. It predicts that there are a minimum of zero positive cases as the min is below zero. The predicted maximum number of positive cases is 5532.4. The median predicted number of positive cases is 2609.

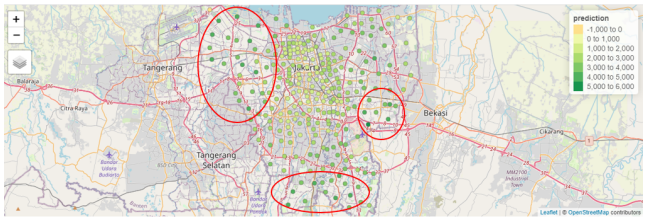

Figure 4.11: Map distribution of Jakarta's predicted postive cases

Figure 4.11 plots predicted the number of positive cases in Jakarta based on the selected independent variables. Generally, East, South and North-West of Jakarta have higher predicted positive cases as represented by the darker green spatial points. Closer observation can be done on these regions by the governments in order to minimise the spread of positive COVID cases.

5. DISCUSSION

The audience will be able to learn which independent variables play a more significant role in the spread of COVID-19 in Jakarta. The system allows users to constantly identify variables that are significant and re-calibrate the model iteratively. With the prediction model, it allow users to compare the predicted value with the actual value from EDA to see the accuracy of the model based on selected independent variables.

Overall, significant independent variables identified are proximity to attractions, restaurants, malls, healthcare facilities and railways. The prediction model with the independent variables mentioned shows that the South, East and North-west region of Jakarta have higher predicted positive cases. Closer observation can be done by the government, and if needed, engage the respective stakeholders in order to minimise the spread of the virus.

6. FUTURE WORK

Firstly, our current application can be considered as comprehensive for users that are using it for research purposes and are aware of the terms used within the analysis and the models. We acknowledge that this could result in a bigger gap in understanding for users that are not as research-focused and use the application mainly for data visualization purposes. Thus, one possible extension beyond our current application's capabilities would be to separate the tabs into research usage and visualization usage, allowing different users to fulfil their needs more efficiently.

Secondly, while our current application allows users to import their own dataset and indicate the respective projection system, we can possibly look into allowing users to import their geospatial layers as well, such that the map visualization can be more zoomed in to the areas of interest.

7. REFERENCES

- A.I.Middya, S.Roy (2021, April 12). Geographically varying relationships of COVID-19 mortality with different factors in India. Retrieved October 9, 2021, from https://www.nature.com/articles/s41598-021-86987-5
- R.C.Urban, L.Y.K.Nakada (2020). GIS-based spatial modelling of COVID-19 death incidence in São Paulo, Brazil. Retrieved October 9, 2021, from https://journals.sagepub.com/doi/pdf/10.1177/0956247820963962
- Sood, A. S. (2021, July 14). Indonesia Covid-19: Almost half of Jakarta's population may have caught the virus, survey finds. Retrieved October 9, 2021, from https://edition.cnn.com/2021/07/13/asia/indonesia-antibody-s index.html
- Worldometers. (n.d.). COVID Live Update: 237,632,869 Cases and 4,851,284 Deaths from the Coronavirus - Worldometer. Retrieved October 8, 2021, from https://www.worldometers.info/coronavirus/#countries