Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

Mapping Diarrhea in Districts of Bandung, Indonesia using Geographically Weighted Quantile Regression

Wara Alfa Syukrilla

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :
Prof. dr. Anneleen VERHASSELT

SUPERVISOR :
Prof.Dr. Yudhie ANDRIYANA

2019
2020

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics

Master's thesis

Mapping Diarrhea in Districts of Bandung, Indonesia using Geographically Weighted Quantile Regression

Wara Alfa Syukrilla

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Anneleen VERHASSELT

SUPERVISOR :

Prof.Dr. Yudhie ANDRIYANA

Acknowledgment

All the praises and thanks be to Allah for granting me the opportunity to continue my study to the master program with scholarship. With His power and merciful, I able to complete the master study and accomplish this master thesis.

I would like to thank my internal supervisor, Prof. Anneleen Verhasselt. She has been generously shared her time for frequent, meaningful meetings and explains in detail when I have confusion.

I would also thank my external supervisor, Yudhie Andriyana, Ph.D., from Padjajaran University in Indonesia. He always finds time to respond to my questions through Skype calls or Whatsapp chats. Thank you for your guidance, support, and advice until the completion of this thesis.

I express my gratitude to the Flemish Interuniversity Council (VLIR) for granting me VLIR-UOS scholarship. The scholarship gives me the privilege to study from expert professors and enable to focus on my study. To all lecturers during my study in the master program at Hasselt University, the knowledge you have shared is precious to me.

I also grateful and thankful to my parents and family that always be there to support me and pray for me day and night. Thanks also to my statistics classmates and Indonesian friends for the helpful discussion during the thesis making process.

Wara

Diepenbeek, June 2020

Abstract

Background: Diarrhea is the second most killer of children under five years old in the world. One of the top 15 countries with the highest contribution to the world's children death due to diarrhea is Indonesia. In Indonesia, the West Java province is the province with the greatest number of diarrhea cases, with Bandung (the capital of the province) is one of the top 5 cities with the most diarrhea cases in this province. To reduce the diarrhea occurrences in Bandung, information on diarrhea risk and its associated covariates is needed. There are several factors that cause high diarrhea incidences. Some of them are lack of access to drinking-safe water sources, poor sanitation, and low practice of hygienic habit. Diarrhea infection is contagious and if not controlled properly it can spread to the other geographical areas. In practice, the heterogeneity of demography in study areas leads to a non-stationary diarrhea risk across locations. Thus modeling using a global spatial model might not fully explain the real variation and a spatial model that varies across locations should be considered. Also, to direct the diarrhea control action to be right on target, we are interested in exploring the full distribution of toddlers' diarrhea and spotting the high-risk and low-risk districts in Bandung using quantile regression.

Objective: To explore the effects of three predictors (the percentage of clean water usage, hand washing habit, and healthy toilet ownership) on the toddlers' diarrhea risk at different districts of Bandung and at several quantiles of the diarrhea risk.

Methodology: The so-called Geographically Weighted Quantile Regression (GWQR) is used to explore the relationship between the diarrhea risk and the covariates both at five quantiles of the response distribution and at every district. Adaptive bisquare kernel is used to determine the spatial weight.

Results: The results show that covariates importance differs between districts and between quantiles of the diarrhea risk distribution. At the 0.25th, 0.50th, 0.75th quantiles, all covariates significantly affect the diarrhea risk, but none of the covariates is found to have an effect on the diarrhea risk at the tails of the response distribution (0.05th and 0.95th quantiles). Although all covariates are significant at three quantiles but their importance differs between districts, some covariates are important at some districts but not important at the other districts.

Keywords: Quantile regression, geographically weighted regression, diarrhea, spatial analysis.

Contents

List of Figures

List of Tables

1 Introduction

Diarrhea is a condition when the faeces is frequently excreted from bowel and not in a solid form. Diarrhea remains a leading cause of mortality for children under age 5 years old (U5) in the world (UNICEF, 2019). It highly affects children living in low- and middle-income countries including Indonesia. Although diarrhea is a treatable and preventable medical problem, diarrhea often leads to death for children under five due to dehydration and late first aid from parents. According to IVAC (2018), Indonesia is one of the top 15 countries with the highest contribution to the world's children death due to pneumonia and diarrhea. The severity of the diarrhea threat on children encouraged UNICEF and WHO to launch Global Action Plan for the Prevention of Pneumonia and Diarrhea (GAPPD) in 2009 (IVAC, 2018). With this action plan, WHO and UNICEF aim to reduce and prevent childhood death from diarrhea and pneumonia effectively, especially in those identified 15 high burden countries.

According to WHO (2011), some of the risk factors that cause the high prevalence of diarrhea are the unhygienic environment and unhealthy living behavior, such as the lack of access to drinking-safe water sources, poor sanitation, the low practice of hand washing habit, and open defecation. Open defecation refers to defecation in the open place or the existence of toilets without septic tank. It makes people vulnerable to worms and bacteria that are transmitted through the soil and causes the underground-water to be contaminated by the feces bacteria. Around 8 million out of 250 million Indonesian practice open defecation hence Indonesia suffers from high diarrhea cases (Indonesia Ministry of Health, 2019). Among the cities with high percentage of open defecation practice, Bandung is one of them. Bandung is the capital of West Java province in Indonesia, in which this province has the highest diarrhea infection in Indonesia (Indonesia Ministry of Health, 2017). Out of all cities in West Java province, Bandung is one of the top 5 cities with the highest diarrhea incidence (West Java Department of Health, 2017). Therefore Bandung needs serious control and treatment to reduce diarrhea occurrence. Besides the mentioned risk factors, there are additional risk factors for diarrhea in children under 5, such as insufficient nutrition due to not optimal breastfeeding process, low economic prosperity, and allergy (Black et al., 2019).

Diarrhea occurrence is infectious. It can spread to other houses and neighbouring districts especially where there are many people in small regions and many shared public facilities with low hygiene (CDC, 2019). This implies that when diarrhea infection happens in one

area, the other close areas will be at higher risk to be infected too, meaning that spatial effect needs to be taken into account. If the spatial effect is not taken into account, it will lead to bias and inefficiency in the estimates (Stakhovych et al., 2012). A commonly used spatial approach is to assume that the effect of predictors on the response is the same everywhere within the study region, which is called spatial stationarity (Fotheringham et al., 2003). This standard spatial approach will produce a single regression model with one global coefficient across geographical study areas after accounting for spatial dependency. However, this standard spatial approach is not always sufficient to describe the real variation in reality. The effect of some predictors on the response might be different between geographical locations. It might be happen that one predictor is essential at one location but not at the other location. When the configuration within the data cannot be explained using a single "global" spatial model, a varying coefficient spatial model called Geographically Weighted Regression (GWR) model can be used (Brunsdon et al., 1996). GWR allows the relationship of a set of variables to be different at different areas by producing a regression model with varying coefficients over space. This assumption is termed as spatial nonstationarity. Using GWR enables one to investigate deeper in which location certain predictors have a bigger/smaller effect on the response.

Applying GWR in the diarrhea data implies that we will obtain the estimates of predictors' effects on the mean function of the diarrhea risk. However, it is also of interest to see the predictors' effects on the high-risk (upper tail) and low-risk (lower tail) areas of diarrhea risk distribution and not only on the mean area of the risk distribution. For this aim, quantile regression which was introduced by Koenker and Bassett Jr (1978) is used in the analysis.

Quantile regression is an extension of classical regression technique in which it estimates the conditional quantile of the response modelled as a function of covariates. Thus in quantile regression, an increase/decrease in the predictors will increase/decrease the conditional quantile of the response (Hao and Naiman, 2007). Median regression is another name of the quantile regression when the quantile is 0.5 (Koenker, 2005). The use of quantile regression is powerful especially when there is heterogeneity in the response distribution since we can obtain an overall idea on how predictors affect the response at many parts of the response distribution by assuming a regression model on various quantiles. Moreover, quantile regression is appropriate to be used when the response distribution is skewed. When the response distribution is skewed, the regression based on mean will likely under/overestimate the ef-

fect of predictors on the tail probability of the response, but not with the quantile regression since it is less sensitive to the tail behavior of the response distribution. This characteristic of quantile regression makes quantile regression also robust to outliers (Alsayed et al., 2020).

Combining GWR and quantile regression results in the so-called Geographically Weighted Quantile Regression (GWQR) method (Chen et al., 2012). This method accounts for spatial nonstationarity as well as heterogeneity of the response distribution at the same time.

1.1 Objective

In this project, we illustrate GWQR on diarrhea risk data specifically toddlers' diarrhea risk in Bandung city, Indonesia. It is of interest to investigate the effects of three predictors (the percentage of clean water usage, hand washing habit, and healthy toilet ownership) on the toddlers' diarrhea risk at different districts of Bandung and at several quantiles of the diarrhea risk.

2 Data Description

The analysed dataset was obtained from the public health department in Bandung. The response (Y) is the diarrhea risk of children under five years old of 30 districts in Bandung in 2015. Diarrhea risk is the percentage of children under five years old who experienced diarrhea during 2015. It is calculated by dividing the number of toddlers diarrhea incidents at each district over the total toddlers in the district then multiplied by 100. In this thesis we use three predictors. These predictors are believed to be associated with diarrhea infection based on the diarrhea bulletin from Indonesia Ministry of Health (2011) and based on Millennium Development Goals (MDGs) from the United Nations which promote the so-called WASH (water, sanitation, and hygiene) campaign. Those three predictors are:

1. X_1 is the percentage of households that use clean water in each district. It is obtained by counting the households that have clean water access divided by the number of houses in a specified district then multiplied by 100.

2. X_2 is the percentage of households with hand washing habit using soap and clean water. To obtain this, houses that practice hand washing habit are recorded and divided by the number of houses in the district then multiplied by 100.

3. X_3 is the percentage of households with healthy toilet. The healthy toilet is described as toilet that satisfy the following criteria: odorless and feces cannot be touched by insects or rats, water and cleaning tools are available, equipped with protective walls and roof, easy to clean and safe to use (Indonesia Ministry of Health, 2014). For obtaining this predictor, the number of houses with healthy toilet is divided by the total houses in the district then multiplied by 100.

3 Methodology

3.1 Quantile Regression

Let Y be the response variable and $X_1, ..., X_p$ be a set of predictor variables. The response is associated to predictors through parameters $\beta_1, ..., \beta_p$. In the classical linear mean regression, we consider the following model

$$Y = \boldsymbol{X}^\intercal \boldsymbol{\beta} + \varepsilon, \tag{1}$$

with $\boldsymbol{X} = (1, X_1, ..., X_p)^\intercal$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\intercal$, and ε is the error term which is assumed $E(\varepsilon) = 0$. Suppose we have n independent observations $(X_{11}, ..., X_{p1}, Y_1), ..., (X_{1n}, ..., X_{pn}, Y_n)$ from $(X_1, ..., X_p, Y)$. The parameter estimation in the linear mean regression is obtained by minimizing the residuals sum of square $\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\beta})^2$ with $\boldsymbol{X}_i^\intercal = (1, X_{1i}, ..., X_{pi})$ and $i = 1, ..., n$.

In the quantile regression one can model the association of predictors not only with the conditional central tendency (median) of the response but also with the other conditional quantiles of the response. This makes quantile regression more beneficial than classical mean regression since it can provide full visualization on how predictors related to the response at all parts of the response distribution. When one wants to explore the response-predictors association only at the upper or lower percentile of the response, quantile regression is suitable to be used. In addition, quantile regression is more robust to outlying observations (Andriyana et al., 2014).

Quantile function is the inverse of cumulative distribution function (CDF). The CDF of Y given $\boldsymbol{X}$ can be expressed as $F_Y(y|\boldsymbol{X}) = P(Y \leq y|\boldsymbol{X})$ with the inverse function $F_Y^{-1}(\tau) = \inf\{y : F(y|\boldsymbol{X}) \geq \tau\} = Q_Y(\tau|\boldsymbol{X})$ (for $\tau \in [0, 1]$) called the τth conditional quantile of the response Y given covariate $\boldsymbol{X}$ (Koenker, 2005). The linear model for quantile regression can be written as:

$$Y = \boldsymbol{X}^\intercal \boldsymbol{\beta}^\tau + \varepsilon, \tag{2}$$

where the vector $\boldsymbol{\beta}^\tau = (\beta_0^\tau, ..., \beta_p^\tau)^\intercal$ is the regression coefficient vector at the τth quantile. ε is the error term in which we assume that the τth quantile of ε given $\boldsymbol{X}$ is equal to zero ($Q_\varepsilon(\tau|\boldsymbol{X}) = 0$) hence $Q_Y(\tau|\boldsymbol{X}) = \boldsymbol{X}^\intercal \boldsymbol{\beta}^\tau$ (Koenker, 2005). The quantile regression parameters are estimated by minimizing the following quantile loss function with respect

to $\boldsymbol{\beta}$

$$\sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\beta}), \tag{3}$$

where $\rho_\tau(.)$ is a check-function which defined as

$$\rho_\tau(e) = \begin{cases} e(\tau) & \textit{if } e > 0, \\ e(\tau - 1) & \textit{otherwise.} \end{cases} \tag{4}$$

The check function is a V-shaped piecewise linear function which can be illustrated as Figure 1.

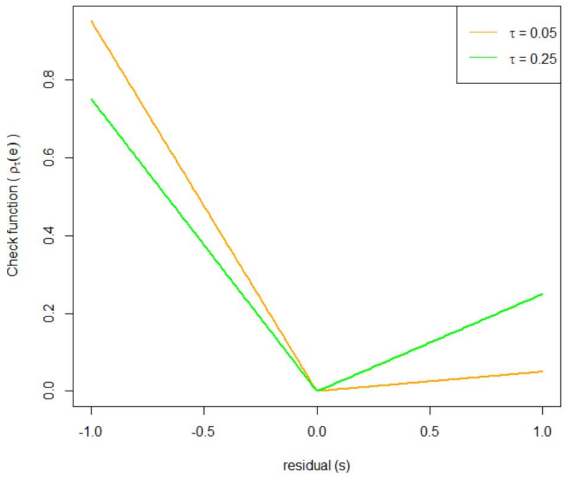

Figure 1: The illustration of check function for quantile regression $\tau = 0.05, 0.25$.

Figure 1 illustrates that the quantile objective function (3) is not differentiable. Thus, as proposed by Koenker (2005), quantile objective function could be translated into a linear programming optimization problem to find the optimal solution.

3.2 Geographically Weighted Regression (GWR)

Geographically Weighted Regression (GWR) is basically a regression model with weighted least squares estimation. GWR is an extension of global mean regression in which a spatial component is added and producing locally mean regression such that it takes into account spatial nonstationarity in the data. The underlying model for GWR is

$$Y = \beta_0(u, v) + \sum_{k=1}^{p} X_k \beta_k(u, v) + \varepsilon = \boldsymbol{X}^\intercal \boldsymbol{\beta}(u, v) + \varepsilon, \tag{5}$$

where $\boldsymbol{\beta}(u,v) = (\beta_0(u,v), ..., \beta_p(u,v))^\intercal$ represent the regression coefficients that are estimated for location with geographical coordinate (u,v), u stand for latitude and v stand for longitude. ε is a random error term that is normally distributed with mean zero and common variance σ^2. Fotheringham et al. (2003) explained that GWR works by adapting the moving window regression principle and enclosing geographical weight component. Suppose that we have a circular window with regression point (u_0, v_0) as the midpoint and there are several observations inside the circle around the midpoint (u_0, v_0). These observations inside the circle are considered near to the regression point and will get a nonzero spatial weight, while the observations outside the circle will have zero weight since they are considered distant from the midpoint (u_0, v_0). Among those observations with nonzero weight, a smaller distance from the regression point (u_0, v_0) will results in a higher weight than those with greater distance. When an observation located at the same location as the regression point, the weight will be unity. A linear mean regression model is then fitted for regression point (u_0, v_0) using all observations within the circle and the parameters are estimated with geographically weighted least squares method. This procedure is then repeated for all regression points.

The spatial weight assigned to each observation i is determined based on a kernel method. In this study, the spatial weight is expressed by $w_{i0} = K(\frac{d_{i0}}{h})$ with $K(.)$ is a kernel function. A kernel function involves distance (d_{i0}) and bandwidth parameter (h). The distance is measured using the Euclidean distance i.e. between the ith observation with coordinate (u_i, v_i) and the regression point (u_0, v_0) then $d_{i0} = ||(u_i, v_i) - (u_0, v_0)||$. The bandwidth parameter $(h \geq 0)$ is a measure of distance-decay in the weighting process as well as a controller of the smoothness of the resultant coefficients. The commonly used kernel functions in GWR are Gaussian kernel and Bi-square kernel. Details about these kernels are described in the Section 3.3.2.

3.3 Geographically Weighted Quantile Regression (GWQR)

As described in the previous section, Geographically Weighted Regression (GWR) is a locally mean regression model that can accomodate spatial non-stationarity. Since this study interest is to see how covariates affect the response at the high risk and low risk of diarrhea, quantile regression is incorporated and results in a Geographically Weighted

Quantile Regression (GWQR) model which is formulated as follows:

$$Y = \beta_0^\tau(u,v) + \sum_{k=1}^{p} X_k \beta_k^\tau(u,v) + \varepsilon = \boldsymbol{X}^\intercal \boldsymbol{\beta}^\tau(u,v) + \varepsilon, \tag{6}$$

where $\beta_k^\tau(u,v)$ is the regression coefficient at τth quantile at the location with geographical coordinate (u,v). Following the quantile regression scheme, the τth conditional quantile function of the observed response Y can be expressed as $Q_Y(\tau|\boldsymbol{X},u,v) = \boldsymbol{X}^\intercal\boldsymbol{\beta}^\tau(u,v) + Q_\varepsilon(\tau|\boldsymbol{X},u,v)$. Since it is assumed that $Q_\varepsilon(\tau|\boldsymbol{X},u,v) = 0$ hence $Q_Y(\tau|\boldsymbol{X},u,v) = \boldsymbol{X}^\intercal\boldsymbol{\beta}^\tau(u,v)$ (Koenker, 2005).

3.3.1 Estimation in GWQR

Suppose that the GWQR coefficient vector $\boldsymbol{\beta}^\tau(u,v) = [\beta_0^\tau(u,v), ..., \beta_p^\tau(u,v)]^\intercal$ has second continuous partial derivative with respect to location latitude u and longitude v. Then at a regression point (u_0,v_0), the vector $\boldsymbol{\beta}^\tau(u,v)$ can be approximated using Taylor's expansion by

$$\boldsymbol{\beta}^\tau(u,v) \approx \boldsymbol{\beta}^\tau(u_0,v_0) + \boldsymbol{\beta}^{\tau(u)}(u_0,v_0)(u-u_0) + \boldsymbol{\beta}^{\tau(v)}(u_0,v_0)(v-v_0), \tag{7}$$

for (u,v) in the neighborhood of (u_0,v_0), with $\boldsymbol{\beta}^{\tau(u)}(u_0,v_0)$ is the vector of partial derivative vector of $\boldsymbol{\beta}^\tau(u,v)$ at (u_0,v_0) with respect to u and $\boldsymbol{\beta}^{\tau(v)}(u_0,v_0)$ is the vector of partial derivative with respect to v. Based on linear approximation (7) and local modelling principles, the GWQR estimates can be obtained through minimization of the following weighted loss function

$$\sum_{i=1}^{n} \rho_\tau \left(Y_i - \boldsymbol{X}_i^\intercal \left[\boldsymbol{\beta}^\tau(u_0,v_0) + \boldsymbol{\beta}^{\tau(u)}(u_0,v_0)(u_i-u_0) + \boldsymbol{\beta}^{\tau(v)}(u_0,v_0)(v_i-v_0) \right] \right) w_{i0}, \tag{8}$$

or it can also be expressed as

$$\sum_{i=1}^{n} \rho_\tau \left(Y_i - \tilde{\boldsymbol{X}}_i^\intercal \boldsymbol{\theta}^\tau(u_0,v_0) \right) w_{i0}, \tag{9}$$

where $\tilde{\boldsymbol{X}}_i = [1, X_{1i}, ..., X_{pi}, (u_i-u_0), X_{1i}(u_i-u_0), ..., X_{pi}(u_i-u_0), (v_i-v_0), X_{1i}(v_i-v_0), ..., X_{pi}(v_i-v_0)]^\intercal$ and $\boldsymbol{\theta}^\tau(u_0,v_0) = [\beta_0^\tau(u_0,v_0), ..., \beta_p^\tau(u_0,v_0), \beta_0^{\tau(u)}(u_0,v_0), ..., \beta_p^{\tau(u)}(u_0,v_0), \beta_0^{\tau(v)}(u_0,v_0), ..., \beta_p^{\tau(v)}(u_0,v_0)]^\intercal$. Using linear programming optimization, the parameter vector $\boldsymbol{\theta}^\tau(u_0,v_0)$ has estimate $\hat{\boldsymbol{\theta}}^\tau(u_0,v_0) = [\hat{\beta}_0^\tau(u_0,v_0), ..., \hat{\beta}_p^\tau(u_0,v_0), \hat{\beta}_0^{\tau(u)}(u_0,v_0), ..., \hat{\beta}_p^{\tau(u)}(u_0,v_0), \hat{\beta}_0^{\tau(v)}(u_0,v_0), ..., \hat{\beta}_p^{\tau(v)}(u_0,v_0)]^\intercal$.

Instead of using the local linear approximation (7), one can use the local constant approximation:

$$\boldsymbol{\beta}^{\tau}(u, v) \approx \boldsymbol{\beta}^{\tau}(u_0, v_0). \tag{10}$$

Using the local constant approximation (10), the GWQR parameters can be estimated by minimizing the following weighted quantile loss function :

$$\sum_{i=1}^{n} \rho_{\tau} \left[Y_i - \boldsymbol{X}_i^{\intercal} \boldsymbol{\beta}^{\tau}(u_0, v_0) \right] w_{i0}, \tag{11}$$

where $w_{i0} = K(\frac{d_{i0}}{h})$ is the spatial weight. The vector of parameter $\boldsymbol{\beta}^{\tau}(u_0, v_0)$ is the regression coefficients at quantile τ and location coordinate (u_0, v_0) with the local constant estimator is notated by $\hat{\boldsymbol{\beta}}^{\tau}(u_0, v_0)$. The spatial weight w_{i0} does not depend on τ indicating that the bandwidth is identical at every quantile.

3.3.2 Kernel in GWQR

A Kernel function is used to determine the weight in GWQR to obtain smooth local parameter estimates. Two commonly used kernel function are Gaussian kernel and Bi-square kernel.

- Gaussian Kernel

$$w_{i0} = \exp\left(-\frac{1}{2} \left(\frac{d_{i0}}{h} \right)^2 \right),$$

- Bi-square Kernel

$$w_{i0} = \begin{cases} \left(1 - \left(\frac{d_{i0}}{h} \right)^2 \right)^2 & \text{if } d_{i0} \leq h, \\ 0 & \text{if } d_{i0} > h. \end{cases}$$

The above two kernels have fixed bandwidth in which the kernel size remains the same across geographical regions. However, some regions might be surrounded by many observations while others are sparse. When there are only very few observations at a region and regression is fitted at this region using only these few observations, then the parameter estimates will likely have large standard error (Fotheringham et al., 2003). Thus it would be sensible to assign kernel with adaptive bandwidth in which it will adjust the kernel size according to the density of observations at a region. This study uses the adaptive bi-square kernel

$$w_{i0} = \begin{cases} \left(1 - \left(\frac{d_{i0}}{h_i}\right)^2\right)^2 & \text{if } d_{i0} \leq h_i, \\ 0 & \text{if } d_{i0} > h_i. \end{cases} \tag{12}$$

According to Fotheringham et al. (2003), any choice of kernel gives relatively the same results, but different bandwidth will rise different results. Thus selecting optimal bandwidth is necessary. Several criteria can be used in selecting the optimal bandwidth: (1) Cross Validation, (2) Generalized Cross Validation, (3) Akaike Information Criterion, (4) Bayesian Information Criterion. In this paper, Leave One Out Cross Validation (CV) approach will be used with formula:

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau \left[Y_i - \hat{Q}_Y^{(-i)}(\tau|\boldsymbol{X}_i, u_i, v_i)\right], \tag{13}$$

where $\hat{Q}_Y^{-i}(\tau|\boldsymbol{X}_i, u_i, v_i)$ is the estimate of $Q_Y(\tau|\boldsymbol{X}_i, u_i, v_i)$ at location i obtained from estimation process that uses all observations except the observation at location i itself $(Y_i, \boldsymbol{X}_i, u_i, v_i)$. The bandwidth with the smallest CV score will be chosen.

3.3.3 Significance Test for GWQR parameter

To construct a hypothesis testing procedure, we need to estimate the standard error of the parameters. Recall the equation (11) can be expressed in other form

$$\sum_{i=1}^{n} \rho_\tau \left[Y_i w_{i0} - \boldsymbol{X}_i^\intercal \boldsymbol{\beta}^\tau(u_0, v_0) w_{i0}\right]. \tag{14}$$

According to Chen et al. (2012), having equation (11) and (14), and based on the asymptotic theorems described in Koenker (2005), the estimate $\hat{\boldsymbol{\theta}}^\tau(u_0, v_0)$ is asymptotically normal distributed with the mean is the true regression coefficients $\boldsymbol{\theta}^\tau(u_0, v_0)$ and the covariance matrix

$$\text{cov}\left(\hat{\boldsymbol{\theta}}^\tau(u_0, v_0)\right) = \frac{\tau(1-\tau)}{n} \boldsymbol{D}_1^{-1} \boldsymbol{D}_0 \boldsymbol{D}_1^{-1}, \tag{15}$$

where
$\boldsymbol{D}_0 = E[\tilde{\boldsymbol{X}}_i^{\,*} \tilde{\boldsymbol{X}}_i^{\,*\intercal}]$,
$\boldsymbol{D}_1 = E[\tilde{\boldsymbol{X}}_i^{\,*} \tilde{\boldsymbol{X}}_i^{\,*\intercal}\, f(Q_Y(\tau|\tilde{\boldsymbol{X}}_i^{\,*}, u_i, v_i))]$,
$\tilde{\boldsymbol{X}}_i^{\,*} = \tilde{\boldsymbol{X}}_i w_{i0}$, and
$f(Q_Y(\tau|\tilde{\boldsymbol{X}}_i^{\,*}, u_i, v_i))]$ is the conditional density of Y_i for $\tilde{\boldsymbol{X}}_i^{\,*}$ at the geographical coordinate

(u_i, v_i) estimated at the τ-th conditional quantile.

To estimate the covariance for $\hat{\boldsymbol{\theta}}^{\tau}(u_0, v_0)$, we use $\hat{\text{cov}}\left(\hat{\boldsymbol{\theta}}^{\tau}(u_0, v_0)\right) = \frac{\tau(1-\tau)}{n}\hat{\boldsymbol{D}}_1^{-1}\hat{\boldsymbol{D}}_0\hat{\boldsymbol{D}}_1^{-1}$ and we can use it for constructing the test statistic for hypothesis testing. The estimation involves
$\hat{\boldsymbol{D}}_0 = \frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_i^*\tilde{\boldsymbol{X}}_i^{*\intercal}$,
$\hat{\boldsymbol{D}}_1 = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(\hat{Q}_Y(\tau|\tilde{\boldsymbol{X}}_i^*, u_i, v_i))\tilde{\boldsymbol{X}}_i^*\tilde{\boldsymbol{X}}_i^{*\intercal}$,
$\hat{f}(\hat{Q}_Y(\tau|\tilde{\boldsymbol{X}}_i^*, u_i, v_i)) = \frac{2h_n}{\tilde{\boldsymbol{X}}_i^{*\intercal}\hat{\boldsymbol{\theta}}^{\tau+h_n}(u_0,v_0) - \tilde{\boldsymbol{X}}_i^{*\intercal}\hat{\boldsymbol{\theta}}^{\tau-h_n}(u_0,v_0)}$,
with h_n is a bandwidth parameter in which for a large sample size ($n \to \infty$) the bandwidth parameter will be close to zero. The determination of the bandwidth h_n is via the Hall-Sheather bandwidth rule suggested by Koenker (1994)

$$h_n = n^{-1/3} z_\alpha^{2/3} \left[\frac{1.5\phi^2\left(\Phi^{-1}(\tau)\right)}{\left(2\Phi^{-1}(\tau)\right)^2 + 1}\right]^{1/5}, \tag{16}$$

with Φ be a standard normal distribution function, with density function denoted by ϕ and z_α satisfying $\Phi(z_\alpha) = 1 - \frac{\alpha}{2}$. α is the significance level for hypothesis testing. It is worth to note that $\hat{f}(\hat{Q}_Y(\tau|\tilde{\boldsymbol{X}}_i^*, u_i, v_i))$ is not necessarily always positive and may cause problem, hence it is replaced by the quantity advised by Koenker and Machado (1999)

$$\hat{f}(\hat{Q}_Y(\tau|\tilde{\boldsymbol{X}}_i^*, u_i, v_i)) = \max\left[0, \frac{2h_n}{\tilde{\boldsymbol{X}}_i^{*\intercal}\hat{\boldsymbol{\theta}}^{\tau+h_n}(u_0, v_0) - \tilde{\boldsymbol{X}}_i^{*}\hat{\boldsymbol{\theta}}^{\tau-h_n}(u_0, v_0) - \eta}\right], \tag{17}$$

where η is a small positive number to avoid division by zero, in this thesis $\eta = 0.01$.

The significance of GWQR parameter estimates can be tested using partial t-test with hypothesis:
$H_0 : \beta_k^{\tau}(u_0, v_0) = 0, \quad k = 1, ..., p$
H_1: $\beta_k^{\tau}(u_0, v_0) \neq 0, \quad k = 1, ..., p$
and the test statistic:

$$t_k(u_0, v_0) = \frac{\hat{\beta}_k^{\tau}(u_0, v_0)}{\hat{\sigma}\left(\hat{\beta}_k^{\tau}(u_0, v_0)\right)},$$

where standard error of the k-th coefficient $\hat{\sigma}\left(\hat{\beta}_k^{\tau}(u_0, v_0)\right)$ is obtained from the $(k+1)$th diagonal element of the estimated covariance matrix $\hat{\text{cov}}\left(\hat{\boldsymbol{\theta}}^{\tau}(u_0, v_0)\right)$. H_0 is rejected when $t_k(u_0, v_0) > z_{\alpha/2}$ with $\alpha = 0.05$.

3.3.4 Assessment of Spatial Nonstationarity

In this thesis, we follow the approach in Chen et al. (2012) to assess the spatial nonstationarity in GWQR coefficients. At a specified quantile, we compare the Inter Quartile Range (IQR) of the GWQR coefficients with the standard error of the traditional quantile regression. When the IQR is twice larger than the standard error of the traditional quantile regression then it is said that spatial non-stationarity is present in the relationship between predictors and response.

3.4 Software

The data analysis was executed using R version 4.0.0. There is no R package available for GWQR yet. The GWQR analysis in this thesis utilized a modified version of `GWmodel`.

4 Results

In this section we presents the results and interpretation of the results from the application of Geographically Weighted Quantile Regression on toddlers' diarrhea data. We start with exploratory data analysis, then present the diarrhea mapping based on GWQR modeling results.

4.1 Exploratory Data Analysis

To get some illustration about the data, this subsection presents an overview of the spread of the diarrhea risk (the response), a check on spatial heterogeneity in the data using Breusch Pagan test, and plots of predictors against the response.

Figure 2 illustrate the spread of the response (diarrhea risk) over 30 districts in Bandung city in 2015. The diarrhea risk ranges from 5 to 32 percent with average 13.69 which means that on average 14 out of 100 of children under 5 years old in Bandung experienced diarrhea. According to this picture, the level of diarrhea risk varies across districts. 12 out of 30 districts have small diarrhea risk (risk less than 10) in which most of them located side by side in the middle part of Bandung city while some others spread at the western part. The eastern part of the map seems to have darker color, indicating that diarrhea risk is high at districts in this area. The highest diarrhea risk happens in the Panyileukan district, expressed with the darkest color at the east side of the map, amounted to 31.80. Not only at the eastern part, but at the center of the map there is also a district with dark color around 6.9°S latitude and 107.61°E longitude. This district is called Bandung Wetan and is the location of the central government building of the West Java province.

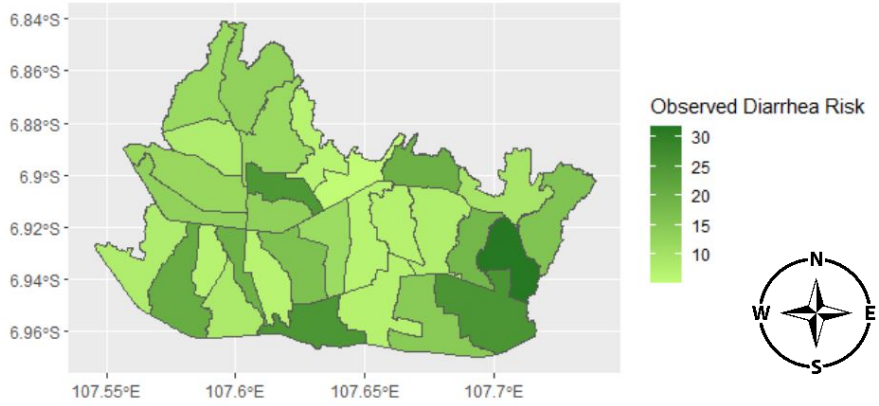

Figure 2: The map of the observed diarrhea risk

In Figure 3, a boxmap is presented to mimic a boxplot in term of geographical map visualization hence the boxmap will help to indicate if outliers are present in the response. In case outliers exist in the response, caution is needed if someone is going to perform modelling based on conditional mean. The use of quantile regression is one option to have a more robust modeling technique than mean-based model. Based on the boxmap there is no outlier in the response.

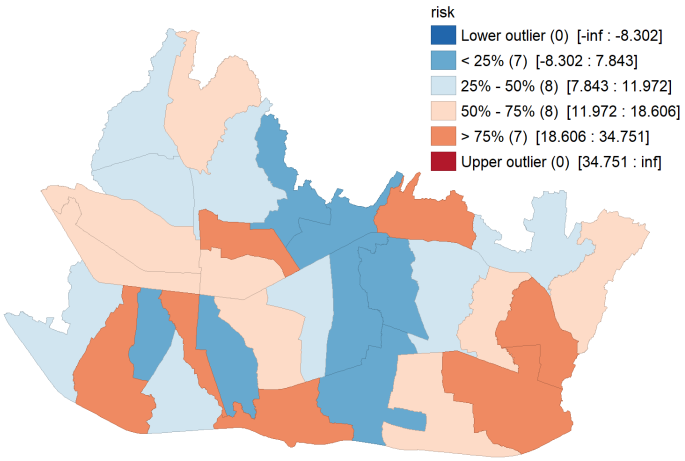

Figure 3: Boxmap of the observed diarrhea risk

The histogram of the response in Figure 4 shows that the diarrhea risk variable is asymmetrically distributed and no outlier is observed. In quantile regression there is no assumption for the distribution of the response hence quantile regression can handle this skewed data.

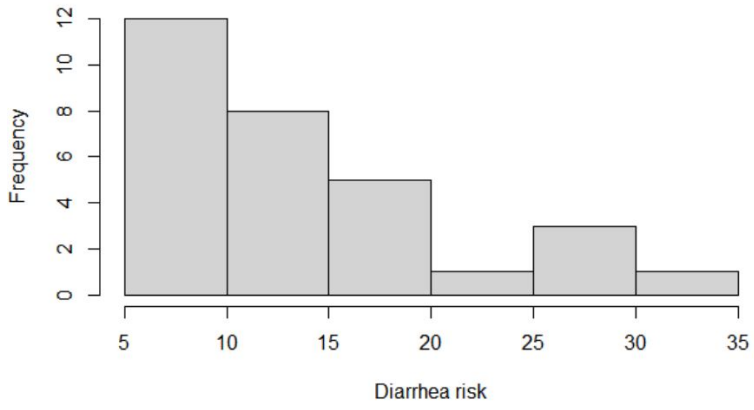

Figure 4: Histogram of the observed diarrhea risk

A test for spatial heterogeneity was performed using Breusch-Pagan test after taking spatial effect into account. The Breusch Pagan p-value is equal to 0.043 (significant at $\alpha = 0.05$) which means that spatial heterogeneity is present in the data. This is an indication that the relationship between covariates and the response seems to vary over space and the use of locally varying spatial model such as Geographically Weighted model would be logical to be used.

Figure 5 shows that there are more points at the higher values of predictors compared to the smaller values of predictors, thus the variability seems increasing as the predictors values increase. For X_1 (the percentage of houses with clean water access), there seems small variability for this predictor because almost all of the points located at the percentage 99-100% although the diarrhea risk varies from 5 to 32 percent. A similar case happens with the predictor X_2 (percentage of hand washing habit) in which the majority of the districts have hand washing habit percentage above 90% even though the diarrhea risk is spread widely from low to high. The third predictor seems to have a larger variability than the previous 2 predictors since the points spread from low to high values of the percentage of healthy toilet. Majority of the points located at the bottom right indicating that many districts have high percentage of healthy toilet and low percentage of diarrhea. Further, it is observed that there is one point which has low values for percentage of clean water, hand washing habit, and healthy toilet, but has low diarrhea risk.

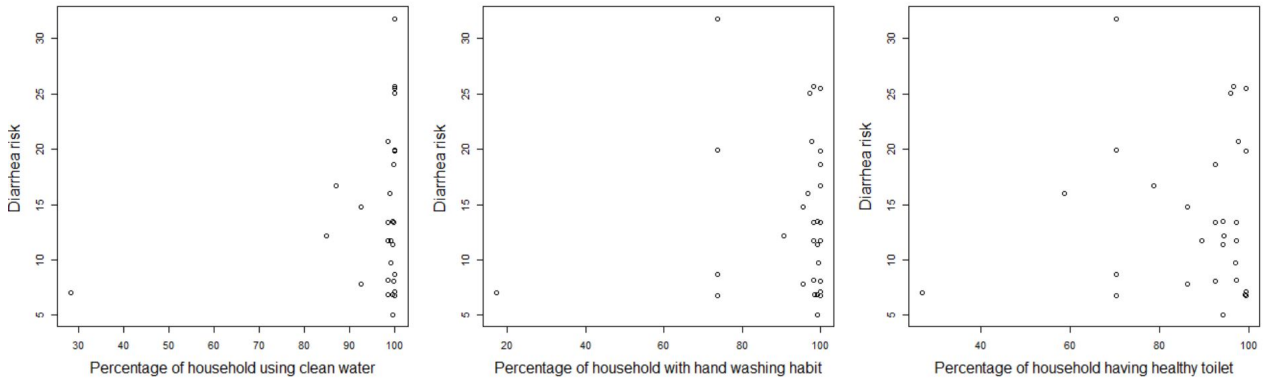

Figure 5: Scatter plot of predictors (left to the right: X_1-X_3) against the response

4.2 Geographically Weighted Quantile Regression

GWQR with bisquare kernel is used in the analysis and using adaptive bandwidth approach the optimal bandwidth obtained from cross validation is $h = 28$ for all the five quantiles.

Table 1 shows the summary of the parameter estimates of GWQR for the five chosen quantiles $\tau = 0.05, 0.25, 0.50, 0.75, 0.95$ with an additional column for the standard error of the global quantile regression coefficients. There is no theoretical cut off for diarrhea risk which classify the risk of diarrhea into high, low, or other categories. Hence in this thesis we use conventional classification for the response i.e. very low risk of diarrhea is 5th percentile, low risk of diarrhea is 25th percentile, median risk of diarrhea is 50th percentile, high risk of diarrhea is 75th percentile, and very high risk of diarrhea is 95th percentile.

For initial identification of whether spatial non-stationarity is exist in the GWQR estimates, we follow the approach from Chen et al. (2012). A predictor is said to have a non-stationary relationship with the response across space if the IQR of the GWQR parameter estimates is twice larger than the standard error of the global quantile regression coefficients. Based on Table 1, all predictors have a spatial non stationary relationship with the response, at least in one quantile. This suggests that the relationships between the toddlers' diarrhea risk and the clean and healthy living behaviors indeed vary over the districts of Bandung. The predictor X_1 is detected to have IQR of the GWQR coefficients twice larger than the standard error of the global quantile regression coefficient at 4 out of 5 quantiles ($\tau = 0.05, 0.25, 0.50, 0.95$). This indicates that the relationship between X_1 and Y is not the same across different districts and at different sections of the risk distribution. The other two predictors (X_2 and X_3) are detected to have non-stationary process across districts at the lower tail of the diarrhea risk (5th percentile).

Figure 6 shows the predictions from GWQR model based on five chosen quantiles. Starting from the medium percentile i.e. 50th percentile, the map of predictions from the GWQR $\tau = 0.50$ model shows there is one district at the eastern part of the map with the darkest color. This indicates that, at the median, this district predicts to have the highest diarrhea risk of all districts, amounted to 31.8 and identified as Panyileukan district. At this district, 50 out of 100 toddlers have a risk of diarrhea of 31.8 percent or less. Panyileukan district again predicts to have the highest diarrhea risk of all districts at the lower percentile i.e. 5th percentile. At this very low percentile where all districts predicts small values of diarrhea risk, the map from GWQR $\tau = 0.05$ model predictions shows there are noticeably 3 districts with huge values of diarrhea risk predictions, indicated by their obvious dark color. At these 3 districts, 95 out of 100 toddlers have a diarrhea risk of 16 or higher. These 3 districts, identified as Panyileukan and the other two districts, may be considered as regions with high risk of toddlers diarrhea and could be recommended to the government for a control

Table 1: Summary of GWQR estimates

Parameter	GWQR					Traditional QR	Nonstationarity
	Min	Median	Max	IQR	SE	SE	Status
$\tau = 0.05$							
Intercept	2.252	12.568	13.013	7.762	10.775	2.877	non stationary
β_1	-0.168	-0.051	0.490	0.551	0.210	0.050	non stationary
β_2	-0.242	0.416	0.437	0.503	0.291	0.058	non stationary
β_3	-0.699	-0.375	-0.153	0.226	0.283	0.049	non stationary
$\tau = 0.25$							
Intercept	2.050	9.983	21.992	3.256	8.299	8.603	stationary
β_1	-0.207	-0.094	0.810	0.590	0.162	0.148	non stationary
β_2	-0.223	0.279	0.354	0.181	0.224	0.173	stationary
β_3	-0.925	-0.163	-0.120	0.125	0.218	0.148	stationary
$\tau = 0.50$							
Intercept	-6.900	8.624	9.059	6.635	6.907	9.084	stationary
β_1	-0.233	-0.126	0.754	0.437	0.135	0.157	non stationary
β_2	-0.595	0.267	0.283	0.304	0.186	0.182	stationary
β_3	-0.516	-0.123	0.297	0.112	0.181	0.156	stationary
$\tau = 0.75$							
Intercept	-7.910	1.457	4.067	7.326	12.945	25.148	stationary
β_1	-0.251	0.191	0.632	0.579	0.253	0.434	stationary
β_2	-0.538	-0.038	0.096	0.327	0.349	0.505	stationary
β_3	0.003	0.203	0.367	0.307	0.340	0.432	stationary
$\tau = 0.95$							
Intercept	-5.126	-2.594	-2.594	2.532	16.376	11.674	stationary
β_1	0.104	0.104	0.556	0.452	0.320	0.201	non stationary
β_2	-0.322	-0.190	-0.190	0.132	0.442	0.234	stationary
β_3	0.073	0.372	0.372	0.299	0.430	0.200	stationary

and treatment regarding healthy environment and living habit.

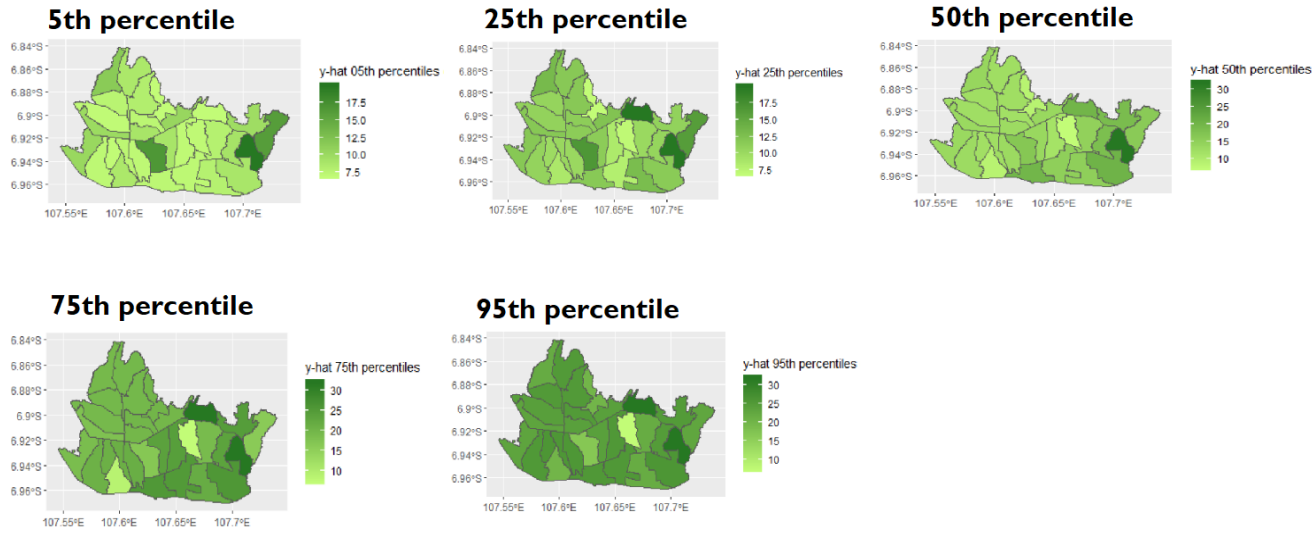

Figure 6: GWQR predictions

Looking at the very high percentile i.e. 95th percentile at Figure 6, the predictions from the GWQR $\tau = 0.95$ model shows remarkably one district with very light color. At this very high percentile where all districts predict big diarrhea risk, this district predicts low diarrhea risk. That is 95 out of 100 toddlers in this district are predicted to have 10 or less diarrhea risk. This district, called Antapani district, might be believed as the region with low risk of toddlers diarrhea and government could refer to Antapani as a role model district for improving the health quality of the other districts especially in diarrhea issue. Furthermore, Panyileukan district is consistently found as the district with the highest prediction of diarrhea risk at all quantiles hence this district might require serious attention from the health institute regarding the diarrhea issue.

The maps of the GWQR parameter estimates evaluated at different quantiles is presented at Figure 7, 8, and 9. The darker the color indicates the higher positive effect of a covariate on the response and the lighter the color indicates the smaller positive effect or the covariate effect is more negative. Locations with white color are the areas in which the effect of the predictor is not significant at 5% significance level.

Figure 7 illustrates how the percentage of houses that use clean water (X_1) is associated to the diarrhea risk. Starting from the median, we can see that based on GWQR $\tau = 0.50$

model there seems 3 groupings in the effect of the percentage of clean water towards the response, and the effect is bigger on the eastern regions rather than on the western regions. It is also noticed that moving from central/median percentile to both left and right, the coefficients are smaller especially at the tails 5th and 95th percentiles. Moreover, the effect of percentage of clean water at these tails are not significant at all districts. At the 75th percentile, there is negative effect of the percentage of clean water towards the diarrhea risk at the western part of the map. Furthermore, there is one location in which the percentage of clean water is not statistically significant affecting the diarrhea risk. This predictions show that the association of covariate and response varies across districts, that is a covariate can be important in some districts but not important at the other districts.

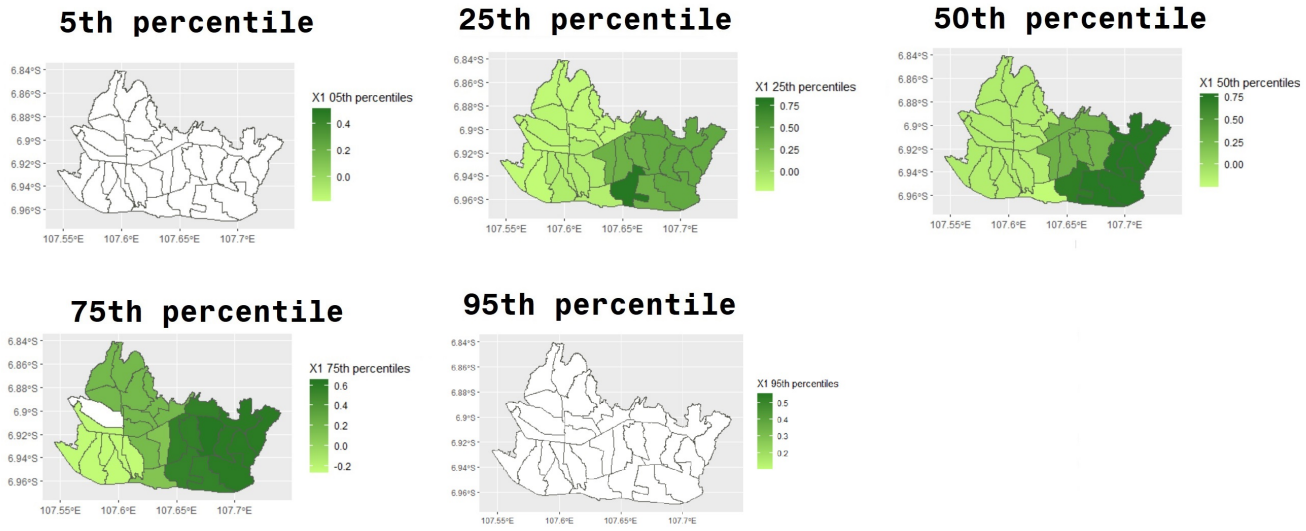

Figure 7: Maps of GWQR Estimates for X_1 (Percentage of Houses with Clean Water) significant at $\alpha = 5\%$

Figure 8 illustrates the geographical map of significant parameter estimates for the percentage of hand washing habit (X_2) evaluated at five chosen quantiles. At all quantiles, the percentage of hand washing habit is predicted to have negative relationship with the diarrhea risk at the eastern part of Bandung, and its effect is gradually increasing (more positive) as the geographical districts located at the western part. However, these positive coefficients are not statistically significant at the 75th percentile. Moving from median quantile (50th percentile) to the left (lower percentiles), the coefficients tend to be more positive, but when moving to the right the coefficients are more negative. Similar to the case in the first predictor, the percentage of hand washing habit is not proven to have an im-

portant effect on the diarrhea risk at the very low and very high percentiles ($\tau = 0.05, 0.95$).

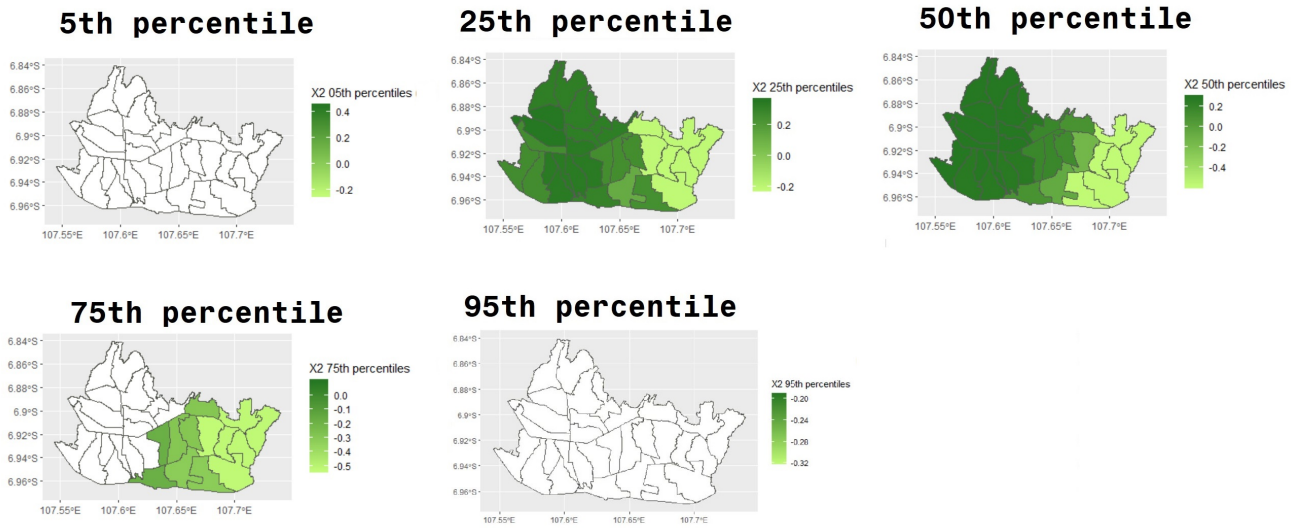

Figure 8: Maps of GWQR Estimates for X_2 (Percentage of houses hand washing habit) significant at $\alpha = 5\%$

Figure 9 shows the geographical map of GWQR coefficients for the percentage of houses with healthy toilet (X_3) that are significant at 5% significance level. At the 50th percentile, the effect of the percentage of healthy toilet is not statistically significant at one district, but significant at the other districts. This covariate has negative effects on the response at the middle part and western part of Bandung. Moving from median to the left, there is stronger negative effects of this covariate at all areas of Bandung. Based on GWQR $\tau = 0.25$ model, increasing healthy toilet percentage would decrease the diarrhea risk at all districts with the strongest effect is on the district with very light color at the southern-middle part of the map, 6.96°S latitude and 107.65°E longitude. Conversely, moving from median to the upper percentiles, the effect of this predictor is more positive. For the 75th percentile, the percentage of healthy toilet is proven to be an important predictor on the diarrhea risk at both western and eastern part of Bandung city, but not important at the middle part of Bandung. There is no effect of the percentage of healthy toilet on the response at the tails of the response distribution i.e. 5th and 95th percentiles.

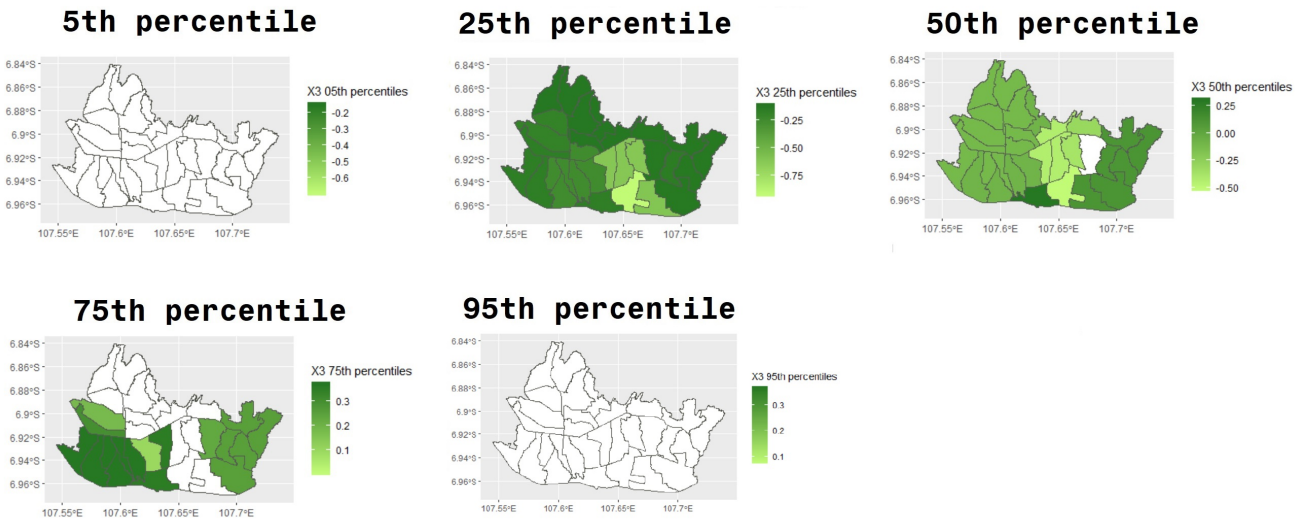

Figure 9: Maps of GWQR Estimates for X_3 (Percentage of Houses with Healthy Toilet) significant at $\alpha = 5\%$

5 Discussion

The Geographically Weighted Quantile Regression (GWQR) model is employed to answer the research questions in the study: how is the relationship between the predictors and the diarrhea risk for children under 5 years old at various spatial location, evaluated at the lower, middle, and upper quantiles of the diarrhea risk distribution.

The presented case study is about diarrhea risk for children under five in Bandung city, Indonesia. As earlier mentioned, some of the factors that causing diarrhea are unhygienic sanitation, bacteria in the water access system, bacteria in soil, unhygienic lifestyle, and lack of some nutrition. In Bandung, not all houses subscribe to water service from the Municipal Water Corporation. Some of them rely on water from the well. For houses that subscribed to water service from the Municipal Water Corporation, the water distribution service is grouped into three sub-regions, so does the water pipeline systems. The variation in water access over districts in Bandung leads to a non-stationary risk of diarrhea. Furthermore, many districts in Bandung have not reach ODF (Open Defecation Free) status yet because some houses do not have septic tank for their toilets and the feces disposal of these houses are flowed to the river. This fact suggests that those areas with the geographical location near the river seem at higher risk of soil and underground-water contamination, and it leads to non-stationary diarrhea risk across districts. To better identify the non-stationary process in the association of the covariates and the response, as well as better recognize the locations with high and low-risk of diarrhea, the GWQR is used for the analysis.

In the context of modeling the locally varying relationship, geographically weighted regression (GWR) model is commonly used, but this model regresses the predictors to the conditional mean of the response. However, using the average measure cannot characterize the full distribution of toddlers' diarrhea risk. In this study we are interested in the tails of the response distribution, not just the central tendency. The government mission for solving toddlers' diarrhea issue is targeting more on the toddlers with high risk of diarrhea, hence modeling that focus on upper quantile of the diarrhea risk distribution seems more relevant rather than modeling on the average of the diarrhea risk. This aim can be addressed by extending GWR to enable for the computation of conditional quantiles hence allows the researcher to obtain a wider picture on predictor-response relationship at various geographical areas and at various percentiles of the response. Furthermore, GWQR has an advantage that it has no distribution assumption on the error term. Hence the results in

this thesis have no issue regarding that assumption although the response variable used in this thesis seems skewed which tends to violate the assumption if someone uses GWR model based on the mean.

Our results show that the importance of predictors differ between locations. The predictor of clean water percentage is considered as important explanatory variable at the eastern areas while the percentage of hand washing habit has more influence at the western side. Moreover, not only the location of important predictors is disparate, but also the magnitude and sign of the coefficients differ between quantiles. This implies that the influence of hygienic lifestyle is spatially varying at different diarrhea risk distribution.

The results of this thesis has shortcomings. Some coefficients at some quantiles show negative relationship between predictors and the response while some others show positive relationship. The positive coefficients seem contradictory to the intuition since an increase in the percentage of houses with clean water access, healthy toilet, and hand washing habit, we expect the diarrhea risk would decrease. Reason of this might be because there is an issue with data measurement especially with the predictor variables. As stated in the exploratory part (Figure 5), the diarrhea risk varies between districts from small to high values but the predictors have high percentage of clean water, hand washing habit, and healthy toilet at the majority of the districts regardless the level of diarrhea risk. The percentage reaches above 98% for X_1 and X_2 at almost all districts including districts with very high diarrhea risk. In addition, there seems a trend between X_3 and Y where the response increasing as X_3 increases. Furthermore, there is one district with low percentage in all predictors but has low diarrhea risk. These situations cause the model to tend to predict a positive association between covariates and the response. Moreover, according to Black et al. (2019), a nutrition deficiency due to non-optimal breastfeeding or due to low economic prosperity is also another risk factor of diarrhea in toddlers. However, the data concerning this information is not available. These rationales might be the contributing reasons of why some coefficients have inappropriate sign. A future study that investigate diarrhea for toddlers is recommended to include another risk factors of toddlers' diarrhea such as breastfeeding information and mothers' educational status (Yilgwan et al., 2012). Also, it is advised for the institution that provide the data to evaluate the measurement process when collecting the data of clean water, healthy toilet, and hand washing habit.

6 Conclusion

The application of GWQR on diarrhea risk dataset allows us to identify the geographical location of high and low risk of diarrhea, and at the same time explore the effect of predictors on the full distribution of toddlers diarrhea risk while allowing its effect to vary spatially across different districts. The results show that the effect of predictors are different at different districts. This information is valuable and would allow the government to solve the diarrhea problem right on target based on which predictor that has substantial effect at a specific district of interest. Based on the estimation of GWQR, the district that is at high risk for toddlers' diarrhea is the district located at the eastern side of Bandung called Panyileukan district, while the district that is at low risk for diarrhea is the district at the middle part of Bandung called Antapani district. This knowledge can be used by the government to put more attention on Panyileukan district regarding treatments to fix diarrhea issue, and refer to Antapani district as a good example for diarrhea issue management. Furthermore, the predictors have relatively different effect for different quantiles. For example, all predictor variables are found to be important at three quantiles (0.25, 0.50, 0.75 quantiles) but none of the predictors are statistically significant affecting the response at the tails of the response distribution, 0.05th and 0.95th quantiles. At three percentiles where the predictors are significant, the location of its importance is differ between percentiles. For example, at the 75th percentile, X_3 is important at the eastern and western part of Bandung but not important at the middle part of Bandung.

References

Alsayed, A. R., Isa, Z., Kun, S. S., and Manzi, G. (2020). Quantile regression to tackle the heterogeneity on the relationship between economic growth, energy consumption, and co 2 emissions. *Environmental Modeling & Assessment*, 25(2):251–258.

Andriyana, Y., Gijbels, I., and Verhasselt, A. (2014). P-splines quantile regression estimation in varying coefficient models. *Test*, 23(1):153–194.

Black, R., Fontaine, O., Lamberti, L., Bhan, M., Huicho, L., El Arifeen, S., Masanja, H., Walker, C. F., Mengestu, T. K., Pearson, L., et al. (2019). Drivers of the reduction in childhood diarrhea mortality 1980-2015 and interventions to eliminate preventable diarrhea deaths by 2030. *Journal of Global Health*, 9(2):020801.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298.

CDC (2019). Controlling the spread of infections in evacuation centers: Facts for residents about diseases that cause diarrhea and/or vomiting. `https://www.cdc.gov/disasters/disease/infectevac.html`.

Chen, V. Y.-J., Deng, W.-S., Yang, T.-C., and Matthews, S. A. (2012). Geographically weighted quantile regression: An application to us mortality data. *Geographical Analysis*, 44(2):134–150.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, Chichester.

Hao, L. and Naiman, D. Q. (2007). *Quantile regression*. Number 149. Sage, California.

Indonesia Ministry of Health (2011). Diarrhea situation in indonesia. *Health Data and Information Window Bulletin*, 2(2).

Indonesia Ministry of Health (2014). Regulation of the minister of health of the republic of indonesia number 3/2014 on community-led total sanitation. `https://peraturan.bkpm.go.id/jdih/userfiles/batang/Permenkes_3_2014.pdf`.

Indonesia Ministry of Health (2017). Profil kesehatan indonesia 2017. `https://pusdatin.kemkes.go.id/resources/download/pusdatin/profil-kesehatan-indonesia/Data-dan-Informasi_Profil-Kesehatan-Indonesia-2017.pdf`.

Indonesia Ministry of Health (2019). Monitoring and evaluation of community-led total sanitation. `http://monev.stbm.kemkes.go.id/`.

IVAC (2018). Pneumonia and diarrhea progress report 2018. `https://stoppneumonia.org/wp-content/uploads/2018/10/JHSPH_PDPR_2018_Final_small.pdf`.

Koenker, R. (1994). Confidence intervals for regression quantiles. In *Asymptotic Statistics*, pages 349–359. Springer-Verlag Berlin Heidelberg.

Koenker, R. (2005). *Quantile Regression (Econometric Society Monographs)*. Cambridge University Press, Cambridge.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310.

Stakhovych, S., Bijmolt, T. H., and Wedel, M. (2012). Spatial dependence and heterogeneity in bayesian factor analysis: A cross-national investigation of schwartz values. *Multivariate Behavioral Research*, 47(6):803–839.

UNICEF (2019). Global annual report 2018: Goal area 1 every child survives and thrives. `https://www.unicef.org/reports/global-annual-results-2018`.

West Java Department of Health (2017). Profil kesehatan jawa barat 2017. `http://diskes.jabarprov.go.id/dmdocuments/01b3018430a412a520e2b4a4b9d9864f.pdf`.

WHO (2011). Water, sanitation and hygiene interventions and the prevention of diarrhoea. `https://www.who.int/elena/titles/bbc/wsh_diarrhoea/en/`.

Yilgwan, C. S., Okolo, S., et al. (2012). Prevalence of diarrhea disease and risk factors in jos university teaching hospital, nigeria. *Annals of African Medicine*, 11(4):217.

Appendix - R Codes (Selected Exploratory and Results)

```
library(lmtest,car,quantreg,rgdal,spdep,spatialreg,np,raster,tidyr)
library(ape,rdist,spgwr,ggplot2,GWmodel,mvoutlier,moments,dplyr)

#Exploratory Data Analysis
#Plot X vs Y,histogram of Y
plot(x=data$X1m,y=data$risk,xlab="Percentage of household using clean water",
    ylab="Diarrhea risk ")
hist(data$risk,xlab="diarrhea risk")

#Longitude latitude
BANDUNG <-readShapePoly("Bdg.shp")
coords<-coordinates(BANDUNG)

#map of y_observed
shp.risk15=st_read("risk15.shp")
gabung.risk2015=left_join(shp.risk15,data,by="POLY_ID")
plot.yobs = ggplot(data=gabung.risk2015) +
  geom_sf(aes(fill = risk)) +
  scale_fill_gradient("Observed Diarrhea Risk",low = "#c2fc77", high = "#267822")
plot.yobs

#check spatial heterogeneity Breusch Pagan
W = as.matrix(1/dist(coords))
Listw= spdep::mat2listw(W)
error.risk<-spatialreg::errorsarlm(risk~X1m+X2m+X3m,data=shp.risk15,listw = Listw)
summary(error.risk)
spatialreg::bptest.sarlm(error.risk)

#prepare GWQR
attach(data)
dspat = data.frame(risk,X1m,X2m,X3m)
x = cbind(X1m,X2m,X3m)
y = cbind(risk)
```

```
z = coords
z = as.matrix(z)
m1 = matrix(1,nrow(data),1)
datanew = SpatialPointsDataFrame(z,dspat)
dist.mat<-gw.dist(dp.locat=z, focus=0, p=2, theta=0, longlat=F)

#GWQR ANALYSIS
#Compute Bandwidth Simultant
inputtau = c(0.05,0.25,0.50,0.75,0.95)
bandwidth = c()
for(i in 1:length(inputtau)){
  bandwidth[i] = bwrq.gwr(y~x,data=datanew,approach="CV",kernel="bisquare"
                          ,adaptive=T,p=2,longlat=FALSE,ntau=inputtau[i],method="fnb")
                          $Bandwidth
  if(i==length(inputtau))  {
    bandwidth=data.frame(t(bandwidth))
    colnames(bandwidth)=inputtau
  }
}
bandwidth #Check bandwidth for each Tau/quantile
#Compute RQGWR
inputtau = c(0.05,0.25,0.50,0.75,0.95)
rqgwrresult = list()
for(i in 1:length(inputtau)){
  rqgwrresult[[i]] =
    rqgwr.basic(y~x,data=datanew,bw=bandwidth[,i],kernel="bisquare",adaptive=T,p=2,
                theta=0, longlat=F,dMat=dist.mat,F123.test=T,cv=T, W.vect=NULL
                ,ntau=inputtau[i],method="fnb")
}
AddSignificance = function(fit,alpha=0.05){
  fitsdf1 = data.frame(fit$SDF)
  fitsdf = fitsdf1[,-((ncol(fitsdf1)-3):(ncol(fitsdf1)))]
  fitlm = fit$lm
  numvariable = NROW(fitlm$coefficients)
  numdata = nrow(fitlm$x)
  sig = c()
```

```
  sigsdf = list()
  for(i in 1:numvariable){
    for(j in 1:numdata){
      sig[j] = ifelse(abs(fitsdf[j,(ncol(fitsdf)-numvariable+i)])<=qnorm(1-(alpha/2)),
                      "Not Significant","Significant")
       if(j==numdata){
        move = data.frame(sig)
        colnames(move) = gettextf("X%d_Sig",i,domain=NA)
        sigsdf[[i]] = move
      }
    }}
  fitlast = cbind(fitsdf1,data.frame(sigsdf))
  return(fitlast)
}

#RESULTS
#Tau = 0.05
result1.bisq = AddSignificance(rqgwrresult[[1]],alpha=0.05)
qr0.05=rq(risk~X1m+X2m+X3m,data=data,tau=0.05)
summary.rq(qr0.05,se="iid")$coef

#Tau = 0.25
result2.bisq = AddSignificance(rqgwrresult[[2]],alpha=0.05)
qr0.25=rq(risk~X1m+X2m+X3m,data=data,tau=0.25)
summary.rq(qr0.25,se="iid")$coef

#Tau = 0.50
result3.bisq = AddSignificance(rqgwrresult[[3]],alpha=0.05)
qr0.50=rq(risk~X1m+X2m+X3m,data=data,tau=0.50)
summary.rq(qr0.50,se="iid")$coef
```