UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Quantile Spatial Durbin Model

Neilshan Loedy
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :
Prof. dr. Anneleen VERHASSELT

SUPERVISOR :
Prof.Dr. Yudhie ANDRIYANA

2020
2021

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

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Quantile Spatial Durbin Model

Neilshan Loedy

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :

Prof. dr. Anneleen VERHASSELT

SUPERVISOR :

Prof.Dr. Yudhie ANDRIYANA

Acknowledgement

First and foremost, praises and thanks to the God, the Almighty, for His showers of blessings in allowing me to successfully complete this master thesis.

I would like to thank my internal supervisor, Professor dr. Anneleen Verhasselt from Master of Statistics program at Hasselt University for her direction, patience, and attention throughout the entire working process of this master thesis. Her attentiveness, genuineness, and generosity have allowed me to have a number of fruitful encounters and discussions. Her guidance has helped me finish my thesis on time during this difficult period of the COVID-19 pandemic. My gratitude also to my external supervisor, Dr. Yudhie Andriyana from Department of Statistics, Universitas Padjadjaran, Bandung, Indonesia, for his constant guidance and support. Working with him has been a fantastic experience.

I would like to express my gratitude to Vlaamse Interuniversitaire Raad (VLIR) for awarding me VLIRUOS scholarship. The scholarship allows me to study with expert professors and expand my knowledge by applying the theories into practice. Having the opportunity to study abroad and earn a Master's degree has been the greatest achievement of my life.

Last but not least, I would to give a special thanks to my beloved family for the love, prayers, supports and sacrifices. Also, to all my friends both in Indonesia and Belgium for the great positivity and support. Thank you, thank you, thank you!

Loedy Neilshan

Diepenbeek, June 2021

Abstract

Dengue fever is one of the most common diseases in Southeast Asia, with Indonesia having one of the highest dengue case counts in the region. In Java, notably Central Java, the number of dengue fever cases is considered high. The spread of dengue fever throughout districts may affect one another because the mosquitoes could fly between areas, and the virus can be transported via infected person mobility. Because mosquitoes can fly between locations and the virus can be transferred by infected people moving around, the spread of dengue fever between districts may impact one another, indicating that spatial effect needs to be considered. Furthermore, the estimated dengue fever incidence rate in Central Java greatly varies. We develop a Quantile Spatial Durbin Model (QSDM) method to explore the risk of dengue disease at multiple quantile levels, considering the spatially lagged response and explanatory variables to analyze the incidence rate of dengue fever across Central Java. We investigate whether the incidence rate of dengue fever for different quantiles in one area is related to the characteristics of neighboring areas, and if so, how they are related. Population density per km^2, percentages of households with adequate sanitation, percentages of households with well-drinking sources, percentages of low-income population, number of floods, and number of submerged households during floods are the six predictors that are considered. When considering the total effect, the dengue fever incidence rate at the lower, middle, and upper quantiles is always inversely proportional to population density and low-income percentages. In contrast, the number of flood-affected households has a positive effect on dengue incidence rate. According to the findings, the eastern part of Central Java has a higher dengue fever incidence rate. Regions with a low population density, stable economy, vital water infrastructure, and weak flood management are more likely to increase the dengue fever incidence rate.

Keywords: Spatial durbin model, quantile regression, dengue fever, Indonesia.

Contents

1 Introduction

Dengue fever is considered a critical public health problem, especially in tropical and subtropical countries, which ranks second only to malaria among deadly mosquito-borne diseases. Dengue disease originates from a virus in the Flaviviridae family with four known serotypes, namely DENV-1, DENV-2, DENV-3, and DENV-4, transmitted by the Aedes mosquitoes (i.e., aegypti and albopictus) infected with the dengue virus (Beasley and Barret, 2008). Even though a vaccine to prevent dengue fever has been discovered, it only has an efficacy level of 60.3% and considered costly, which makes dengue fever still classified as a dangerous disease (WHO, 2016). Roughly 2.5 billion people live in dengue fever endemic countries, with about 1.8 billion (more than 70%) in Southeast Asia and the Western Pacific Region (WHO, 2009), which carries nearly 75% of the current global disease burden due to dengue. Indonesia has one of the highest dengue case burdens in Southeast Asia, with an estimated 10 million clinical cases and 3000 deaths each year (Stanaway *et al*, 2016). The number of dengue fever cases is widely spread in East Java, West Java, and Central Java (Haryanto, 2018).

The dengue virus transmission cycle can be summarized as follows: Firstly, the female Aedes mosquito that brings the dengue virus bites a healthy individual to become infected. This individual will get dengue fever. The transmission to other humans can occur when another female Aedes mosquito, which does not carry the virus, bites the infected individual, catch the virus, and bites another healthy individual, so it becomes a cycle of transmission.

According to Whitehorn and Simmons (2011), Aedes inhabits domestic settings, particularly in large residential areas with high population density. Dengue viruses have fully adapted to a human- Aedes aegypti- human transmission cycle in the tropics' big metropolitan centers, where dense human populations coexist with equally dense mosquito populations (Gubler, 2011). However, Thammapalo *et al* (2008) stated that it

might also be present in rural areas. Aedes lays their eggs in artificial, uncovered water containers around human habitations. Surendran *et al* (2012) suggests that frequently used water domestic wells are suitable habitats for development of Aedes in tropical coasts area. Abandoned litter, gutters, and receptacles in open spaces, also serve as mosquito breeding sites. Therefore, water storage vessels can be associated with an increased number of mosquitoes and dengue fever cases (Chen *et al*, 1994). Thus, the risk of disease is higher in a poorly sanitized environment. This condition is positively correlated with low-income populations due to inconsistent or non-existent waste collection (Weiss and McMichael, 2004). Bich *et al* (2011) and Jahan (2011) asserted that the number of dengue fever is also associated with flooding, a common natural disaster in Indonesia as a country with high rainfall intensity events throughout the year.

Spread of the dengue fever between neighboring districts may affect one another because the mosquitoes could fly between areas (although the flight range of Aedes is minimal (50–100 m) (WHO, 1997)), it can be transported from one location to another, or from infected persons' mobility. These transmission factors suggested that when the incidence rate of dengue fever is high in one area, the other close regions will also have a higher incidence rate, meaning that spatial effect needs to be considered. Whereas ignoring the spatial effect can lead to biased estimates (Kostov, 2009).

The expected incidence rate of dengue fever in Central Java is equal to 12 out of 100,000 individuals. The difference between the expected incidence rate in Brebes (the region with the lowest expected incidence rate of dengue fever; 2 for every 100,000 individuals) and Blora (the region with the highest expected incidence rate of dengue fever; 45 for every 100,000 individuals) has such a big gap. This discrepancy clearly shows that in Central Java, the expected incidence rate is severely varied. To identify the factors that determine the differences in the expected incidence rates, statistical procedures that quantify the spatial effects and the heterogeneity of errors should be applied. Using the classical statistical methods that mainly focus on mean values (like the spatial

models based on classical regression) can lead to misleading results on the dengue risk determinants (Trzpiot and Orwat-Acedańska, 2016). To tackle this problem, quantile regression introduced by Koenker and Bassett (1978) is used in the analysis. The spatial model with quantile regression is powerful, especially when there is heterogeneity in the response distribution. Using this method, an overall idea of how predictors affect the response at many parts of the response distribution by assuming a regression model on various quantiles can be obtained.

The objective of the study is to develop a Quantile version of the Spatial Durbin Model as a statistical model that allows one to study dependencies between variables in different quantiles of the response distribution, which takes into account spatial effects within both dependent and independent variables. To our knowledge, this is the first research to develop Spatial Durbin Model into its quantile version. This model is then used to identify the environmental and quality of life factors (population density per km^2, percentages of households with adequate sanitation, percentages of households with well-drinking sources, percentages of the low-income population, number of floods, and number of submerged households during flood per area) associated with the expected incidence rate of dengue fever varying across Central Java.

This study is organized as follows: in section 2, the description of the data used to fit the models, the methods, including Quantile Regression, Contiguity Weights, Moran's I, Spatial Durbin Model (SDM), and Quantile Spatial Durbin Model (QSDM) are defined. In section 3, data exploration, simulation study of QSDM, and the application of the models (SDM and QSDM) on Central Java's data for dengue disease in 35 districts are shown. The discussion is presented in section 4. Finally, the conclusions and further recommendations are presented in section 5.

2 Methodology

2.1 Data Description

The analyzed dataset is obtained from the Central Java Center Bureau of Statistics. The response ($log(Y_i)$) is the log incidence rate of dengue fever per 100,000 individuals in each of 35 districts in Central Java. The log incidence rate is a log transformation on the total number of cases observed in the study period divided by the total number of people at risk. In this study, six predictors were used, which were believed to affect the dengue virus's spread based on several studies. Those six predictors are:

1. **Population density per km^2** (X_1)**:**
   The number of human inhabitants per square kilometer.

2. **Percentages of households with adequate sanitation** (X_2)**:**
   The percentage of households that have access to proper sanitation services is the ratio between the number of households that have access to appropriate sanitation services to the number of households, expressed in percent (%). Proper sanitation facilities are sanitation equipment that meets health requirements. Namely, those facilities are used by households alone or together with other households, equipped with toilets of the type of gooseneck, and a place for final disposal of feces in the form of a septic tank (Wastewater Treatment Plant).

3. **Percentages of households with well-drinking sources** (X_3)**:**
   The percentage of households that use improved drinking water sources is the percentage of households that use proper (quality) drinking water for drinking. This adequate drinking water includes protected drinking water, including tap water, public hydrant, water terminal, rainwater storage or springs and wells protected, borehole or pump well, which is at least 10 meters away from construction sewage, waste collection, and garbage disposal. Not including bottled water, water from mobile vendors, water sold through tanks, and unprotected wells.

4. **Percentages of low-income population** (X_4)**:**

   Percentages of individuals with an average monthly expenditure per capita below the poverty line.

5. **Number of floods** (X_5)**:**

   Flood events include flash floods and river floods in Central Java.

6. **Number of submerged households during floods** (X_6)**:**

   The number of houses submerged in water and mud caused by the natural disaster.

2.2 Exploratory Data Analysis

Data exploration is a fundamental step in data analysis since it helps to examine a variety of aspects of the dataset. In this study, graphical representation such as scatter plot for covariates, incidence risk plot, and quantile incidence risk plot across the areas were used to explore the data.

2.3 Quantile Regression

Quantile regression as introduced by Koenker and Bassett (1978) quantifies the relationship between dependent and independent variables across various quantiles of the conditional distribution of the dependent variable. Quantile regression is more advantageous than classical mean regression considering its ability to provide complete visualization on the relationship between predictors and the response at all parts of its distribution. In addition, quantile regression is better suited to examining a dependent variable with changes in the distribution and more robust to outlying observations (McMillen, 2012).

Quantile function is the inverse of cumulative distribution function (CDF). Let Y be the response variable and $X_1, \ldots, X_p$ be a set of predictor variables. The CDF of Y given $\boldsymbol{X} = (1, X_1, \ldots, X_p)^\intercal$ can be expressed as $F_Y(y|\boldsymbol{X}) = P(Y \leq y|\boldsymbol{X}) = \tau$ with the inverse function $F_Y^{-1}(\tau|\boldsymbol{X}) = Q_Y(\tau|\boldsymbol{X}) = \inf\{y : F_Y(y|\boldsymbol{X}) \geq \tau \in [0, 1]\}$ called the τ-th conditional quantile of the response Y given covariate $\boldsymbol{X}$ (Koenker, 2005). The linear

model for quantile regression can be formulated as

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

where the vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^{\intercal}$ is the regression coefficient vector and ε defined as the error term in which we assume that the τ-th quantile of ε given $\boldsymbol{X}$ is equal to zero. The distribution of the error (ε) is left unspecified, which is one of the virtues of the method on the robustness. The conditional quantile function of the response Y given $\boldsymbol{X}$ can be written as $Q_Y(\tau|\boldsymbol{X}) = \boldsymbol{X}^{\intercal}\beta^{\tau}$ (Koenker, 2005).

Suppose we have n independent observations $(X_{11}, \ldots, X_{p1}, Y_1), \ldots, (X_{1n}, \ldots, X_{pn}, Y_n)$ from $(X_1, \ldots, X_p, Y)$. The quantile regression parameters are estimated by minimizing a sum of asymmetrically weighted absolute residuals (by giving different weights to positive and negative residuals) with respect to β:

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

where the check function, $\rho_{\tau}(u) = u(\tau - I(u < 0))$ is the tilted absolute value function illustrated in Figure 1 that yields to the τ-th sample quantile as its solution (Koenker and Hallock, 2001). The quantile objective function in (1) is not differentiable, therefore the quantile objective could be elucidated into a linear programming optimization problem to determine the optimal solution (Koenker, 2005).

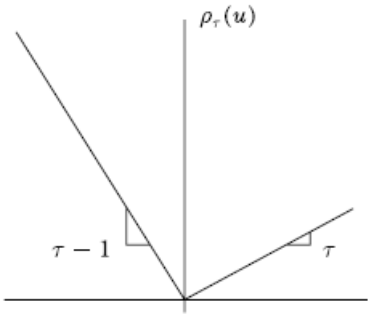


Figure 1: Tilted absolute value function

2.4 Contiguity Weights

Spatial proximity matrices (neighborhood structure) are a key component in the analysis of data with spatial dependence. Formally, the weights express the neighbor structure between the observations as $n \times n$ matrix ($\mathbf{W}$) which contains weights (w_{ij}) for each pair of areas that determines the closeness or spatial relations of these locations (Zhou and Lin, 2008). There are several ways to define the neighborhood structure. For the purpose of this study, queen contiguity weights matrix was used. This neighborhood matrix includes all of each area's immediate neighbors that share common points with the particular area, which can be seen in Figure 2 below.

	C_1	B_2	C_2	
	B_1	A	B_3	
	C_4	B_4	C_3	

Figure 2: Queen contiguity *(share a common edge or vertex)*

In its simplest form, the spatial weights matrix represents the existence of a neighbor relation as a binary association, with weights 1 and 0. Formally, each spatial unit is represented in the matrix by a row i, and the possible neighbors by the columns j, with $j \neq i$. The existence of a neighborhood connection between the spatial unit

corresponding to row i and the one matching column j follows then as $w_{i,j} = \boldsymbol{W}_{i,j} = 1$ and 0 otherwise. Additionally, The elements w_{ii} are commonly set to 0, implying there is no neighborhood structure between an area and itself.

2.5 Moran's I

Spatial autocorrelation is an integral concept in spatial analysis. It refers to the dependency of observation from a certain location to another observation in a different geographical location. Its measurement depends on the defined spatial adjacency matrix. Test of spatial dependency can be done using spatial autocorrelation test with Moran's I method. This measure, formulated by Moran (1950), can be used if the variable of interest is continuous. It is calculated based on the cross products of the deviations of the observation from the overall mean with associated weights. The formula is given by

$$I = \frac{n \sum_i \sum_j \boldsymbol{W}_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\left(\sum_i \sum_j \boldsymbol{W}_{ij}\right) \sum_i (Y_i - \bar{Y})^2}$$

where n is the total number of observations, $\bar{Y}$ is the overall mean, and the weights $\boldsymbol{W}_{ij}$ are obtained from the proximity matrix.

Under the assumption of independence, as shown in Moran (1950), I has an asymptotic normal distribution with mean

$$E[I] = -\frac{1}{n-1}$$

and variance

$$Var(I) = \frac{n^2 \sum_{ij} \boldsymbol{W}_{ij}^2 + 3\left(\sum_{ij} \boldsymbol{W}_{ij}\right)^2 - n \sum_i \left(\sum_j \boldsymbol{W}_{ij}\right)^2}{(n^2 - 1)\left(\sum_{ij} \boldsymbol{W}_{ij}\right)^2}$$

Positive spatial autocorrelation occurs when $I > E[I]$ which means that areas are more alike and negative autocorrelation is when $I < E[I]$. Testing for spatial autocorrelation can be done in two ways: by (i) randomization and (ii) normal approximation. For

randomization, observations are assigned randomly several times in the different areas. The null hypothesis of independence is then rejected when the observed I lies in the tail of this distribution. Using normal approximation, the test statistic

$$Z(I) = \frac{I - E(I)}{\sqrt{VAR(I)}}$$

can be used. Under independence and for large n ($n > 25$), $Z(I)$ is normally distributed with mean 0 and variance 1. If the observed $Z(I)$ lies in the tails of the standard normal distribution then the assumption of independence is rejected.

2.6 Spatial Durbin Model

According to Manski (1993) and Elhorst (2010), various types of interaction effects can be distinguished in modeling spatial data, such as:

1. An endogenous interaction relationship indicates that the dependent variable for a spatial unit depends on the outcomes for other spatial units.

2. A correlated relationship refers to the situation where unobserved environmental characteristics lead to similar outcomes across spatial units (i.e., error terms).

3. An exogenous interaction relationship implies that the outcome for a spatial unit is associated with the covariates of the outcome in other spatial units.

Traditionally, spatial lag and spatial error models have focused on the endogenous interaction and correlated relationship on the error term (Anselin, 1988). However, since explanatory variables also show a spatial pattern, a Spatial Durbin Model (SDM) that includes both a spatially lagged dependent variable and spatially lagged explanatory variables (LeSage and Pace, 2009) is preferred.

The Spatial Durbin Model is an extension of Spatial Autoregressive Regression (SAR) model augmented by spatially lagged explanatory variables (Fischer and Wang, 2011). Via the Spatial Durbin Model we will examine whether the dengue's incidence rate in

one area is also related to neighboring areas' characteristics and, if so, how they are associated by including lags of the dependent variable and the independent variables to quantify the magnitude of direct and indirect effects, which can be written in vector form, as

$$\boldsymbol{Y} = \rho \boldsymbol{W}\boldsymbol{Y} + \alpha \boldsymbol{i}_n + \boldsymbol{X}\beta + \boldsymbol{W}\boldsymbol{X}\gamma + \boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$ represents a $(n \times 1)$ vector of dependent variable, $\boldsymbol{W}$ is a $(n \times n)$ spatial weight matrix, ρ is the parameter that represents the strength of the spatial autoregressive relation in the response, $\boldsymbol{i}_n$ is an $n-$dimensional column vector of ones, α is the intercept coefficient, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)'$ is the $n \times (p-1)$ non-intercept explanatory variable matrix with $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{i(p-1)})$ with $i = 1, \ldots, n$, β is a $(p-1)\times 1$ vector of parameters that measures the marginal impact of the explanatory variables from neighbouring areas on the dependent variable $\boldsymbol{Y}$, γ is the $(p-1)\times 1$ coefficient vector for the exogenous spatially lagged independent variables, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is the error terms, which in the mean regression setting can be defined as a vector with multivariate normal distribution $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$, with $\boldsymbol{I}_n$ as an identity matrix. Multiplying $\boldsymbol{X}$ by $\boldsymbol{W}$ produces spatially lagged covariates that reflect an average of neighbouring observations (Fischer and Wang, 2011).

Rearranging the equation (2) yields:

$$\boldsymbol{Y} = (\boldsymbol{I}_n - \rho \boldsymbol{W})^{-1}(\alpha \boldsymbol{i}_n + \boldsymbol{X}\beta + \boldsymbol{W}\boldsymbol{X}\gamma + \boldsymbol{\varepsilon})$$

The matrix of partial derivatives for this model is:

$$\begin{pmatrix} \frac{\partial Y_1}{X_{1k}} & \ldots & \frac{\partial Y_1}{X_{nk}} \\ \vdots & \ldots & \vdots \\ \frac{\partial Y_n}{X_{1k}} & \ldots & \frac{\partial Y_n}{X_{nk}} \end{pmatrix} = [\boldsymbol{I}_n - \rho \boldsymbol{W}]^{-1} \begin{pmatrix} \beta_k & \ldots & w_{1n}\gamma_k \\ \vdots & \ldots & \vdots \\ w_{1n}\gamma_k & \ldots & \beta_k \end{pmatrix}$$

$$= [\boldsymbol{I}_n - \rho \boldsymbol{W}]^{-1}[\beta_k \boldsymbol{I}_n + \boldsymbol{W}\gamma_k]$$

The $n \times n$ matrix has non-zero elements off the major diagonal, implying the existence of spatial spillovers, which suggests the model parameters are estimated under the explicit assumption of dependency between observations; changes in values for one observation will "spillover" to affect values of other observations. By expressing the power series expansion of the matrix $(\boldsymbol{I}_n - \rho \boldsymbol{W})^{-1}$ as

$$(\boldsymbol{I}_n - \rho \boldsymbol{W})^{-1} = \boldsymbol{I}_n + \rho \boldsymbol{W} + \rho^2 \boldsymbol{W}^2 + \rho^3 \boldsymbol{W}^3 + \dots$$

Golgher and Voss (2016) shows that the model generates a process of "global spillover," designating that changes in an independent variable anywhere in the study domain will influence the value of the dependent variable on the other areas, even when the matrix $\boldsymbol{W}$'s characterization of neighborhood impacts indicates 1st-order of contiguity.

In order to overcome the difficulties to interpret the spillover effect within Spatial Durbin Model, LeSage and Pace (2009) presents a means of summarizing the direct, indirect, and total effects in such models through an averaging process. The direct effect is the value of the main diagonal elements of the partial derivatives matrix (β_k). In other words, the direct effect represents the expected average change across all observations for the dependent variable in a particular area due to an increase of one unit for covariates in this area. The non-diagonal cross-partial derivatives represent the indirect spillovers, which influence the dependent variable $\boldsymbol{Y}_i$ in a region from one unit increase in an explanatory variable in another region $\boldsymbol{X}_j$. Moreover, for the spatial Durbin model, the indirect effect can be divided into two parts, namely, the local effects due to the $\boldsymbol{\gamma}$ coefficient, and the global effect arising from the inverse matrix involving ρ (Elhorst, 2010). The first is local because their impact is only on its adjacent neighbors, while the global influence affects all areas through $(\boldsymbol{I}_n - \rho \boldsymbol{W})^{-1}$ matrix (Golgher and Voss, 2016).

2.7 Quantile Spatial Durbin Model

QSDM model of order τ combines the two approaches mentioned in subsection 2.3 and 2.6. It has a similar form with model (2) with an assumption on the τ-th quantile of ε given $\boldsymbol{X}$ is equal to zero, instead of $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$. Vector $\boldsymbol{\varepsilon}$ contains independent and identically distributed random variables whose distribution is not specified (Trzpiot and Orwat-Acedańska, 2016).

Due to the endogeneity problems (there are spatial lags of the dependent variable $\rho \boldsymbol{WY}$ on the right hand side of the equation) in the model (2), using traditional method to estimate the parameters on Spatial Durbin Model and Quantile Spatial Durbin Model can not be done directly. In QSDM case, an alternative approach based on generalized method of moments method allows the model to be estimated using a variant of two-stage least squares (2SLS) (Kelejian and Prucha, 1998). Chernozhukov and Hansen (2005) extend this approach by using instrumental variables. Though somewhat more complicated, this approach is more robust and has a straightforward formulation of covariance matrix estimates (McMillen, 2012). Below we briefly describe the method:

1. The endogenous variable, $\boldsymbol{WY}$ is regressed on a set of instruments ($\boldsymbol{Z}$). In this study we use $(\boldsymbol{X}, \boldsymbol{WX})$ as a set of instruments for $\boldsymbol{WY}$ as suggested by Kelejian and Prucha (1998).

2. The predicted values of $\widehat{\boldsymbol{WY}}$ are then utilized as an explanatory variable for the quantile regression $\boldsymbol{Y} - \rho \boldsymbol{WY}$ on $\boldsymbol{X}$ and $\widehat{\boldsymbol{WY}}$ with τ quantile, with a grid of alternative values is used for $\rho \in [-1, 1]$. The estimated value of ρ is the value that gives the coefficient on $\widehat{\boldsymbol{WY}}$ that is closest to zero.

3. After finding $\hat{\rho}$, the estimated values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are calculated by a quantile regression of $\boldsymbol{Y} - \hat{\rho} \boldsymbol{WY}$ on $\boldsymbol{X}$ and $\boldsymbol{WX}$ using the same quantile τ.

The motivation of these steps is, if the chosen instruments are optimal, the coefficient on $\widehat{\boldsymbol{WY}}$ will be zero when both the actual variable, $\boldsymbol{WY}$, and the instrumental variable are

incorporated in a regression (Kim and Muller, 2004). The standard errors and p-values for the estimates are calculated using the residual bootstrap as suggested by Trzpiot and Orwat-Acedańska (2016). This method is chosen to ensure that the spatial structure of the data is kept (Anselin, 1988). Kecojevic (2011) described the method as follows:

1. We fit a quantile regression model and calculate $\hat{Q}_Y(\tau|\boldsymbol{X}) = \boldsymbol{X}^\intercal \hat{\beta}^\tau$ to obtain the residuals

$$\boldsymbol{e}^\tau = \boldsymbol{Y} - \hat{Q}_Y(\tau|\boldsymbol{X})$$

   to get a set of empirical residuals on τ-th level $\boldsymbol{e}^\tau = e_1^\tau, e_2^\tau, \ldots, e_n^\tau$ from which a bootstrap re-sample $\boldsymbol{e}_b^* = e_{b,1}^{\tau^*}, e_{b,2}^{\tau^*}, \ldots, e_{b,n}^{\tau^*}$ is drawn with replacement.

2. A bootstrapped vector of the response variable for this re-sample is generated by adding the re-sampled vector of residuals to the vector of fitted response values, from the sample:

$$\boldsymbol{Y}_b^* = \hat{Q}_Y(\tau|\boldsymbol{X}) + \boldsymbol{e}_b^{\tau^*}$$

3. These bootstrapped responses $\boldsymbol{Y}_{b,i}^*$ are then used to estimate quantile regression coefficient and to calculate the standard error for each regression coefficient.

3 Results

3.1 Exploratory Data Analysis

Data exploration is a fundamental step in data analysis since it examines various aspects of the dataset. In this study, graphical representation of log incidence rate for dengue fever spread will be illustrated and test on spatial heterogeneity in the observations using Moran's I index was performed. Additionally, plots of predictors against the response variable are shown to describe the relationship between the variables.

The histogram illustrated in Figure 3 suggests that in 2018, the log incidence rate of dengue fever in Central Java shows a right-skewed distribution. This indicates the need to apply quantile regression as a more robust technique to handle this situation, since there is no assumption of the response's distribution was made on the method.

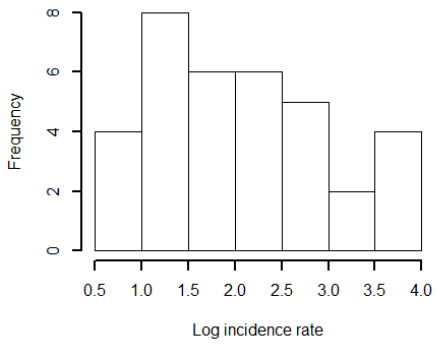


Figure 3: Histogram of the log incidence rate of dengue fever based on data from the 2018 in Central Java

Figure 4 shows that there are more points at the lower values of predictors for population density, number of floods, and number of submerged households during floods. In comparison, percentages of households with well sanitation exhibit more points in higher values, which implies well sanitation on most districts in Central Java. This figure also shows more variability on the response compared to the population density, number of

floods, and the number of submerged households during flood. Moreover, households with well drinking sources and percentages of the low-income population are scattered, with high variability on both the response and the predictors.

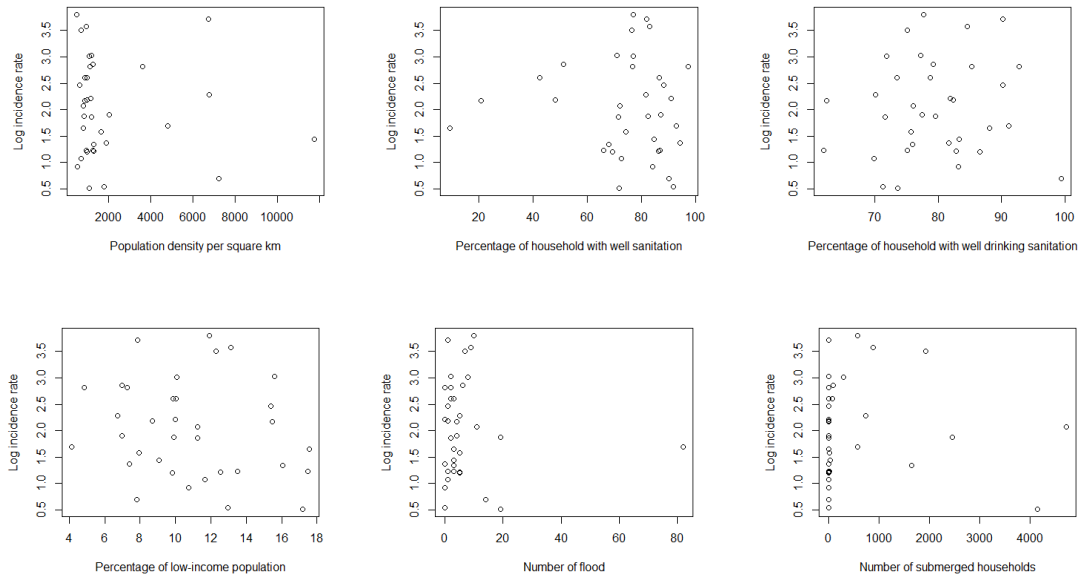


Figure 4: Scatter plot of predictors (left to the right: $X_1 - X_6$) against the response

Figure 5 illustrates the spread of dengue fever's incidence rate and the log incidence rate, over 35 districts in Central Java in 2018. The Incidence rate (left) ranges from 1 to 45 cases, with an average of 12 cases for every 100,000 individuals. According to the figure, the incidence rate of dengue fever varies across districts. 23 out of 35 districts have a low dengue fever risk (risk less than 10 cases for every 100,000 individuals), most of them located side by side in the west part of Central Java. The middle and east parts of the map have a darker color, indicating that dengue fever incidence rate is higher in districts in these areas. The highest incidence rate can be observed in both the middle and east parts of the map, corresponding to Magelang and Blora region, respectively. On the right side of Figure 5, the log incidence rate of dengue disease also show a big variation across districts, with the majority of darker colors observed in the east part of the map.

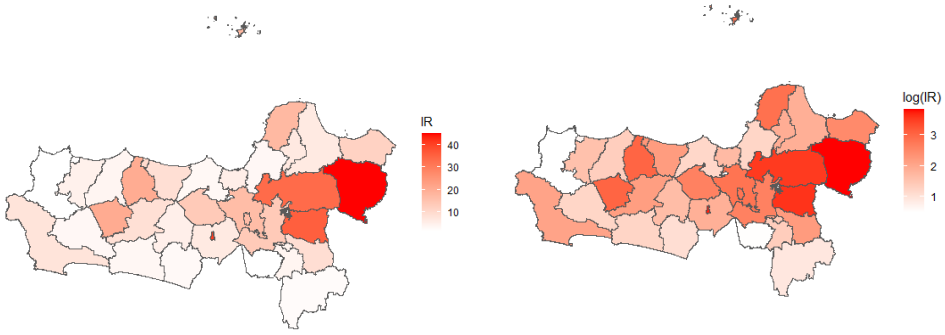


Figure 5: Map of the observed incidence rate per 100,000 individuals *(left)* and Map of the log incidence rate *(right)* for the spread of dengue disease in Central Java on 2018

Figure 6 shows the illustration of the distribution of each covariates across Central Java, which suggests the presence of spatial autocorrelation for the covariates of the data. In addition, Moran's I test was done to check the existence of spatial autocorrelation, which refers to the dependency of an observation from a particular location to another observation in different geographical areas. This Moran's I tests resulted in a p-value equal to 0.0061 for the response, which implies spatial autocorrelation in the log incidence rate of dengue fever. These findings indicates that the presence of spatial autocorrelation within covariates and the response, hence the use of Spatial Durbin Model would be logical to be used.

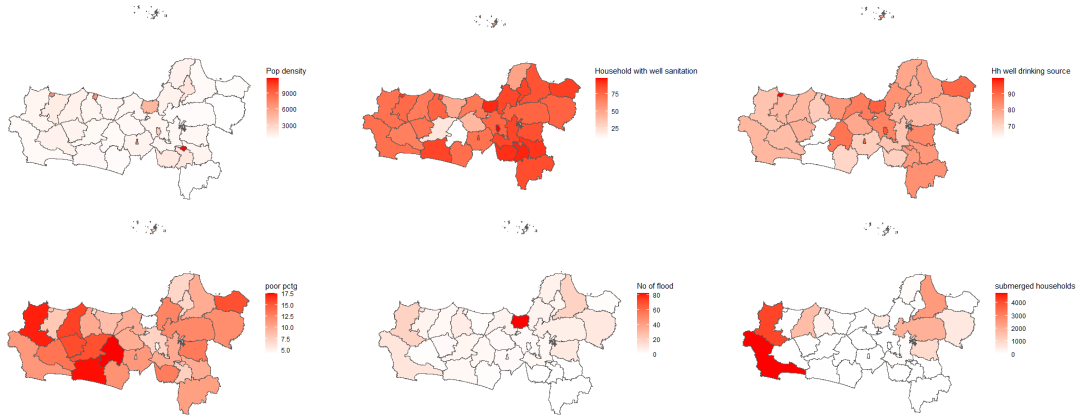


Figure 6: Map of the covariates (left to the right: $X_1 - X_6$) with lower value of the covariates when the color is lighter (*white*) and darker color when the value is higher (*red*).

Figure 7 represents a map with the quantile of log incidence rate in Central Java. As can be seen from the figure, the log incidence rate of dengue fever varies across the region. Based on all the aforementioned findings, it can be concluded that the use of a Spatial Durbin Quantile Regression Model is reasonable.

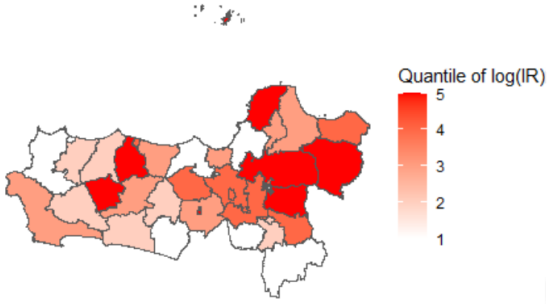


Figure 7: Map of the quantiles of log incidence rate with *Quantile 1* (≤20%) - [0; 1.2217), *Quantile 2* (20%-40%) - [1.2217; 1.6753), *Quantile 3* (40% - 60%) - [1.6753; 2.1963), *Quantile 4* (60% - 80%) - [2.1963; 2.8240), and *Quantile 5* (≥80%) - [2.8240; 3.8068)

3.2 Simulation study of Quantile Spatial Durbin Model

In this subsection, the Quantile Spatial Durbin Model (QSDM) performance will be evaluated. 100 simulations were conducted for $\rho = 0.1, 0.2, \ldots, 0.9$ and $\tau = 0.1, 0.2, \ldots, 0.9$. Let θ denotes the parameters to be assessed (e.g., β, γ, and ρ). Root Mean Square Error for θ will be calculated (RMSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}$) with N is the number of simulations, to evaluate the performance of each parameter estimates. Using the formula on (2), we chose a value of each parameter, $\alpha = 1.439$, $\beta = 3.942$, and $\gamma = -0.182$. $\boldsymbol{\varepsilon}$ are i.i.d sampled from a standard normal distribution with a sample size of 35, $\boldsymbol{X_1} \sim N(10, 1)$ are sampled from a normal distribution with mean 10 and standard deviation of 1, and $\boldsymbol{W}$ is the queen spatial weight matrix for Central Java. The resulting model can be written as

$$\boldsymbol{Y} = \rho \boldsymbol{W}\boldsymbol{Y} + 1.439 + 3.942\boldsymbol{X_1} - 0.182\boldsymbol{W}\boldsymbol{X_1} + \boldsymbol{\varepsilon}$$

Figure 8 shows the results of the model performance in terms of estimating spatial autocorrelation (ρ) in different level of quantiles ($\tau = 0.1, 0.2, \ldots, 0.9$). The horizontal and vertical axis show the level of quantiles and ρ, respectively. It can be seen that the estimation of ρ in each level of quantiles are close to the true parameter with relatively small variability and small numbers of outliers. It can also be observed that the median of $\hat{\rho}$ is considerably close to the actual value of ρ.

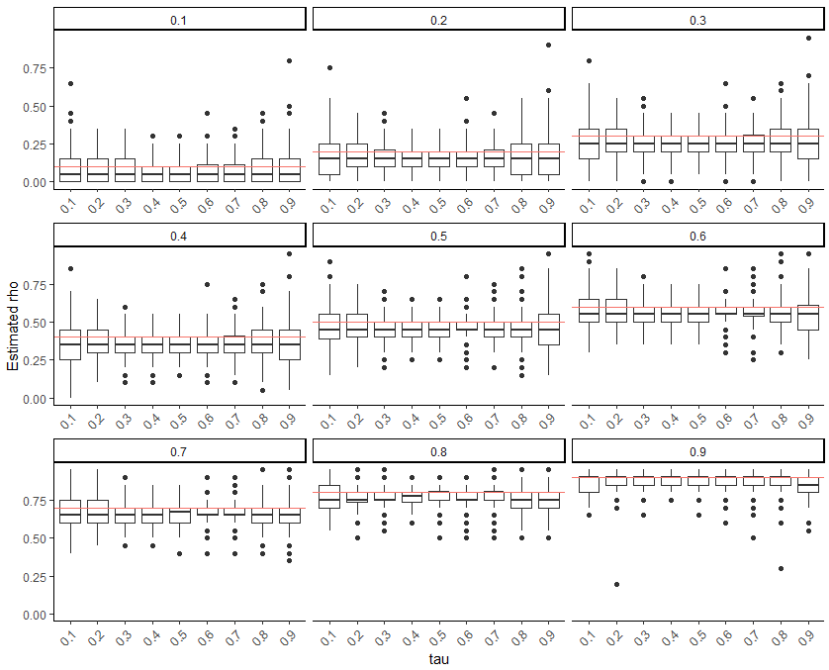


Figure 8: Boxplot of the distribution of $\hat{\rho}$ for $\rho = 0.1, 0.2, \ldots, 0.9$ (from *upper left* to *bottom right*) within 100 simulations. Each small boxes represent the distribution of estimated spatial autocorrelation ($\hat{\rho}$) in different level of quantiles ($\tau = 0.1, 0.2, \ldots, 0.9$).

A closer look at the model performance can be obtained from Figure 9, which provides the RMSE value and its distribution for $\hat{\rho}$, $\hat{\beta}$ and $\hat{\gamma}$ within 100 simulations. It can be observed from the line figures *(left)* that all parameter estimates have small mean of RMSE in every level of quantiles. Moreover, $\hat{\rho}$ and $\hat{\beta}$ suggest a clear U-shaped patterns, indicating smallest RMSE in the median and higher values of RMSE in the first and last level of τ. A same conclusion can be noticed from the boxplot *(right)* of $\hat{\rho}$ and $\hat{\beta}$. For the estimation of the spatial lag parameter ($\hat{\gamma}$), the boxplot suggests small RMSE

for the median and higher variability in the first and last level of τ, although no clear pattern can be concluded.

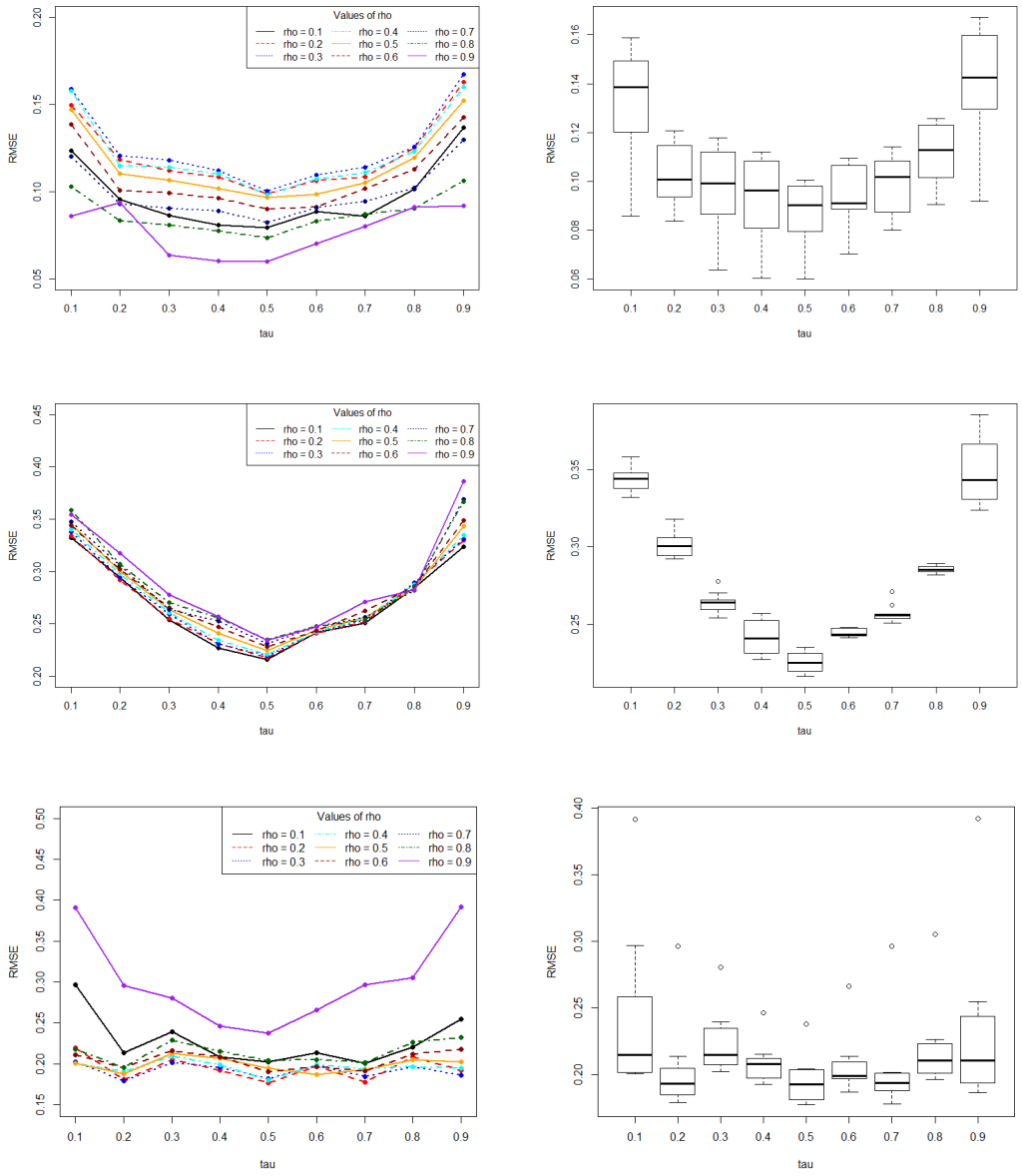


Figure 9: RMSE value *(left)* and its distribution *(right)* on $\hat{\rho}$ *(first row)*, $\hat{\beta}$ *(second row)*, $\hat{\gamma}$ *(third row)* for $\rho = 0.1, 0.2, \ldots, 0.9$ in each $\tau = 0.1, 0.2, \ldots, 0.9$ within 100 simulations.

3.3 Spatial Durbin Model

Spatial Durbin Model (SDM) with log-transformed response variable is fitted to ensure the resulting incidence risks to be positive. Parameter estimates on SDM are presented on Table 1. It can be seen that low income is significantly affecting the log-incidence rate of dengue cases in Central Java, and there is a significant ($\alpha = 5\%$) global spatial autocorrelation ($\rho = 0.5815$) in the data.

Table 1: Parameter estimate of Spatial Durbin Model

Variable	Estimate	Std. Error	z-value	Pr(>|z|)
Population_density_per_sq_km (β_1)	-0.00008	0.00006	-1.3640	0.1725
Household_well_sanitation_pctg (β_2)	-0.00933	0.00795	-1.1733	0.2460
Household_well_drinking_source_pctg (β_3)	-0.00987	0.01916	0.5152	0.6064
Poor_pctg (β_4)	-0.11798	0.04675	-2.5235	0.0116
No_of_Flood (β_5)	-0.00717	0.01066	-0.672	0.5011
Submerged_households (β_6)	0.00002	0.00012	0.1723	0.8631
lag.Population_density_per_sq_km (γ_1)	-0.00023	0.00018	-1.2560	0.2091
lag.Household_well_sanitation_pctg (γ_2)	0.01265	0.14661	0.8632	0.3880
lag.Household_well_drinking_source_pctg (γ_3)	-0.02662	0.05195	-0.5124	0.6083
lag.Poor_pctg (γ_4)	0.06197	0.12636	0.4905	0.6238
lag.No_of_Flood (γ_5)	-0.02536	0.03738	-0.6784	0.4975
lag.Submerged_households (γ_6)	0.00011	0.00034	0.3287	0.7423
$\rho = 0.58151$				0.0151

Since direct interpretation of the Spatial Durbin Model's coefficients is difficult because they do not represent actual partial derivatives, the estimates from this model cannot be interpreted as partial derivatives in a typical regression model. Therefore, the signs and magnitudes resulting from changes in the explanatory variables are reported as summary measures of total, direct and indirect effects shown in Table 2. Table 2 indicates that the direct effect coefficient of percentages of households with well-drinking sources and number of submerged households during floods has a positive effect on local log incidence rate. At the same time, the spatial spillover effect coefficient of the percentages of households with adequate sanitation and number of submerged households during floods will positively affect the log incidence rate of dengue of its surrounding areas. From the total effect in Table 2, it can be observed that the log incidence rate of dengue in Central

Java will increase as the percentages of households with adequate sanitation and number of submerged households during floods increases. Moreover, decomposing the marginal effects into direct (own-area) and indirect (spillover) will also examine the effect that significantly impacting the total value.

Table 2: Decomposition estimates of the total, direct and indirect effects of Spatial Durbin Model (SDM)

Variable	Direct Effect	Indirect Effect	Total Effect
Population_density_per_sq_km	-0.0001	-0.0006	-0.0007
Household_well_sanitation_pctg	-0.0079	0.0158	0.0079
Household_well_drinking_source_pctg	0.0057	-0.0458	-0.4002
Poor_pctg	-0.1192	-0.0145	-0.1338
No_of_Flood	-0.1293	-0.0648	-0.0777
Submerged_households	0.00004	0.00027	0.0003

3.4 Quantile Spatial Durbin Model Application to Dengue Disease in Central Java

In the following section, Quantile Spatial Durbin Model for dengue disease in 35 districts on Central Java is utilized. From Figure 10, it can be seen that the spatial estimated autoregressive coefficient $\hat{\rho}$ is relatively lower at $\tau < 0.5$ and higher at $\tau \geq 0.5$. This shows that in districts with lower log incidence rate of dengue disease ($\tau < 0.5$), any changes in an independent variable that happen in the neighborhood anywhere, will have low impact on the value of the dependent variable on the other areas.

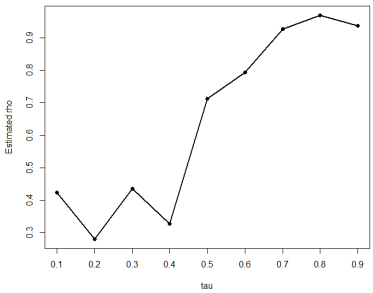


Figure 10: Quantile estimates of spatial autoregressive coefficients ($\hat{\rho}$) for dengue disease in Central Java based on Quantile Spatial Durbin Model.

Considering the influence factors and its lagged variables illustrated in Figure 11, we obtain the following results:

- The coefficient of variable *Population density per km*^{2} shows small variability and resulting in negative values across all quantiles, except for $\tau = 0.1$. This shows a negative correlation between the covariate and log incidence rate of dengue disease in Central Java, that is, the risk ratio of dengue disease for areas with lower population density per km^2 is higher. The same pattern can be obtained from the *Lag of population density per km*^{2}, indicating that an increase in population density per km^2 on direct neighbors of an area will have a negative effect of dengue disease log incidence rate in Central Java.

- The estimates of the parameter of the variable *Percentages of households with adequate sanitation* are negative for all $\tau < 0.9$ and the estimates are relatively smaller at the lower quantile. This indicates that the percentages of households with adequate sanitation increases the log incidence rate of dengue disease in Central Java. On the other hand, *Lag of percentages of households with adequate sanitation* show positive results for all level of quantiles, suggesting that a better sanitation on the adjacency neighbors will positively impact the log incidence rate of dengue disease in a particular area.

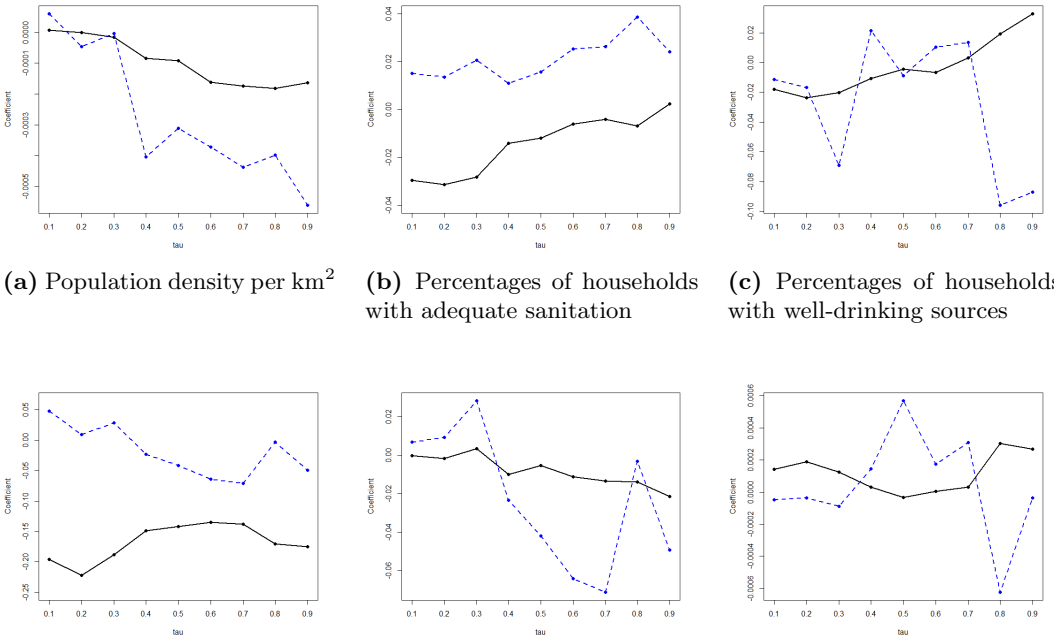
- The influence of *Percentages of households with well-drinking sources* shows a similar pattern with the percentages of households with adequate sanitation. This estimates shows a negative association in $\tau < 0.7$, and positive association otherwise. This suggests that there is a higher risk ratio of dengue disease for areas with more households having well-drinking sources. Moreover, *Lag of percentages of households with well-drinking sources* shows no clear pattern.

- The estimates of the parameter of the variable *Percentages of low-income population* are negative for all quantiles. This indicates that the level of economic development reduces the log incidence rate of dengue disease in Central Java. The *Lag of percentages of low-income population* also indicates that, when the neighbor

of an area has a higher percentage of low-income population, log incidence rate of dengue disease in Central Java will decrease.

- The estimates of the variable *Number of floods* shows small changes across quantiles with negative values at all quantiles, except for $\tau = 0.3$. This shows a negative association between the number of floods and log incidence rate of dengue disease in Central Java, that is, the risk ratio of dengue disease for areas with more floods is lower. The *Lag of number of floods* implies a positive association for smaller quantiles ($\tau < 0.4$) and negative association otherwise. This indicates that the number of floods in the neighborhood areas reduces the risk ratio of dengue disease in a particular area in Central Java.

- For $\tau \neq 0.5$, the *Number of submerged households during floods* shows a positive association with dengue disease log incidence rate in Central Java. On the other hand, *Lag of number of submerged households during floods* shows negative values on $\tau = 0.1, 0.2, 0.3, 0.8$ and 0.9. This suggests that the lower and higher values of the number of submerged households during floods in the neighborhood areas will reduces the log incidence rate of dengue disease in a particular area in Central Java.



(a) Population density per km^2

(b) Percentages of households with adequate sanitation

(c) Percentages of households with well-drinking sources

(d) Percentages of low-income population

(e) Number of floods

(f) Number of submerged households during floods

Figure 11: Regression coefficients of different influential factors for dengue disease in Central Java based on Quantile Spatial Durbin Model. The black line and blue dashed line shows the regression coefficients of different influential factors and lag of influential factors in each $\tau = 0.1, 0.2, \ldots, 0.9$, respectively

As mentioned earlier, direct interpretation of the Spatial Durbin Model parameter estimates is difficult because of the spillover effect, which also applies in Quantile Spatial Durbin Model. As a result, the values that arise from changes in the explanatory variables on three chosen quantiles ($\tau = 0.2, 0.5, 0.7$) are presented as a summary of total, direct, and indirect effects in Table 3. These three levels of quantiles are chosen to represent the lower, middle, and upper value of the dependent variable. Dengue fever risk is negatively correlated with population density and poor percentages, whereas submerged households gives a positive association across the chosen quantiles. A negative value for direct effects of the number of flood cases across quantiles indicates a negative relationship between dengue fever and flood cases in that quantile, whereas a positive value for indirect effect on the lower quantile indicates that flooding events in the sur-

rounding areas will increase the dengue risk of that quantile. It's also worth noting that an area surrounded by households with good sanitation has a higher dengue fever incidence risk across quantiles. Households with good sanitation in a given district, on the other hand, have a negative relationship with dengue fever incidence rates in the middle and lower quantiles. Lastly, households with good drinking sources negatively impact dengue disease in the low and middle quantiles while having a positive impact in the high quantiles.

Table 3: Decomposition estimates of the total, direct, and indirect effects of Quantile Spatial Durbin Model (SDM) on $\tau = 0.2, 0.5, 0.7$

Variable	Direct Effect	Indirect Effect	Total Effect
	$\tau = 0.2$		
Population_density_per_sq_km	-0.000003	-0.00006	-0.00006
Household_well_sanitation_pctg	-0.0309	0.0063	-0.0246
Household_well_drinking_source_pctg	-0.0254	-0.0306	-0.0561
Poor_pctg	-0.2261	-0.0687	-0.2948
No_of_Flood	-0.0011	0.0118	0.0108
Submerged_households	0.0002	0.00002	0.0002
	$\tau = 0.5$		
Population_density_per_sq_km	-0.0002	-0.0012	-0.0014
Household_well_sanitation_pctg	-0.0099	0.0226	0.0127
Household_well_drinking_source_pctg	-0.0079	-0.0379	-0.0458
Poor_pctg	-0.1585	-0.1815	-0.3401
No_of_Flood	-0.0187	-0.1447	-0.1635
Submerged_households	0.0001	0.00173	0.0019
	$\tau = 0.7$		
Population_density_per_sq_km	-0.0007	-0.0077	-0.0084
Household_well_sanitation_pctg	0.0156	0.2883	0.3039
Household_well_drinking_source_pctg	0.0179	0.2137	0.2317
Poor_pctg	-0.1684	-0.4433	-0.6117
No_of_Flood	-0.0864	-1.0697	-1.1562
Submerged_households	0.0003	0.0043	0.0046

Figure 12 shows the predictions from QSDM model based on three chosen quantiles ($\tau = 0.2, 0.5, 0.7$). Since the response was modeled by a logarithm function for every 100,000 individuals, exponentiating the prediction of the model is required to compute the incidence rate for each region. For $\tau = 0.2$, the map of predictions shows there is one

district at southwest part of the map with the lightest (white) color. This suggests that, this district (Kebumen) predicts the lowest log incidence rate of dengue. This can be interpreted as, 20 out of 100 individuals have a risk of dengue fever of $e^{2.679}/100000 = 0.0004983\%$. From the medium percentile i.e. 50th percentile, the map of predictions from the QSDM shows almost similar log incidence rate of dengue across Central Java. Lastly, for $\tau = 0.7$, the map of predictions shows that the districts located at the eastern part of Central Java have a higher log incidence rate of dengue fever, with the top 3 highest predictions are located in Blora, Pati, and Rembang with the the predictions of incidence rate are 4.3%, 2.9%, and 2.63%, respectively.

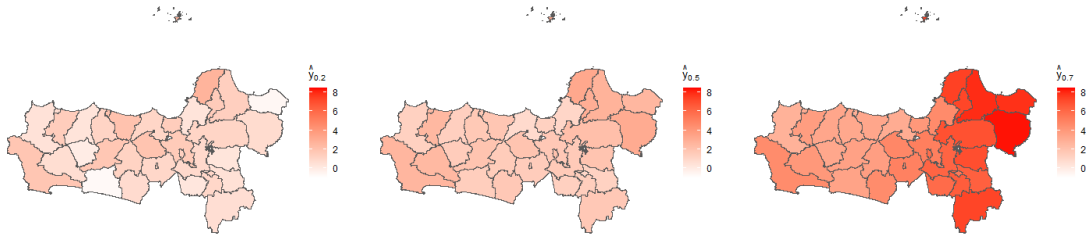


Figure 12: Map of the predictions from QSDM model on $\tau = 0.2, 0.5, 0.7$ (*left to right*), respectively, with lower value of the prediction when the color is lighter (*white*) and darker color when the value is higher (*red*).

4 Discussion

The Quantile Spatial Durbin Model (QSDM) is used to answer the research question in the study, that is, to model both a spatially lagged dependent variable and spatially lagged explanatory variables in various quantiles for dengue fever across Central Java.

The case study provided here is about the danger of dengue fever in Central Java, Indonesia. Residential areas, the availability of water storage vessels, environmental sanitation systems, low-income populations, and flooding events are all variables that contribute to the spread of dengue fever. In Central Java, some districts do not have well sanitation and drinking water supplies for their residents. Economy inequality also raises the risk of dengue transmission by increasing exposure to risk variables associated with dengue transmission, as the vector prefers urban construction, human agglomeration, and a considerable increase in population density. Moreover, floods can also raise the danger of diseases spread by vectors or carriers, such as dengue fever. Although the flight range of Aedes is limited, it can be transmitted from one region to another or from infected humans' mobility. This implies the spread of dengue fever between surrounding districts may have an impact on one another. To better identify the risk of dengue disease across the locations with high and low risk of dengue fever, the QSDM is utilized for the analysis.

In the setting of examining the dengue's incidence rate in one area and its relation to neighboring areas' characteristics, Spatial Durbin Model (SDM) is commonly used. However, this model regresses the predictors to the conditional mean of the response, which cannot characterize the entire distribution of dengue fever risk. We are interested in evaluating the dengue fever risk on various portions of the distribution in this study so that some recommendations may be made to the government for places with high dengue fever risk, emphasizing the upper quantile of dengue fever risk. This aim can be accomplished by utilizing QSDM to allow for the computation of conditional quantiles, allowing the researcher to have a deeper understanding of the predictor-response rela-

tionship across districts and percentiles of the response. Furthermore, unlike the SDM model, QSDM has the benefit of not having distribution assumptions on the error term.

The findings suggest that the eastern areas of Central Java have a greater dengue fever incidence rate for higher quantiles. This is unsurprising, given that Blora has had one of the highest incidence rates in recent years (Ginandra, 2015). Looking at the total effect as a combined result from direct and indirect effects, dengue fever incidence risk in lower, middle and upper quantile is constantly inversely proportional to population density and low-income percentages, whereas the number of submerged households due to floods has a positive association. An unexpected discovery was discovered for well sanitation percentages, meaning that it raises the risk of dengue fever in the middle and higher quantiles, as well as the percentages of well drinking sources in the higher quantiles. The frequency of floods also has a surprising finding for the middle and higher quantiles, indicating a negative association for dengue fever incidence risk.

Schmidt *et al* (2011) stated that dengue transmission may be more suited in sparsely populated areas than in densely populated places. In high-density areas, the vector/host ratio may be less conducive to intense transmission, but absolute case numbers can still be substantial. Dengue spreads in waves through regions, amplifying sites with high vector/host ratios, such as low-density areas. The water quality can affect the laying of eggs by the mosquito Aedes aegypti (Marques *et al*, 2013). Mosquito eggs will grow successfully under suitable water characteristics. In Indonesia, including Central Java, tiled tub *(bak mandi)*, built in the corner, which always filled with water for bathing and drinking purposes are widely used. Many immature mosquitoes were found in or around houses with these tiled water tub, producing more pupae per house than all other outdoor containers combined (Nelson *et al*, 1976). These good quality of stagnant water in tiled tubs might explain why the incidence of dengue fever is higher in areas with good sanitation and drinking water. Apart from lower water quality, locations with a significant low-income population are more likely to have unprotected drinking sources,

such as unprotected dug wells, ponds, and spring waters, which provide a natural habitat for Aedes larvae predators. The survival rate of Aedes' larva was found to be drastically reduced when these natural predators were present (Couret *et al*, 2020). Furthermore, our findings for the amount of flood and submerged households confirm Few *et al* (2004) arguments, stating that dengue fever is unlikely to be a major issue caused by flood since many Aedes breeding sites are likely to be overwhelmed by floodwaters. Recessing floodwaters, on the other hand, may provide perfect breeding environments in the post-onset phase.

5 Conclusion

The application of QSDM on the dengue fever dataset in Central Java allows us to identify the incidence rate of the disease by taking into account spatially lagged response and explanatory variables in various quantiles. The results reveal that predictor effects varied depending on quantile. This information is crucial because it will help the government tackle the dengue problem precisely by determining which predictor has a significant influence, especially at higher quantile levels. Based on the estimation of QSDM, the districts that are at high risk for dengue are located in the eastern areas of Central Java (e.g., Pati, Rembang, and Blora). The government can use this information to focus more on these areas to reduce dengue fever cases. To summarize, a low-density area with a stable economy, good water infrastructure, and poor flood management will be at higher risk of dengue disease. For further studies, it is suggested that other risk factors for dengue disease, such as the larva-free and healthy-clean practices in households indexes, can be considered. Additionally, expanding the model into a spatio-temporal model by utilizing data for periods of time may result in a better outcome since more data is being used for the study.

References

Anselin, L. 1988, Spatial Econometrics: Methods and Models. Dordrecht: Kluwer.

Beasley, D. W., & Barrett, A. D. (2008). The infectious agent. *Dengue: Tropical Medicine*. Volume, 5, 29-73.

Bich, T. H., Quang, L. N., Thanh Ha, L. T., Duc Hanh, T. T., & Guha-Sapir, D. (2011). Impacts of flood on health: epidemiologic evidence from Hanoi, Vietnam. *Global Health Action*, 4(1), 6356.

Chen, Y. R., Hwang, J. S., & Guo, Y. J. (1994). Ecology and control of dengue vector mosquitoes in Taiwan. *The Kaohsiung Journal of Medical Sciences*, 10, S78-87.

Chernozhukov, V., & Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, 73(1), 245-261.

Chernozhukov, V., & Hansen, C. (2008). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142(1), 379-398.

Couret, J., Notarangelo, M., Veera, S., LeClaire-Conway, N., Ginsberg, H. S., & LeBrun, R. L. (2020). Biological control of Aedes mosquito larvae with carnivorous aquatic plant, Utricularia macrorhiza. *Parasites & Vectors*, 13, 1-11.

Elhorst, J. P. (2010). Applied spatial econometrics: raising the bar. *Spatial Economic Analysis*, 5(1), 9-28.

Few, R., Ahern, M., Matthies, F., & Kovats, S. (2004). Floods, health and climate change: a strategic review.

Fischer, M. M., & Wang, J. (2011). *Spatial data analysis: models, methods and techniques.* Springer Science Business Media.

Ginandra, I. W. (2015). *Hubungan Tingkat Pengetahuan Kepala Keluarga Dengan Perilaku Pencegahan Demam Berdarah Dengue Di Desa Sendangmulyo Kabupaten Blora* (Doctoral dissertation, Universitas Muhammadiyah Surakarta).

Gubler, D. J. (2011). Dengue, urbanization and globalization: the unholy trinity of the 21st century. *Tropical Medicine and Health*, 39(4), S3-S11.

Golgher, A. B., & Voss, P. R. (2016). How to interpret the coefficients of spatial models: Spillovers, direct and indirect effects. *Spatial Demography*, 4(3), 175-205.

Haryanto, B. (2018). Indonesia dengue fever: Status, vulnerability, and challenges. *Current Topics in Tropical Emerging Diseases and Travel Medicine*, 5, 81-92.

Jahan, F. (2011). Dengue fever (DF) in Pakistan. *Asia Pacific family medicine*, 10(1), 1-4.

Kelejian, H. H., & Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17(1), 99-121.

Kecojevic, T. (2011). *Bootstrap Inference for Parametric Quantile Regression*. The University of Manchester (United Kingdom).

Kim, T. H., & Muller, C. (2004). Two-stage quantile regression when the first stage is based on quantile regression. *The Econometrics Journal*, 7(1), 218-231.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4), 143-156.

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

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33-50.

Kostov, P. (2009). A spatial quantile regression hedonic model of agricultural land prices. *Spatial Economic Analysis*, 4(1), 53-72.

LeSage, J., & Pace, R. K. (2009). Introduction to spatial econometrics. Chapman and Hall/CRC.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3), 531-542.

Marques, G. R., Chaves, L. S. M., Serpa, L. L. N., Arduíno, M. D. B., & Chaves, F. J. M. (2013). Água de abastecimento público de consumo humano e oviposição de Aedes aegypti. *Revista de Saúde Pública*, 47, 579-587.

McMillen, D. P. (2012). Quantile regression for spatial data. Springer Science Business Media.

Moran, P.A.P, 1950, Notes on continuous stochastic phenomena, *Biometrika* 37:17-23.

Nelson, M. J., Pant, C. P., Self, L. S., & Usman, S. (1976). Observations on the breeding habitats of Aedes aegypti (L.) in Jakarta, Indonesia. *The Southeast Asian Journal of Tropical Medicine and Public Health*, 7(3), 424-429.

Schmidt, W. P., Suzuki, M., Thiem, V. D., White, R. G., Tsuzuki, A., Yoshida, L. M., ... & Ariyoshi, K. (2011). Population density, water supply, and the risk of dengue fever in Vietnam: cohort study and spatial analysis. *PLoS Med*, 8(8), e1001082.

Stanaway, J. D., Shepard, D. S., Undurraga, E. A., Halasa, Y. A., Coffeng, L. E., Brady, O. J., ... & Murray, C. J. (2016). The global burden of dengue: an analysis from the Global Burden of Disease Study 2013. *The Lancet Infectious Diseases*, 16(6), 712-723.

Surendran, S. N., Jude, P. J., Thabothiny, V., Raveendran, S., & Ramasamy, R. (2012). Pre-imaginal Development of Aedes Aegypti in Brackish and Fresh Water Urban Domestic Wells in Sri Lanka. *Journal of Vector Ecology: Journal of the Society for Vector Ecology,* 37(2), 471-473.

Thammapalo, S., Chongsuvivatwong, V., Geater, A., & Dueravee, M. (2008). Environmental factors and incidence of dengue fever and dengue haemorrhagic fever in an urban area, Southern Thailand. *Epidemiology Infection*, 136(1), 135-143.

Trzpiot, G., & Orwat-Acedańska, A. (2016). Spatial quantile regression in analysis of healthy life years in the European Union countries. *Comparative Economic Research*, 19(5), 179-199.

Weiss, R. A., & McMichael, A. J. (2004). Social and environmental risk factors in the emergence of infectious diseases. *Nature Medicine*, 10(12), S70-S76.

Whitehorn, J.,& Simmons, C. P. (2011). The pathogenesis of dengue. *Vaccine*, 29(42), 7221-7228.

World Health Organization. (1997). Dengue haemorrhagic fever: diagnosis, treatment, prevention and control. World Health Organization.

World Health Organization, Special Programme for Research, Training in Tropical Diseases, World Health Organization. Department of Control of Neglected Tropical Diseases, World Health Organization. Epidemic, & Pandemic Alert. (2009). *Dengue: guidelines for diagnosis, treatment, prevention and control. World Health Organization.*

WHO. (2016) Dengue vaccine: WHO position paper. *Weekly Epidemiological Record*, 93, 457–476.

Zhou, X. & Lin H. (2008) Spatial Weights Matrix. In: Shekhar S., Xiong H. (eds) Encyclopedia of GIS. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-35973-1_1307

Appendix

Table 4: Parameter estimate of Quantile Spatial Durbin Model (QSDM) for $\tau = 0.2, 0.5, 0.7$ with $nboot = 10,000$.

Variable	Estimate	Std. Error	z-value	Pr(>|z|)
$\tau = 0.2$				
Population_density_per_sq_km (β_1)	-0.0000006	0.00008	-0.0071	0.9943
Household_well_sanitation_pctg (β_2)	-0.0312	0.01024	-3.0507	0.0022
Household_well_drinking_source_pctg (β_3)	-0.0238	0.02251	-1.0554	0.2912
Poor_pctg (β_4)	0.2224	0.05918	-3.7571	0.0002
No_of_Flood (β_5)	-0.0017	0.01577	-0.1052	0.9161
Submerged_households (β_6)	0.0019	0.00015	1.1784	0.2686
Lag_Poulation_density_per_sq_km (γ_1)	-0.00005	0.0002	-0.2136	0.8308
Lag_Household_well_sanitation_pctg (γ_2)	0.01353	0.0181	0.7479	0.4544
Lag_Household_well_drinking_source_pctg (γ_3)	-0.0166	0.0579	-0.2866	0.7744
Lag_Poor_pctg (γ_4)	0.01009	0.1568	0.0643	0.9486
Lag_No_of_Flood (γ_5)	0.0094	0.0438	0.2153	0.8295
Lag_Submerged_households (γ_6)	-0.00004	0.0004	-0.08651	0.9311
ρ	0.2800	0.2079	1.3463	0.1781
$\tau = 0.5$				
Population_density_per_sq_km (β_1)	-0.000092	0.0001	-0.8132	0.4161
Household_well_sanitation_pctg (β_2)	-0.0120	0.0137	-0.8705	0.3841
Household_well_drinking_source_pctg (β_3)	-0.0043	0.0338	-0.1288	0.8975
Poor_pctg (β_4)	-0.1416	0.0808	-1.7514	0.0798
No_of_Flood (β_5)	-0.0053	0.0161	-0.3304	0.7411
Submerged_households (β_6)	-0.00003	0.0002	-0.1647	0.8692
Lag_Poulation_density_per_sq_km (γ_1)	-0.0003	0.0003	-0.9431	0.3456
Lag_Household_well_sanitation_pctg (γ_2)	0.0156	0.0245	0.6368	0.5243
Lag_Household_well_drinking_source_pctg (γ_3)	-0.0088	0.0881	-0.1001	0.9201
Lag_Poor_pctg (γ_4)	0.0437	0.2084	0.2096	0.8339
Lag_No_of_Flood (γ_5)	-0.0417	0.0672	-0.6216	0.5334
Lag_Submerged_households (γ_6)	0.0006	0.0005	1.0114	0.3118
ρ	0.712	0.3617	1.9686	0.0490
$\tau = 0.7$				
Population_density_per_sq_km (β_1)	-0.0002	0.0002	-0.8018	0.4226
Household_well_sanitation_pctg (β_2)	-0.0041	0.0259	-0.1585	0.8741
Household_well_drinking_source_pctg (β_3)	0.0033	0.0625	0.0527	0.9579
Poor_pctg (β_4)	-0.1381	0.1515	-0.9107	0.3624
No_of_Flood (β_5)	-0.0132	0.0303	-0.4356	0.6631
Submerged_households (β_6)	0.0003	0.0004	0.0786	0.9373
Lag_Poulation_density_per_sq_km (γ_1)	-0.00043	0.0005	-0.8134	0.4159
Lag_Household_well_sanitation_pctg (γ_2)	0.02631	0.0465	0.5652	0.5718
Lag_Household_well_drinking_source_pctg (γ_3)	0.01361	0.1599	0.0851	0.9321
Lag_Poor_pctg (γ_4)	0.09341	0.3875	0.2411	0.8095
Lag_No_of_Flood (γ_5)	-0.07118	0.1303	-0.5459	0.5851
Lag_Submerged_households (γ_6)	0.00031	0.0011	0.2761	0.7824
ρ	0.9270	0.1874	4.9456	$< .0001$

R Code

```
#-- Define neighborhood based on sharing boundaries--#
belqnb <- poly2nb(shape,queen=T)

# Plot of neighborhood structure
plot(shape,border="black")
plot(belqnb,coord,add=TRUE,col="red")

# create neighbours list with row-standardized weigths
q.W <- nb2listw(belqnb, style = "W")

# Calculate Moran's I statistics
moran(x = plotvar, listw = q.W,
      n = narea, S0 = Szero(q.W))

# Spatial Durbin Model
#-------------------------
durbin <- lagsarlm(log(risk) ~ Population_density_per_sq_km +
    Household_well_sanitation_pctg + Household_well_drinking_source_pctg +
    poor_pctg + No_of_Flood + submerged_households, Durbin = T,
    data = data.final, list = q.W, tol.solve = 1.0e-20)

summary(durbin)
impacts(durbin, listw = q.W)

# SDM Quantile Reg
#-----------------------

data.durbin <- data.final %>% dplyr::mutate(log.risk = log(risk)) %>%
  dplyr::select(-c(risk, dengue, Population))

form <- log.risk ~ Population_density_per_sq_km + Household_well_sanitation_pctg +
  Household_well_drinking_source_pctg + poor_pctg +
```

```
  No_of_Flood + submerged_households

mt <- terms(form, data = data.durbin)
mf <- lm(form, data = data.durbin, method = "model.frame")
x <- model.matrix(mt,mf)
lag <- create_WX(x, listw = q.W)
form <- update(form, log.risk ~ .+ lag)

qdurbin.u <- function(form, wy = NULL, wmat = NULL, inst = NULL, winst = NULL,
                      shpfile = NULL, tau = 0.5,
                      rhomat = NULL, data = NULL, silent = F, q.W = NULL){
  mt <- terms(form, data = data)
  mf <- lm(form, data = data, method = "model.frame")
  x <- model.matrix(mt,mf)
  lags <- create_WX(x, listw = q.W)

  qdata <- model.frame(form, data = data)
  y <- qdata[, 1]
  q.W <- q.W
  dontneedw <- !identical(wy, NULL) & !identical(inst, NULL) &
    identical(inst, NULL)
  if (identical(wmat, NULL) & dontneedw == FALSE) {
    if (identical(shpfile, NULL)) {
      stop("Shape file needed")
    }
    library(spdep)
    neighbors <- poly2nb(shpfile, queen = TRUE)
    wmat <- nb2mat(neighbors, zero.policy = TRUE)
    q.W <- nb2listw(neighbors, style = "W")
  }
  n <- nrow(wmat)

  if (identical(wy, NULL)) {
```

```
  wy <- as.numeric(wmat %*% as.matrix(y))
}
qdata <- data.frame(qdata, wy)
newform <- as.formula(form, env = qdata)
xmat <- model.matrix(form, data = data)

if (identical(inst, NULL) & identical(winst, NULL)) {
  zmat <- cbind(xmat, wmat %*% xmat[,-1])
}
zmat <- zmat[, -1]
nk = ncol(xmat) + 1
wyboot <- wy
zboot <- zmat
bootdata <- qdata
nrho = length(rhomat)
rhohat <- NULL
if (nrho > 1) {
  rhohat <- rhomat
  fit <- lm(wy ~ zmat)
  qdata$wyhat <- fitted(fit)
  newform <- as.formula(form, env = qdata)
  newform <- update(newform, newy ~ . + wyhat)
  for (i in seq(1:nrho)) {
    qdata$newy <- y - rhomat[i] * qdata$wy
    fit <- rq(newform, tau = tau, data = qdata)
    rhohat[i] = fit$coef[length(fit$coef)]
  }
  j = which(abs(rhohat) == min(abs(rhohat)))
  if (j == 1 | j == nrho) {
    cat("Warning:  rho is at an endpoint of rhomat",
        "\n")
  }
  minrho = rhomat[j]
```

```
    if (silent == FALSE) {
      cat("Coefficients on instrumental variable for WY:",
          "\n")
      print(cbind(rhomat, rhohat))
    }
    qdata$newy <- y - minrho * qdata$wy
    qdata.lag <- cbind(qdata, lags)
    newform <- update(newform, . ~ . - wyhat + lags)
    xmat.lags <- model.matrix(newform, data = qdata.lag)
    nk <- ncol(xmat.lags) +1
    fit <- rq(newform, tau = tau, data = qdata.lag)
    bmat <- c(fit$coef, minrho)
    semat <- c(rep(NA,ncol(xmat.lags)))
    summat <- array(0, dim = c(nk, 4))
    summat[, 1] <- bmat
    summat[, 2] <- semat
    summat[, 3] <- bmat/semat
    summat[, 4] <- 2 * (1 - pnorm(abs(bmat/semat)))
    rownames(summat) <- c(colnames(xmat.lags), "WY")
    colnames(summat) <- c("Coef.", "Std. Err.",
                          "Z-Values", "Pr(>|z|)")
    if (silent == FALSE) {
      cat("Chernozhukov and Hansen IV Quantile Regression Results",
          "\n")
    }
  }
  return(summat)
}

# rho = 0.1
# same code for every rho
#----------------
# simulation
```

```
set.seed(123)
alpha <- as.vector(runif(1, min = 0, max = 5))
beta <- as.vector(runif(1, min = 0, max = 5))
I <- diag(nrow(wmat))
gamma <- as.vector(runif(1, min = -1, max = 1))

for(i in 1:100){
  set.seed(123*i)
  error <- rnorm(n, mean = 0, sd = 1)
  alpha.cov <- replicate(n,1)
  covariates <- matrix(rnorm(1*n, mean = 10, sd = 1),nrow=n)
  # intercept
  y.durbin <- as.matrix(solve(I - rho*wmat)) %*%
  as.matrix(alpha * alpha.cov + covariates %*% beta +
  wmat %*% covariates %*% gamma + error)

  sdm.sim <- data.frame(cbind(y.durbin, covariates))
  colnames(sdm.sim) <- c("risk", "X1")

  mt <- terms(risk ~ X1, data = sdm.sim)
  mf <- lm(risk ~ X1, data = sdm.sim, method = "model.frame")
  x <- model.matrix(mt,mf)
  lags <- create_WX(x, listw = q.W)
  form <- risk ~ X1

  for(j in 1:9){
    quant.sdm <- qdurbin.u(form,
                           wmat = wmat,
                           data = sdm.sim,
                           tau = j/10,
                           rhomat = seq(0,0.99,0.05),
                           silent = T,
                           q.W = q.W)
```

```
    intercept.sdm.q[i,j] <- quant.sdm[1]
    intercept.sdm.q.e[i,j] <- quant.sdm[1] - quantile(error, probs = j/10)
    beta1.sdm.q[i,j] <- quant.sdm[2]
    beta1.lag.sdm.q[i,j] <- quant.sdm[3]
    rho.for.q[i,j] <- quant.sdm[4]
    se.x1[i,j] <- quant.sdm[2,2]
    se.rho[i,j] <- quant.sdm[4,2]
    y.fit <- as.matrix(solve(I - rho.for.q[i,j]*wmat)) %*%
    (intercept.sdm.q[i,j] + beta1.sdm.q[i,j] * covariates[,1] +
                beta1.lag.sdm.q[i,j] * wmat %*% covariates[,1])
  }
}

# dengue data
#----------------
#----------------------------

data.durbin <- data.final %>% dplyr::mutate(log.risk = log(risk)) %>%
  dplyr::select(-c(risk, dengue, Population))

form <- log.risk ~ Population_density_per_sq_km + Household_well_sanitation_pctg +
  Household_well_drinking_source_pctg + poor_pctg +
  No_of_Flood + submerged_households

res.1 <- qdurbin.u(form,
           wmat = wmat,
           data = data.durbin,
           tau = 0.1,
           rhomat = seq(0,1,0.001),
           q.W = q.W,
           silent = F)
```

```
# bootstrap residuals for tau = 0.2
# same code for every tau
#---------------------

for(k in 1:B){
  e.boot <- sample(res, size = n, replace = T)
  y.boot <- as.matrix(
    solve(I - res.2[14] * wmat) %*%
    (res.2[1] + res.2[2] * x[,2] + res.2[3] * x[,3] +
                        res.2[4] * x[,4] + res.2[5] * x[,5] +
                        res.2[6] * x[,6] + res.5[7] * x[,7] +
                        res.2[8] * wmat %*% x[,2] + res.2[9] * wmat %*% x[,3] +
                        res.2[10] * wmat %*% x[,4] + res.2[11] * wmat %*% x[,5] +
                        res.2[12] * wmat %*% x[,6] + res.2[13] * wmat %*% x[,7])) + e.boot
  x.boot <- data.durbin[,-1]
  dat.boot <- cbind(y.boot, x.boot) %>% as.data.frame()
  form.boot <- y.boot ~ Population_density_per_sq_km + Household_well_sanitation_pctg +
    Household_well_drinking_source_pctg + poor_pctg +
    No_of_Flood + submerged_households
  fit.boot <- qdurbin.u(form.boot,
                        wmat = wmatrix$wmat,
                        data = dat.boot,
                        tau = 0.2,
                        rhomat = seq(0,1,0.001),
                        q.W = q.W,
                        silent = T)
  beta0.boot[k] <- fit.boot[1]
  beta1.boot[k] <- fit.boot[2]
  beta2.boot[k] <- fit.boot[3]
  beta3.boot[k] <- fit.boot[4]
  beta4.boot[k] <- fit.boot[5]
  beta5.boot[k] <- fit.boot[6]
  beta6.boot[k] <- fit.boot[7]
```

```
  lag.beta1.boot[k] <- fit.boot[8]
  lag.beta2.boot[k] <- fit.boot[9]
  lag.beta3.boot[k] <- fit.boot[10]
  lag.beta4.boot[k] <- fit.boot[11]
  lag.beta5.boot[k] <- fit.boot[12]
  lag.beta6.boot[k] <- fit.boot[13]
  rho.boot[k] <- fit.boot[14]
}

for(i in 1:9){
  SE.beta0.boot <- sd(beta0.boot)
  SE.beta1.boot <- sd(beta1.boot)
  SE.beta2.boot <- sd(beta2.boot)
  SE.beta3.boot <- sd(beta3.boot)
  SE.beta4.boot <- sd(beta4.boot)
  SE.beta5.boot <- sd(beta5.boot)
  SE.beta6.boot <- sd(beta6.boot)
  SE.lag.beta1.boot <- sd(lag.beta1.boot)
  SE.lag.beta2.boot <- sd(lag.beta2.boot)
  SE.lag.beta3.boot<- sd(lag.beta3.boot)
  SE.lag.beta4.boot <- sd(lag.beta4.boot)
  SE.lag.beta5.boot <- sd(lag.beta5.boot)
  SE.lag.beta6.boot <- sd(lag.beta6.boot)
  SE.rho.boot <- sd(rho.boot)
}

summat.2 <- array(0, dim = c(nrow(res.5), 4))
summat.2[, 1] <- res.2[,1]
summat.2[, 2] <- rbind(SE.beta0.boot, SE.beta1.boot, SE.beta2.boot,
                    SE.beta3.boot, SE.beta4.boot,
                       SE.beta5.boot, SE.beta6.boot,
                       SE.lag.beta1.boot, SE.lag.beta2.boot,
                    SE.lag.beta3.boot, SE.lag.beta4.boot,
```

```
                        SE.lag.beta5.boot, SE.lag.beta6.boot,
                        SE.rho.boot)
summat.2[, 3] <- summat.2[,1]/summat.2[,2]
summat.2[, 4] <- 2 * (1 - pnorm(abs(summat.2[,1]/summat.2[,2])))
rownames(summat.2) <- c(rownames(res.2))
colnames(summat.2) <- c("Coef.", "Std. Err.",
                         "Z-Values", "Pr(>|z|)")

# impact matrix
#------------------
DE <- NULL
IE <- NULL
TE <- NULL
vec <- as.vector(rep(1, nrow(data.durbin)))
i <- 1

for(i in 1:n.beta){
  IqW.inv <- as.matrix(solve(diag(nrow(data.durbin)) - res.2[nrow(res.2)]*wmat.du))
  sec.eq <- as.matrix(res.2)[i+1] * diag(nrow(data.durbin)) + wmat.du *
  as.matrix(res.2)[(n.beta+i+1)]
  SkW <- IqW.inv %*% sec.eq
  DE[i] <- 1/nrow(data.durbin) * sum(diag(SkW))
  TE[i] <- 1/nrow(data.durbin) * sum(rowSums((SkW)))
  IE[i] <- TE[i] - DE[i]
  summat[i,1] <- DE[i]
  summat[i,2] <- IE[i]
  summat[i,3] <- TE[i]
  rownames(summat) <- c(as.matrix(row.names(res.2[2:(n.beta+1),])))
  colnames(summat) <- c("Direct", "Indirect", "Total")
}
```