A bayesian spatial autoregressive model with k-NN optimization for modeling the learning outcome of the junior high schools in West Java

Jaya, Mindra; Toharudin, Toni; Abdullah, Atje Setiawan

Published in:
Model Assisted Statistics and Applications

DOI:
10.3233/MAS-180435

Publication date:
2018

Citation for published version (APA):
Jaya, M., Toharudin, T., & Abdullah, A. S. (2018). A bayesian spatial autoregressive model with k-NN optimization for modeling the learning outcome of the junior high schools in West Java. *Model Assisted Statistics and Applications*, *13*(3), 207-219. https://doi.org/10.3233/MAS-180435

Model Assisted Statistics and Applications 13 (2018) 207–219
DOI 10.3233/MAS-180435
IOS Press

A bayesian spatial autoregressive model with k-NN optimization for modeling the learning outcome of the junior high schools in West Java

I. Gede Nyoman Mindra Jaya[a,b,*], Toni Toharudin[b] and Atje Setiawan Abdullah[c]
[a] *Faculty Spatial Science, University of Groningen, Netherlands*
[b] *Statistics Department, Universitas Padjadjaran, Bandung, Indonesia*
[c] *Computer Science Department, Universitas Padjadjaran, Bandung, Indonesia*

Abstract. Increasing the human capital development index of Indonesia is needed to realize the country's dream to become a developed country in the world. Quality education is needed for that purpose, and this should start from an early age. School is a formal institution for knowledge transfer, which is very useful in building the quality of an Indonesian's character. Since 2000, Indonesia has made enormous effort to improve the quality of education, which is measured by increased learning outcome, which is measured by mean national examination score. Indonesia has focused on three major aspects, namely, improving equity and access, enhancing quality and relevance, and strengthening management and accountability. These three aspects are translated into eight standards accreditation score. Education quality is believed to have spatial characteristics that follow the Tobler law. In general, schools close to each other, especially in one administrative area, have the same quality characteristics. The spatial characteristics need to be included in modeling the national examination score. Because of the normality assumption problem, we use a Bayesian spatial autoregressive model (BSAR) to evaluate the effect of the eight standard school qualities on learning outcomes and use k-nearest neighbors (k-NN) optimization in defining the spatial structure dependence. We use junior high schools data in Wes Java. West Java is one of the largest provinces in Indonesia with the highest number of junior schools. The result shows that the national examination score of the junior high schools in West Java is significantly influenced by the standard of graduate competence, and the standard of assessment. We found that the spatial effect also significant which means the average of the national examination score of the nearest schools influences the national examination of the junior high observed.

Keywords: Bayesian, BSAR, k-NN, learning outcome, ML

JEL: C30, I18

1. Introduction

Indonesia is one of the developing countries in Southeast Asia and has the ambition to be one of the developed countries in the world. For that purpose, human capital development is crucial. Despite the high regional diversity, the quality of education throughout the territory of Indonesia has always been a top priority. According to the Indonesian education system, the grades of school education consist of elementary, junior, and senior high school levels (OECD, 2015; Firman & Tola, 2008). Since 2000, Indonesia has made enormous efforts to improve the quality of

[*] Corresponding author: I. Gede Nyoman Mindra Jaya, Statistics Department, Universitas Padjadjaran, Bandung, Indonesia. E-mail: mindra@unpad.ac.id.

education, which is measured by increased literacy (Tobias et al., 2014). To develop the quality of education, Indonesia has focused on three major aspects, namely, improving equity and access, enhancing quality and relevance, and strengthening management and accountability. These three aspects are translated into eight standard accreditations score, which include standard of contents (Q1), standard of process (Q2), standard of graduate competence (Q3), standard of education (Q4), standard of facilities and infrastructure (Q5), standard of management (Q6), standard of financing (Q7), and standard of assessment (Q8). Improvement of these standards is expected to improve the human capital development index, which, for schools relate to the learning outcome and can be measured by the national examination score. To map the quality of education, in 2000, the government reestablished the national examination that was once abolished.

The national examination (*Ujian Nasional-UN*) is an assessment learning outcome by the government, which has become the benchmark of student success in following the process of learning in school. The purpose of the UN is to know and measure students' level of mastery of a certain subject matter nationally and the quality level of education in every province in Indonesia (Elfiza et al., 2016). The national examination score is a value generated from the national examinations held nationally at the final levels of elementary, junior, and senior high schools. The high level of diversity of the national examination score among junior high schools in Indonesia indicates that the quality of education is not the same in different parts of the country and needs to be improved. Improving the quality of education in Indonesia is a big challenge because it is a vast and diverse country that has the fourth largest population in the world, comprising 34 provinces and over 500 districts with roughly 55 million students, 3 million teachers, and 236,000 schools (MoEC, 2013).

West Java is one of the largest provinces in Indonesia with the highest number of junior high schools. However, a lot of schools still have a medium level of accreditation status. The West Java government has taken great efforts to improve the quality of education; however, these efforts need to be defined accurately. Policy makers should be more focused on the weakest indicators of education quality. Improving the quality indicators of education will increase the national examination score and the percentage of students graduating. A better understanding of the effect of school characteristics, which are measured by the eight standards of quality on learning outcomes, is important because the Ministry of Education and Culture (MoEC) may use this information to decide what policy should be taken to improve learning outcomes. As a first step toward understanding the determinants of learning outcomes in West Java and also Indonesia, this paper focuses on how the eight standards of quality in junior high school influences their academic achievement, which is measured based on the national examination score (Newhouse & Beegle, 2005) in West Java. For that purpose, we build a model that can identify which indicators are most dominant in influencing the national examination score, which represents learning outcomes.

The education quality is believed to have spatial characteristics following the Tobler law. Tobler (1970) introduced the first law of geography; that is, everything is related to everything else, but near things are more related than distant ones. These spatial characteristics are usually measured by means spatial dependence and spatial heterogeneity (Jaya et al., 2018; Elhorst, 2014). In general, schools close to each other, especially in one administrative area, have the same quality characteristics so that modeling the national examination score needs to include the spatial characteristics for reliable estimation (Anselin, 2003; Elhorst, 2014). The spatial econometrics model is a powerful model that can be used to accommodate the spatial characteristics especially for spatial dependence and explain the spillover (i.e., the effect of variable changes at one location on the outcome variables in other locations).

Spatial econometrics are widely used in regional, economic, social, and epidemiological fields, among others (Jaya et al., 2017; Klotz, 2004). There are several models in spatial econometrics (i.e., spatial autoregressive model, spatial error model, and spatial Durbin model). These models are developed based on the structure of spatial dependencies and heterogeneities (Vega & Elhorst, 2015). Based on the Tobler law, we believe that the national examination score has spatial dependence characteristics that can be patterned after the spatial autoregressive model (SAR). The parameters of the spatial econometrics model can be estimated by the mean of the maximum likelihood estimation (Anselin, 1988). However, for large sample size, the maximum likelihood is not a better choice because the complexity in calculating the standard error estimates for hypothesis testing (LeSage & Pace, 2009), the purpose, and the strict assumption of the normality distribution is needed. To overcome these problems, we introduce an alternative using a Bayesian approach. A Bayesian approach has several advantages. It provides a convenient way of combining prior information with data using the appropriate statistical framework. Bayesian inference is conditional on the data and is exact without dependence on asymptotic approximation so that the normality distribution of the error term is not an issue. For practitioners, the result of the Bayesian estimation is more interpretable and easier to understand. It provides understandable interpretations, such as “the true parameter θ has a probability of 0.95 of falling in a 95% credible interval,” and it can be used as an approximation for hypothesis testing. But the Bayesian approach also has a disadvantage. It often comes with a high computational cost (SAS, 2011).

Besides the sample size and normality assumption problem, how to present the spatial dependence structure in spatial econometrics is still a big challenge. The spatial dependence structure is represented as a spatial weight matrix (W). Several methods have been introduced to create a spatial weight matrix, including contiguity, distance, and geostatistical methods. The errors in defining the spatial weight matrix may give misleading results, so we need to be careful. However, no best procedure and criteria have been introduced to define the best spatial structure. This issue is still an open research area. In practice, we usually define W based on a subjective argument. Here, we introduced the optimization procedure based on the k-NN algorithm. The idea is to create a spatial weight matrix that gives the largest spatial autocorrelation coefficient. We assume that the spatial autocorrelation has to be evaluated based on the appropriate spatial structure. We use Moran’s index statistics to measure the spatial autocorrelation. Moran’s index is a standard method that is commonly used and easy to understand. The coefficient of Moran’s index is between -1 and 1. Moran’s index close to 1 means that the determined variable has strong spatial autocorrelation.

This paper focuses on the application optimization of the k-NN approach and Bayesian approach for modeling and testing the effect of the eight quality standards on the national examination score for junior high schools in West Java. The structure of the paper is organized as follows: In Section 2, we discuss the methods that are used in the paper, including the k-NN, Moran’s index, and Bayesian approach. The result is accomplished in Section 3. The paper is closed with a discussion and conclusion in Section 4.

2. Method

2.1. k-Nearest Neighbors algorithm (k-NN)

In pattern recognition, several techniques are usually used (i.e., k-NN, k-Mean, naïve Bayes). k-NN is one of the famous nonparametric methods used for classification (Altman, 1992). It uses a simple algorithm and does not need assumption of distribution but works effectively in practice. The k-NN algorithm has a different behavior based on k. Amaratunga and Cabrera (2003) presented a formal description of k-NN. Let y_j represent the jth training sample and x_j be the class label for y_j. The objective of the k-NN algorithm is to calculate the probability that y categorized to the i-th class by the proportion of k nearest neighbors that belong to the i-th class (Zhang, 2006):

$$\hat{p}(i|y) = \frac{|\{x_j = i|y_j \in N_y\}|}{k} \tag{1}$$

The label class of y is based on the i-th class, which maximizes the probability $\hat{p}(i|y)$.

2.2. Moran’s index

In spatial data analysis, Moran’s index is used to measure the spatial autocorrelation, which detects whether a determined variable has a spatially dependent structure or occurs in a random pattern. The index takes the values of -1 and 1. The large positive value indicates that the determined variable has a strong positive spatial autocorrelation. It indicates the observations that are spatially close have almost similar values. The Moran’s index formulation can be written as (Shekhar & Xiong, 2008)

$$I = \frac{\sum_i \sum_{j\neq i} w_{ij} \left(y_i - \bar{y}\right)\left(y_j - \bar{y}\right)}{S^2 \sum_i \sum_{j\neq i} w_{ij}} \tag{2}$$

where y_i denotes the observed value at area i, $\bar{y}$ is the mean of the y variable over the n areas, and

$$S^2 = \frac{1}{n}\sum_{i}^{n} \left(y_i - \bar{y}\right)^2 \tag{3}$$

2.3. Spatial econometrics model

Spatial econometrics is a subfield of econometrics dealing with spatial effects among geographical units. Its methods were developed to accommodate the spatial interaction in the determined variable related to spatial locations (Anselin, 1988). Several models have been introduced, namely, spatial autoregressive model (SAR), spatial error model (SEM), spatial autoregressive combined model (SAC), spatial Durbin model (SDM), spatial error Durbin model (SEDM), spatial lag exogenous model (SLX), and spatial general nesting model (SGNS). The SAR model is widely applied for spatial lattice data (Li et al., 2007).

2.3.1. Spatial autoregressive model (SAR)

The spatial autoregressive (SAR) model can be presented as (Anselin, Spatial Econometrics: Methods and Models, 1988)

$$y_i = \rho \sum_{j=1}^{n} w_{ij} y_j + \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon_i, \tag{4}$$

where y_i denotes the national examination score for school i-th and denotes the autoregressive spatial coefficient. The autoregressive spatial coefficient represents the magnitude of the effect of the average national examination score of the neighboring junior high school on the national examination scores of the observed junior high school. Parameter models β_0 and β_k denote the coefficient of the intercept and regression slope for the Kth exogenous variable, x_{ik} denotes the value of the Kth exogenous variable at school ith, and ε_i denotes a random error with an independent identical assumption of the normal distribution with a zero mean and a variance of $\sigma^2(\varepsilon_i \sim_{iid} N\left(0,\sigma^2\right)$. The w_{ij} component is an element of a spatial weight matrix that can be determined based on the school's juncture or distance between schools and through optimization methods. This research uses a spatial weight matrix based on the optimization method of k-NN. Several methods can be used to estimate SAR model parameters, such as variable (IV) instrument method, maximum likelihood (ML), and Bayesian method (LeSage & Pace, 2009).

2.3.2. Bayesian spatial autoregressive (BSAR)

Using Bayesian methods through the definition of a prior distribution can solve incomplete information obtained from the data (Congdon, 2013). The Bayesian method gives good results for small sample sizes compared with the ML method and similar results for a large sample size. ML gives unfavorable results for a small sample size because it is difficult to get the optimum value of autoregressive parameters. However, for large sample size, the ML estimation is difficult to use to find the standard error estimate because of the large size of the Hessian matrix. The Bayesian method is well suited in cases of nonnormality, and homoscedasticity assumptions of the error term are violated. Those conditions are very common in spatial data analysis (Anselin, 1988). Equation (4) can be written in the matrix notation as follows:

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\beta + \varepsilon \tag{5}$$

where **y** is the vector of the national examination score of the size ($n \times 1$), is an autoregressive parameter, W is the spatial weight matrix ($n \times n$), **X** is a $n \times p$-sized design matrix including the unit vector, $\beta = [\beta_0, \beta_1, ..., \beta_\mathrm{K}]^\mathrm{T}$ sized ($p \times 1$) is a regression parameter including an intercept β_0 with $p = K + 1$, and ε is a vector of the error component ($n \times 1$) with $\varepsilon \sim_{iid} \left(0, \sigma^2 I\right)$. The Bayesian approach assumes that the parameters β and σ^2 are random variates following the normal and inverse gamma distributions, respectively, and ρ is a random variate following the uniform distribution. The joint posterior distribution formulation is as follows (LeSage & Pace, 2009):

$$p\left(\beta, \sigma^2, \rho | D\right) = \frac{p\left(D|\beta, \sigma^2, \rho\right)\pi\left(\beta, \sigma^2|\rho\right)\pi\left(\rho\right)}{p\left(D\right)} \tag{6}$$

where $D = \{y, X\}$. To obtain the posterior distribution in Eq. (6), it is necessary to determine the likelihood function and priors as follows:

1. The likelihood function $p\left(D|\beta, \sigma^2, \rho\right)$

   By assumption of the learning, outcome follows a normal distribution, the likelihood function can be written as

$$p\left(D|\beta,\rho,\sigma^2\right)=\left(2\pi\sigma^2\right)^{-\frac{n}{2}}|A|\exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{Ay}-\mathbf{X}\beta\right)^{\mathrm{T}}\left(\mathbf{Ay}-\mathbf{X}\beta\right)\right) \tag{7}$$

with $\mathbf{A}=(\mathbf{I}-\rho\mathbf{W})$, and $|\mathbf{A}|$ denotes the determinant matrix of $\mathbf{A}$.

2. Prior distribution, $\pi(\beta|\sigma^2)$ follows normal distribution

$$\begin{aligned}\pi(\beta|\sigma^2)&\sim\mathrm{N}\left(\mathbf{c},\sigma^2\mathbf{T}\right)\\&=\frac{1}{(2\pi)^{\mathrm{p}/2}\left(\sigma^2\right)^{\mathrm{p}/2}|\mathbf{T}|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\beta-\mathbf{c}\right)^{\mathrm{T}}\mathbf{T}^{-1}\left(\beta-\mathbf{c}\right)\right)\end{aligned} \tag{8}$$

3. Prior distribution, $\pi\left(\sigma^2\right)$ follows inverse gamma distribution

$$\begin{aligned}\pi\left(\sigma^2\right)&\sim\mathrm{IG}\left(a,b\right)\\&=\frac{b^a}{\Gamma\left(a\right)}\left(\sigma^2\right)^{-(a+1)}\exp\left(-b/\sigma^2\right)\end{aligned} \tag{9}$$

4. Joint prior distribution, $\pi\left(\beta,\sigma^2\right)$ follows normal inverse gamma.
   Joint prior distribution $\pi\left(\beta,\sigma^2\right)$ is obtained Eq. (8) from Eq. (9)

$$\begin{aligned}\pi\left(\beta,\sigma^2\right)&\sim\mathrm{NIG}\left(\mathbf{c},\mathbf{T},a,b\right)\\&=\pi(\beta|\sigma^2)\pi\left(\sigma^2\right)\\&=\mathrm{N}\left(\mathbf{c},\sigma^2\mathbf{T}\right)\times\mathrm{IG}\left(a,b\right)\\&=\frac{1}{(2\pi)^{\mathrm{p}/2}\left(\sigma^2\right)^{\mathrm{p}/2}|\mathbf{T}|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\beta-\mathbf{c}\right)^{\mathrm{T}}\mathbf{T}^{-1}\left(\beta-\mathbf{c}\right)\right)\\&\quad\times\frac{b^a}{\Gamma\left(a\right)}\left(\sigma^2\right)^{-(a+1)}\exp\left(-b/\sigma^2\right)\\&=\frac{b^a}{(2\pi)^{\mathrm{p}/2}|\mathbf{T}|^{1/2}\Gamma\left(a\right)}\left(\sigma^2\right)^{-\left(a+\left(\frac{\mathrm{p}}{2}\right)+1\right)}\\&\quad\times\exp\left(-\frac{1}{2\sigma^2}\left[\left(\beta-\mathbf{c}\right)^{\mathrm{T}}\mathbf{T}^{-1}\left(\beta-\mathbf{c}\right)+2b\right]\right)\end{aligned} \tag{10}$$

5. Prior distribution $\pi\left(\rho\right)$

$$\begin{aligned}\left(\rho\right)&\sim\mathrm{U}\left(\lambda_{\min}^{-1},\lambda_{\max}^{-1}\right)\\&=\frac{1}{\lambda_{\max}^{-1}-\lambda_{\min}^{-1}}\end{aligned} \tag{11}$$

Based on the likelihood function and prior distribution above, the joint posterior distribution is as follows:

$$\begin{aligned}p\left(\beta,\rho,\sigma^2|D\right)&=\left(2\pi\sigma^2\right)^{-\frac{n}{2}}|A|\exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{Ay}-\mathbf{X}\beta\right)^{\mathrm{T}}\left(\mathbf{Ay}-\mathbf{X}\beta\right)\right)\\&\quad\times\frac{b^a}{(2\pi)^{\frac{\mathrm{p}}{2}}|\mathrm{T}|^{\frac{1}{2}}\Gamma\left(a\right)}\left(\sigma^2\right)^{-\left(a+\left(\frac{\mathrm{p}}{2}\right)+1\right)}\times\exp\left(-\frac{1}{2\sigma^2}\left[\left(\beta-\mathbf{c}\right)^{\mathrm{T}}\mathbf{T}^{-1}\left(\beta-\mathbf{c}\right)+2b\right]\right)\\&\quad\times\left(\frac{1}{\lambda_{\max}^{-1}-\lambda_{\min}^{-1}}\right)\\&\propto\left(\sigma^2\right)^{-\left(a+\frac{n+p}{2}+1\right)}|A|\\&\quad\times\exp\left(-\frac{1}{2\sigma^2}\left[\left(\mathbf{Ay}-\mathbf{X}\beta\right)^{\mathrm{T}}\left(\mathbf{Ay}-\mathbf{X}\beta\right)+\left(\beta-\mathbf{c}\right)^{\mathrm{T}}\mathbf{T}^{-1}\left(\beta-\mathbf{c}\right)+2b\right]\right)\end{aligned} \tag{12}$$

The combined joint posterior distribution Eq. (12) can be written in the following form:

$$p\left(\beta,\rho,\sigma^2|D\right)=\propto\left(\sigma^2\right)^{-\left(a^*+1\right)}|\mathbf{A}|\times\exp\left(-\frac{1}{2\sigma^2}\left[2b^*+\left(\beta-c^*\right)^T\left(T^*\right)^{-1}\left(\beta-c^*\right)\right]\right) \tag{13}$$

where

$$\mathbf{T}^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{T}^{-1}\right)^{-1} \tag{14}$$

$$\mathbf{c}^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{T}^{-1}\right)^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{y} + \mathbf{T}^{-1}\mathbf{c}\right) \tag{15}$$

$$b^* = b + \left(\mathbf{c}^{\mathbf{T}}\mathbf{T}^{-1}\mathbf{c} + \mathbf{y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{y}) - (\mathbf{c}^*)^{\mathrm{T}}(\mathbf{T}^*)^{-1}\mathbf{c}^*\right)/2 \tag{16}$$

$$a^* = a + \frac{n+p}{2} \tag{17}$$

The parameters (β, σ, ρ) can be estimated by the combination of the Gibbs and Metropolis Hasting sampling methods through the sequential sampling method from the conditional posterior distribution of the parameters (β, σ, ρ).

Conditional distribution

1. Conditional distribution of $\beta|\rho$, σ^2

$$p\left(\beta|\rho, \sigma^2_{(0)}\right) \sim \mathrm{N}\left(c^*, \sigma^2_{(0)}\mathbf{T}^*\right) \tag{18}$$

$\beta|\rho$, σ^2 follows a multivariate normal distribution.

2. Conditional distribution of $\sigma^2|\rho, \beta$:

$$p\left(\sigma^2|\beta_{(1),}\rho\right) \sim \mathrm{IG}\left(a^*, b^*\right) \tag{19}$$

3. Conditional distribution of $\rho|\beta_{(1),}, \sigma^2_{(1)}$:

$$\begin{aligned} p\left(\rho|\beta, \sigma^2\right) &= \frac{p\left(\rho, \beta, \sigma^2|\mathbf{D}\right)}{p\left(\beta, \sigma^2|\mathbf{D}\right)} \\ &\propto p\left(\rho, \beta, \sigma^2|\mathbf{D}\right) \\ &\propto |\mathbf{I}_{\mathrm{n}} - \rho\mathbf{W}_\rho| \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{A}\mathbf{y} - \mathbf{X}\beta\right)^{\mathrm{T}}\left(\mathbf{A}\mathbf{y} - \mathbf{X}\beta\right)\right) \end{aligned} \tag{20}$$

Conditional distributions $\rho|\beta, \sigma^2$ do not follow standard distribution forms such as normal distribution, gamma, or the other distributions. The estimation is done by an MCMC procedure, which is a combination of the Gibbs sampling and the metropolis methods.

Computational algorithm

To get an estimation of SAR model parameters, calculate with the algorithm as follows:

1. Define the initial values (*initial value*) { $\beta_{(0)}$, $\rho_{(0)}$, $\sigma^2_{(0)}$ }.
2. Generate parameter $\beta_{(l)} \sim \mathrm{N}\left(\mathbf{c}^*, \sigma^2_{(0)}\mathbf{T}^*\right)$.
3. Generate parameter $\sigma^2_{(l)} \sim \mathrm{IG}\left(a^*, b^*\right)$.
4. Generate parameter $\rho_{(l)}$.
   (a) Define the candidate distribution ρ^* (i.e., the normal distribution).
   (b) Do the sampling process:

$$\psi_H\left(\rho^{\mathrm{c}}, \rho^*\right) = \min\left[1, \frac{p\left(\rho^*|\boldsymbol{\beta}, \sigma\right)}{p\left(\rho^{\mathrm{c}}|\boldsymbol{\beta}, \sigma\right)}\right] \tag{21}$$

with $\rho^* = \rho c + d \cdot N\left(0, 1\right)$, where d is the tuning parameter specified by the acceptance value.

5. Do the 2–4 process as much as M iteration. Generally, M = 100,000.

All the computational processes are done in R software with our own packages. The R code is available upon request

2.3.3. Convergence diagnostic

The parameter inference in a Bayesian method is based on the sequence of MCMC samples that are derived from the true posterior distribution. Validity inference needs to be checked using convergence diagnostic. The convergence

Table 1
Statistics of the national examination score of junior high school in West Java

Statistics	National examination score
Minimum	17.30
1st quartile	21.01
Median	21.58
Mean	23.28
3rd quartile	24.57
Maximum	35.55

Table 2
Statistics of the eight standards accreditation score of junior high school in West Java

Statistics	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Minimum	30.77	26.67	48.75	40.63	50.65	52.87	58.11	59.02
1st quartile	84.90	82.26	77.50	76.39	78.90	83.61	87.84	84.02
Median	90.63	88.71	87.50	84.72	88.64	90.57	93.58	89.75
Mean	89.12	87.19	85.42	81.49	85.34	88.57	91.89	88.33
3rd quartile	94.79	93.55	93.75	89.75	93.18	95.08	97.64	93.85
Maximum	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

diagnostic relates to determining the minimum sample required to ensure a reasonable approximation to the target posterior density. The graphical approach usually used for convergence diagnostic includes trace plot, ergodic mean plot, and autocorrelation plot. The convergence of the algorithm is shown by the stabilizers of those plots after some iterations (see Ioannis, 2009, for detail).

2.3.4. Spillover effect

Spillover effects are mainly interesting in spatial econometrics. It explains a change in covariate in a particular region potentially impacting the outcome in all regions that, according to the spatial weight matrix (W), are unconnected. The spillover effect of the x_j variable can be formulated using the average of the diagonal of $(I_n - W)^{-1}\beta_k$.

3. Results

We use educational data from 2012. These data are complete to be able to support the analysis of the quality of education in junior high schools in West Java, Indonesia.

The national examination score is obtained from four subjects: English, Indonesian, Mathematics, and Natural Sciences. The minimum national examination score from 415 schools is 17.30, and the maximum score is 35.55. The Fig. 1 below shows the distribution of the national examination score for 415 schools in West Java.

In 2012, the majority of junior high schools in West Java got an average of national examination scores that was less than 25. It means that the average scores in four subjects were lower than 6. Figure 1 shows that there is spatial clustering of the national examination score, where schools with a low score become one group and high-score schools become another group.

Table 2 shows that the score for standard of contents (Q1), standard of process (Q2), standard of graduate competence (Q3), standard of education (Q4), standard of facilities and infrastructure (Q5), standard of management (Q6), standard of financing (Q7), and standard of assessment (Q8) varies from very small and large value (26.67 to 100). This means that the quality of education in junior high schools in West Java have large variation.

Figure 2 shows the distribution of the eight standard accreditation score of junior high school in West Java Map. We can see that the map present the spatial clustering. The map informs that the schools have the similar quality will close each other. The hypothesis in this research is that the standard quality education, which is measured by the eight standard scores, influences the national examination score. For this purpose, we develop a spatial autoregressive model (SAR). The SAR model is built based on 415 junior high schools in West Java. In the SAR model, constructing the spatial weight matrix (W) is still the major issue besides the estimation and the inference problems. Here, we introduced the optimization procedure based on the k-NN algorithm. The idea is to create a spatial weight matrix that gives the largest spatial autocorrelation coefficient. We assume that the spatial autocorrelation has to be evaluated

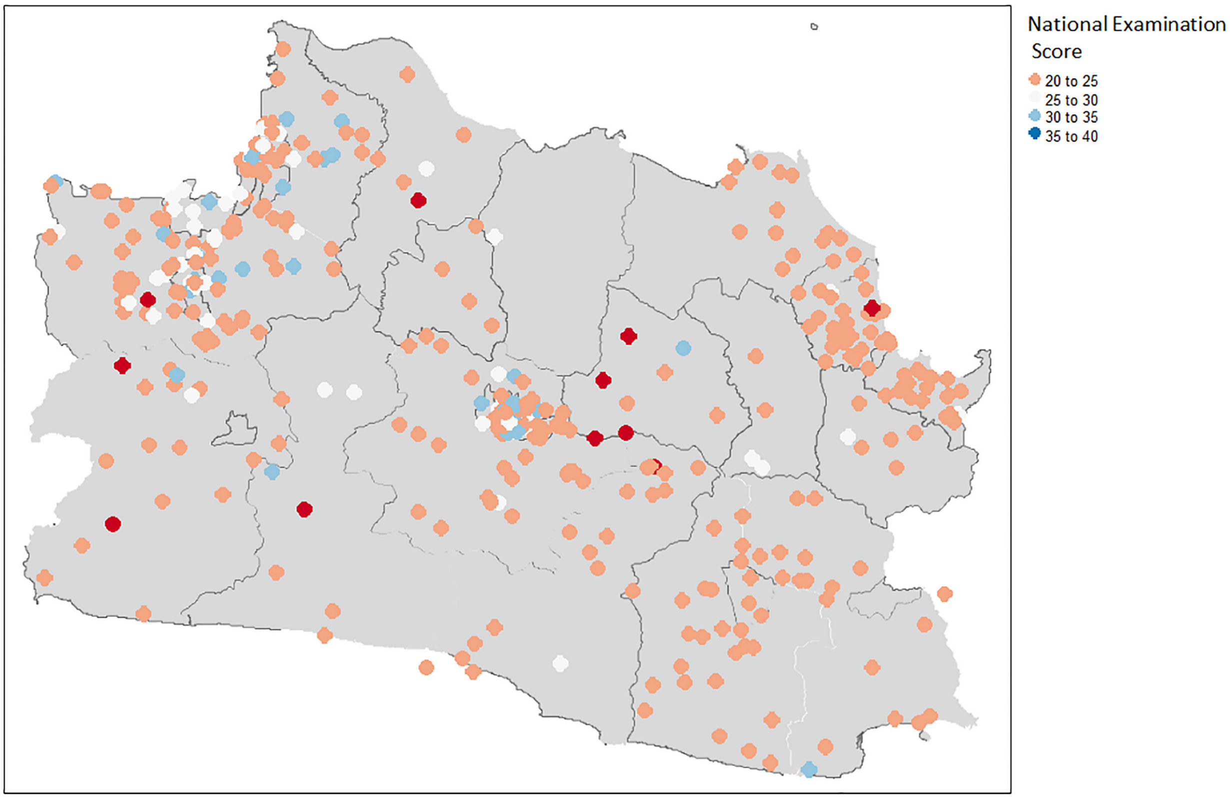


Fig. 1. The distribution of the national examination score of junior high schools in West Java.

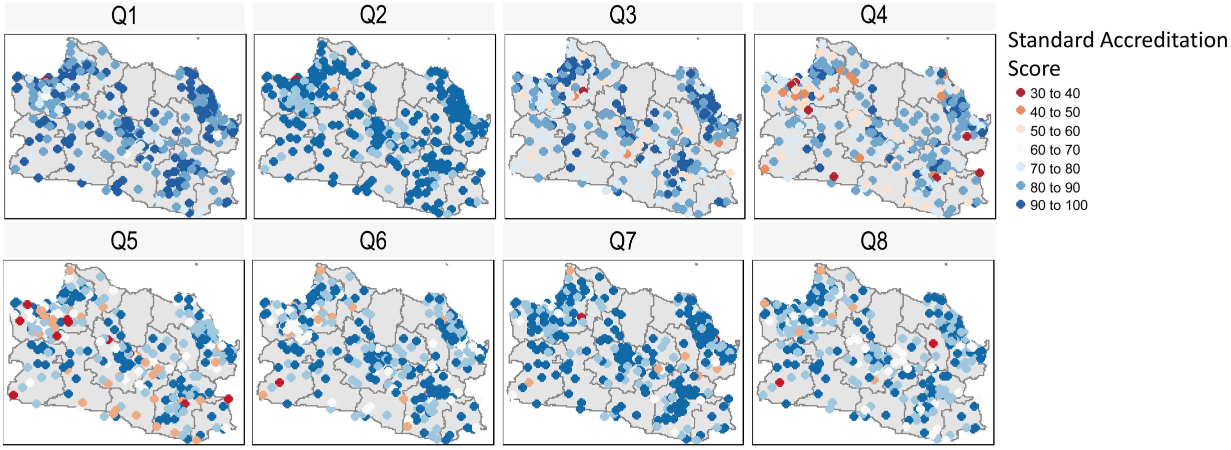


Fig. 2. The distribution of the eight standard accreditation scores of junior high school in West Java.

based on the appropriate spatial structure. We use Moran's index statistics to measure the spatial autocorrelation. We have tried to simulate the eight different values of k $(k = 1, 2, \ldots, 8)$ and find the optimum k based on the Moran's index statistics. The optimum k is the k with the highest Moran's index value. The simulation is presented in Fig. 3 below.

Figure 3 shows the maximum value of Moran's index that is obtained form $k = 1$. It means that the national examination scores for one adjacent school are relatively similar. This phenomenon may be because adjacent schools, especially in one administrative district, get equal treatment from the local government so that they have the same

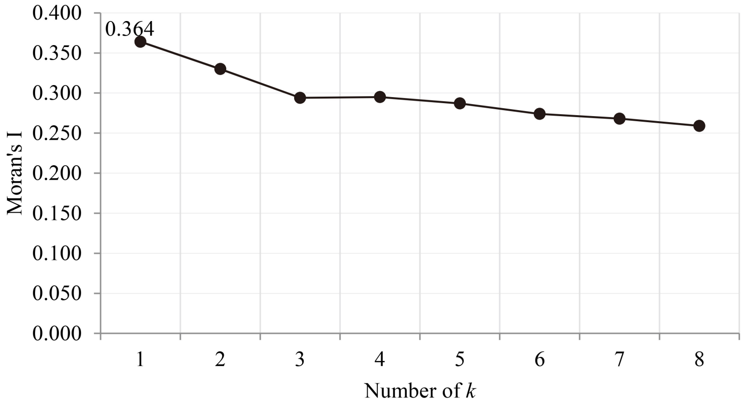
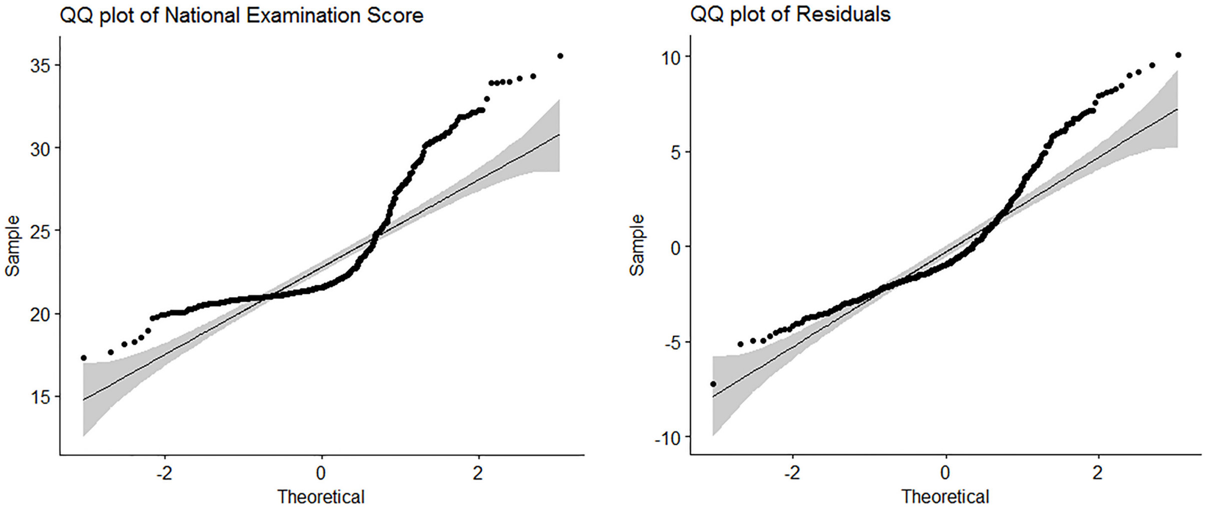


Fig. 3. The simulation to find the optimum k.



(a)

(b)

Fig. 4. Testing of normality assumption. (a) Normality testing for response variable (y); (b) Normality testing for errors (ε).

relative quality. Based on this result, we decide to create a spatial weight matrix (W) based on k-NN with $k = 1$. We need to present the statistical reason why we use the Bayesian approach instead of the maximum likelihood estimation to estimate the parameters of the SAR model. Here, we present that the normality assumption of the response variable (y) and the residuals (ε) for ML estimation are violated.

The maximum likelihood estimation strictly needs the determined response variable and the error term following the normal distribution. If the normality assumption is violated, the estimation result may be invalid and the hypothesis testing with t-student or F-test may be misleading (Harvey, 1990). We use QQ plot, and Shapiro-Wilk normality test to check the normality assumption of the response variable and the error term (Rees, 2001). Figure 4a and b shows QQ plot with the Shapiro-Wilk normality tests for national examination score (y) and Error (ε). The results present that the national examination score and also the error term do not follow a normal distribution. QQ plots show the deviations of theoretical (i.e., normal) and empirical distributions. The Shapiro-Wilk normality tests present those p-values less than 0.05 which implies a null hypothesis that the data and error term follow normal distribution are rejected. Based on this result, the Bayesian approach is the alternative method for estimating and inferencing the effect of the eight quality standards on the national examination score in junior high schools in West Java. Here, we use a combination of the Gibbs sampling and the Metropolis Hasting algorithm to estimate and test the parameters of SAR. We present the convergence diagnostics to evaluate the inference validity of MCMC sampling based on a 100,000 iteration sample, and burn in sampling takes 1,000 samples.

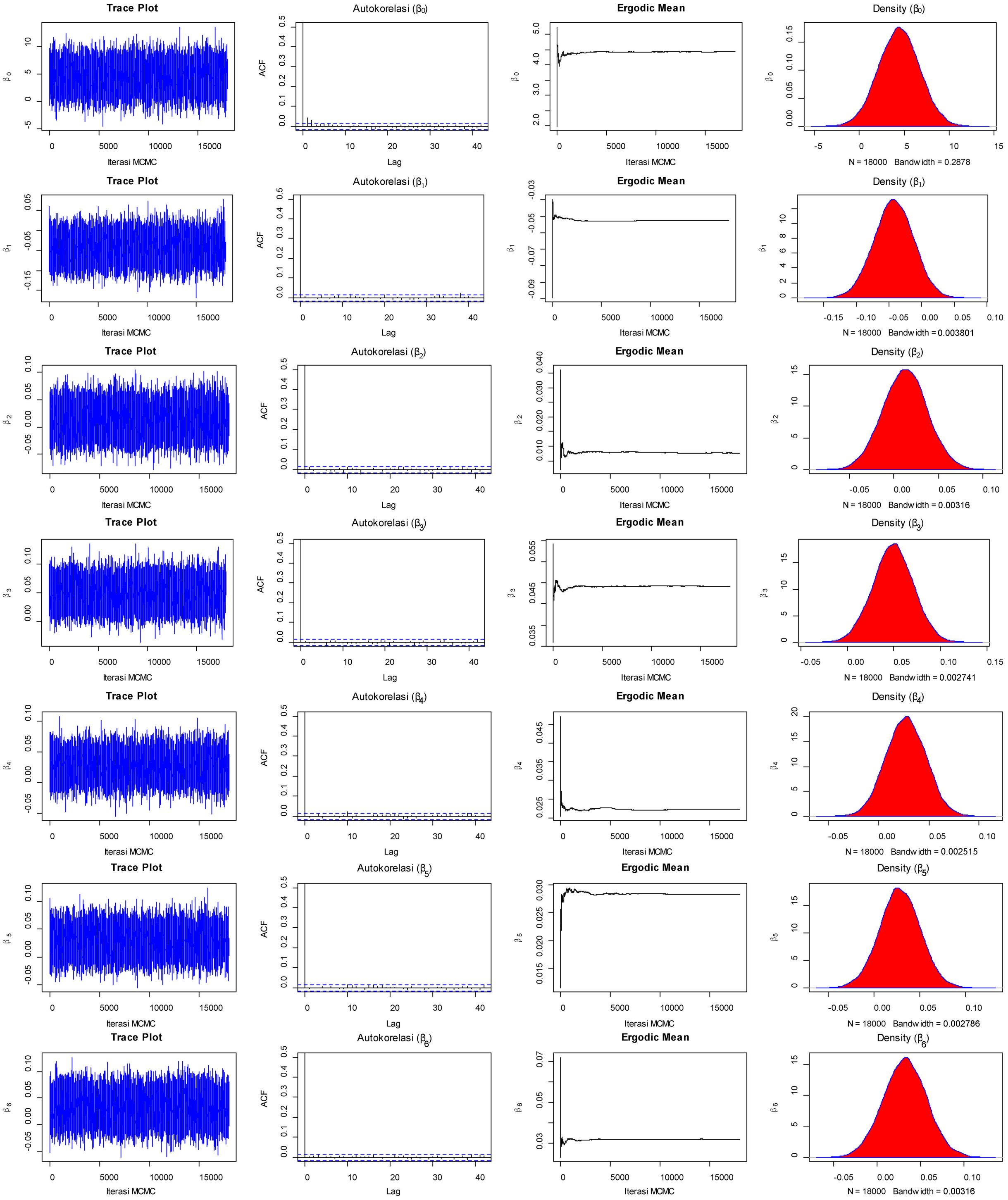


Fig. 5. MCMC of convergence diagnostics.

Figure 5 shows four different MCMC convergence diagnostic plots. The first column of the plots includes trace plots. Trace plots show the values that the parameter took during the runtime of the chain. The convergence patterns of all the trace plots show the ideal pattern. All trace plots present rapid up-and-down variation with no long-term

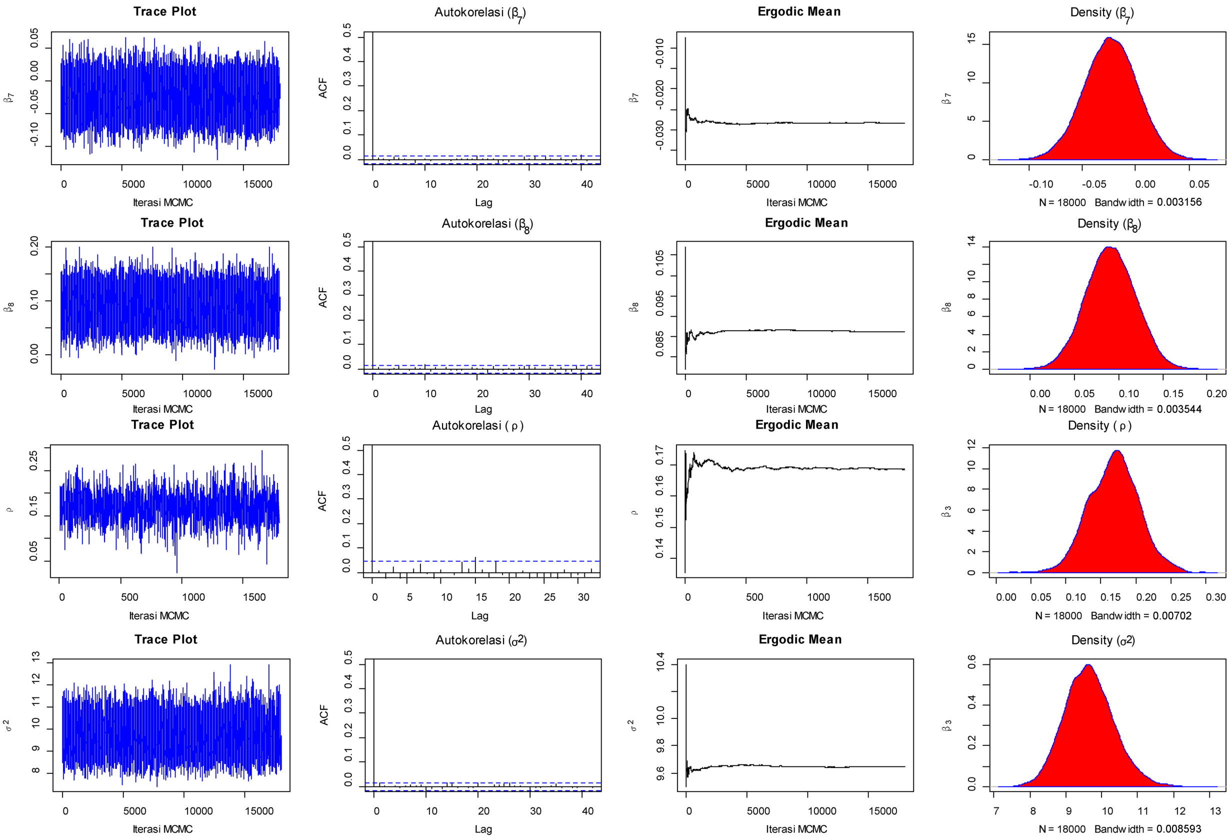


Fig. 5. continued.

trends or drifts. This result indicates that the convergence in distribution takes place rapidly. The next plots are the autocorrelation plots. Those autocorrelation plots also present the ideal pattern, or there is no significant lag. It means that the MCMC sampling process is drawn independently. Another way to check for the convergence of MCMC sampling is the ergodic mean. The third column presents the ergodic mean plots. The ergodic mean plot presents the stationarity condition. After 1,000 iterations, we can see that the ergodic mean has a convergence pattern. The right plot is a marginal density plot. The histograms present the distribution of the values of the parameter in the chain. Here, the distribution is close to the normal distribution. In general, all the diagnostic plots present the convergence pattern. It means that the estimation and inference processes are valid. The parameters estimate of the Bayesian spatial autoregressive model is presented in Table 3.

Table 3 display the parameters estimate of SAR model of junior high school in West Java by means ML and Bayesian approaches. Both estimations present similar results. It means that the Bayesian estimation gives an accurate estimate, which is similar to the maximum likelihood. In other words, the Bayesian estimation may be the best alternative for the ML if the normality assumption is violated. Two standards have a significant effect on the national examination score. The first standard is graduate competence (β_3) and the second is standard of assessment (β_8). Based on the 415 junior high schools in West Java, the graduate competence and assessment standard are important to be concerned with evaluating the national examination score. The interpretable result of the credible interval is the other advantage of the Bayesian approach in this case, besides of overcoming the normality assumption, is that the interpretable result is easier than ML because of the credible interval. For example, the credible interval of the parameter estimate of graduate competence (β_3) can be interpreted as "the parameter estimates of graduate competence lies with a 95% probability interval between 0.0068 and 0.0916."

The advantage of the SAR model compared with that of the standard regression model is that we can calculate the global spillover effect. The global spillover effects of those variables are 0.050 and 0.0926, respectively. It means that

Table 3
Parameters estimate based on the ML and Bayesian approaches

Parameters	ML	Bayesian			
		Mean	q (2.50%)	q (50%)	q (97.50%)
Intercept (β_0)	4.4169	4.4351	−0.0195	4.4316	8.8930
Standard of contents (β_1)	−0.0425	−0.0424	−0.1013	−0.0423	0.0160
Standard of process (β_2)	0.0129	0.0126	−0.0361	0.0127	0.0620
Standard of graduate competence (β_3)	0.0490*	0.0491*	0.0068	0.0493	0.0916
Standard of educational (β_4)	0.0271	0.0272	−0.0115	0.0272	0.0662
Standard of facilities and infrastructure (β_5)	0.0283	0.0284	−0.0148	0.0282	0.0713
Standard of management (β_6)	0.0319	0.0319	−0.0169	0.0320	0.0807
Standard of financing (β_7)	−0.0231	−0.0233	−0.0729	−0.0233	0.0252
Standard of assessment (β_8)	0.0908*	0.0911*	0.0366	0.0909	0.1455
σ^2	9.3552*	10.3683*	8.3855	9.6195	11.0822
ρ	0.17131*	0.1355*	0.1013	0.1693	0.2385
AIC = 241.047 (AIC for ML: 2135.4)					
$R^2 = 0.266$ (R^2 for ML: 0.266)					

*) significant at level $\alpha = 5\%$.

the change of graduate competence and assessment in one school gives impact on the change of learning outcomes in all schools with sizes of 0.050 and 0.0926, respectively.

4. Discussions and conclusions

The model optimization based on the k-NN method found that the appropriate spatial weight matrix (W) is at $k = 1$. It means that the strength spatial autocorrelation exists on one nearest neighbor where the national examination score for adjacent schools is relatively the same. Because of the violation of normality assumption in the ML approach, the estimation and inference of the spatial autoregressive model are resolved by the mean Bayesian approach. The parameters estimate of the ML and Bayesian approaches have similar results. We use a credible interval with 95% interval confidence for testing the null hypothesis. If the credible interval includes the zero values, the null hypothesis is accepted. We found that the national examination score of the junior high schools in West Java is influenced by the average of the national examination score from the first nearest schools and significantly influenced by standard of graduate competence (β_3) and standard of assessment (β_8). The result tells us that increasing the national examination score is relevant in increasing the competence of graduate and school assessment. The result of this research can be improved by accommodating the time effect. A spatial panel data model may be used to explain the temporal trend in the national score exam. This information is needed by schools and the government for studying the quality of education changes in Indonesia. Indonesia needs to improve the quality of its education continuously to improve the human capital development index.

Acknowledgments

This paper is funded by HIU-RKDU Unpad 2017. The authors thank Rector Universitas Padjadjaran and to the anonymous referee for valuable checking to improve this paper.

References

Altman, N. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185.

Amaratunga, D., & Carbera, J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. John Wiley & Sons, Inc.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. California: Springer.

Anselin, L. (2003). Spatial Econometrics. In Baltagi, B.H. *A Companion to Theoretical Econometrics* (pp. 310-330). Germany: Blackwell Publishing Ltd.

Congdon, P. (2013). Bayesian Spatial Statistical Modeling. In Fischer, M.M., & Nijkamp, P. *Handbook of Regional Science* (pp. 141-1434). New York: Springer.

Elfiza, Y., Rusman, & Nasir, M. (2016). Hubungan antara Hasil Uji Kognitif Try Out Ujian Nasional (UN) dengan Hasil Ujian Nasional (UN) Mata Pelajaran Kimia SMA Kota Banda Aceh Tahun Ajaran 2014/2015. *Jurnal Ilmiah Mahasiswa Pendidikan Kimia (JIMPK)*, *1*(3), 35-42.

Elhorst, J.P. (2014). *Spatial Econometrics From Cross-Sectional Data to Spatial Panels*. New York: Springer.

Firman, H., & Tola, B. (2008). The Future of Schooling in Indonesia. *CICE Hiroshima University, Journal of International Cooperation in Education*, *3*(1), 71-84.

Harvey, A.C. (1990). *The Econometric Analysis of Time Series*. United States: First MIT Press.

Ioannis, N. (2009). *Bayesian Modeling Using WinBUGS*. Greece: Wiley.

Jaya, I., Folmer, H., Ruchjana, B.N., Kristiani, F., & Andriyana, Y. (2017). Modeling of Infectious Disease: A Core Research Topic for The Next Hundred Year. In Randall, J., & Peter, S. *Regional Research Frontiers* (pp. 681-701). US: Springer.

Jaya, I., Ruchjana, B.N., Tantular, B., Zulhanif, & Andriyana, Y. (2018). Simulation and Application of the Spatial Autocorrelation Geographically Weighted Regression Model (SAR-GWR). *ARPN Journal of Engineering and Applied Sciences*, *13*(1), 1-9.

Klotz, S. (2004). *Cross Sectional Dependence in Spatial Econometrics Models with an Application to German Start Up Activity Data.* USA: Transaction Publisher.

LeSage, J., & Pace, R.K. (2009). *Introduction to Spatial Econometrics.* USA: Chapman & Hall/CRC.

Li, H., Calder, C.A., & Cressie, N. (2007). Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model. *Geographical Analysis*, *39*, 357-375.

MoEC, M.O. (2013). *Overview of the Education Sector in Indonesia 2012: Achievements and Challenges.* Jakarta: MoEC.

Newhouse, D., & Beegle, K. (2005). The Effect of School Type on Academic Achievement: Evidence from Indonesia. *World Bank Policy Research Working Paper 3604*, 1-44.

OECD, A.D. (2015). *Education in Indonesia: Rising to the Challenge.* Paris: OECD Publishing.

Rees, D. (2001). *Essential Statistics.* USA: Chapman & Hall.

SAS. (2011). *SAS/STAT 9.3 User's Guide Survival Analysis.* USA: SAS Institute Inc.

Shekhar, S., & Xiong, H. (2008). *Encyclopedia of GIS.* USA: Springer.

Tobias, J., Wales, J., Syamsulhakim, E., & Suharti. (2014). *Toward Better Education Quality Indonesia's Promising Path.* London: Overseas Development Institute.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(1), 234-240.

Vega, S.H., & Elhorst, J.P. (2015). The SLX Model. *Journal Regional Science*, 339-363.

Zhang, A. (2006). *Advanced Analysis of Gene Expression Microarray Data.* London: Wordl Scientific.

Appendix

The R code is available upon request.