A Data-driven Analysis to Attest Regional Interlinkages between Economics and Environment: Utilization of SDG Indicators for Spatio-temporal Analysis

Gigih Fitrianto
Graduate School of Economics
Hiroshima University of Economics

Summary of Dissertation

There is no doubt that one of the major causes of the current environmental changes is due to human, especially economic, activity. The purpose of this dissertation is to elucidate the temporal and spatial linkages between indicators for the Sustainable Development Goals (SDG indicators) and to conduct a comprehensive regional analysis. For this purpose, the availability of data according to the number of spatial and temporal data is important. This thesis attempts to develop a data-driven analytical approach under the limitations of data size for Indonesian data.

Chapter 1 provides a framework for data-driven analysis through the use of SDG indicators, which are designed to address socio-economic and environmental issues in an integrated manner in order to build a more comprehensive regional development analysis. In this chapter, we have developed a framework for drawing a general framework between socioeconomic and environmental factors through spatio-temporal analysis. For Indonesia's data, especially for the Gross Regional Product (GRP) by industrial sector, there is a change in the data classification from a base year of 2000 to a data classification with 2010 as the base year. This is a change in the classification of industries counted in the GRP data from 9 to 17 industries. With this new classification, data from Statistics Indonesia will be available from 2010. Thus, the maximum period of systematic availability is 9 years at the time of writing. The basic previous studies in the time series analysis in Chapter 2 and the regression model in Chapter 3 are considered in this chapter.

In Chapter 2, we analyzed the economic structural changes and regional resilience of neighboring municipalities using SDG indicators. A time series analysis of the broken time trend model was conducted using spatial dependency information. We focused on detecting the resilience of the regional spectrum to economic shocks and comparing the economic growth trends of each region before and after the shocks. The Lehman shock was chosen as a case study. Since the Lehman shock is the case study, this study does not use Indonesian data due to data limitations, but compares data from the U.S. and Japan, where larger data sizes are available: data for 51 U.S. states from 2005:Q1-2018:Q2 (number of periods: 54) and data for 47 prefectures of Japan from 2001-2014 (number of periods: 14). We extend the broken time trend model to a spatial autoregressive model and consider several statistical tests for detecting structural changes, such as the F-test and Wald test. In the U.S., the F-test is shown to be better at detecting spatial clusters when the spatial factor is not taken into account, but the Wald test is shown to be better when the spatial factor is taken into account. In Japan, the Wald test is not usable due to the limited number of data.

In the United States, the negative effects of the crisis are concentrated in the West Coast, Southeast, and Great Lakes regions. In the states in these regions, manufacturing, construction, insurance, and finance are the main contributors to the GRP. On the other hand, states dependent on agriculture, forestry, fisheries, fishing, hunting and mining were more resistant to shocks. In Japan, negative impacts were spread across all regions. After the crisis, most of the prefectures recovered, with the exceptions of Nagasaki Prefecture, for example.

Chapter 3 examines the general methodology for variable selection in the spatial regression analysis to demonstrate the inter-regional linkages between SDG indicators for the island of Sumatra, Indonesia. This chapter focuses on spatial factors in regional analysis to elucidate the interlinkages between SDG indicators across the region. As not all information is reliable to explain the data, a variable selection process was used to select the best variables to prove the interlinkages in and between the provinces.

In order to obtain the best variables to prove interregional linkages, we used a linear transformation of the full spatial panel model, the largest model used in spatial panel analysis, to obtain the best variables to prove the linkages. Variable Selection in Spatial Regression (VSSR) is newly developed and used.

The full spatial panel model can be expressed as follows,

$$\begin{aligned} \boldsymbol{y}_t &= \alpha\mathbf{1} + \delta\mathbf{W}\boldsymbol{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}_t\boldsymbol{\theta} + \boldsymbol{\nu}_t \\ \boldsymbol{\nu}_t &= \lambda\mathbf{W}\boldsymbol{\nu}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathrm{i.i.d.\,N}(0, \sigma^2\mathbf{I}), \quad t = 1, \dots, T \end{aligned} \tag{1}$$

where $\boldsymbol{y}_t$ gives $N \times 1$ vector of dependent variable at time t with N being the number of regions. $\mathbf{X}_t = [x_{1t}, \dots, x_{Kt}]$ is $N \times K$ matrix which consists of K number of independent variables. The $\mathbf{W}$ is a spatial neighbor matrix of size $N \times N$ with $w_{ii} = 0$, which stores the neighborhood information between regions. The α represents a common intercept and $\mathbf{1} = [1 \quad , \dots , \quad 1]'$ is $N \times 1$ vector of one. $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are $K \times 1$ vectors of regression coefficients and spatial regression coefficients for explanatory variables respectively. To obtain the best variables to attest regional interlinkage, VSSR used a linearized transformation of equation (1), into

$$(\mathbf{I} - \lambda \mathbf{W})\boldsymbol{y}_t \;= (\mathbf{I} - \lambda \mathbf{W})\{\alpha \mathbf{1} + \delta \mathbf{W} \boldsymbol{y}_t + \mathbf{X}_t \boldsymbol{\beta} + \mathbf{W} \boldsymbol{X}_t \boldsymbol{\theta}\} + \boldsymbol{\epsilon}_t$$

$$\widetilde{\boldsymbol{y}}_t = \tilde{\alpha} \mathbf{1} + \delta \widetilde{\boldsymbol{W}} \boldsymbol{y}_t + \widetilde{\boldsymbol{X}}_t \boldsymbol{\beta} + \widetilde{\boldsymbol{W}} \boldsymbol{X}_t \boldsymbol{\theta} + \boldsymbol{\epsilon}_t \tag{2}$$

The VSSR can be obtained from following procedures,

Step 0. Use the linearized form in equation (2) as base model.

Step 1. Fix λ in the interval $[0, \phi)$ for grid search, where ϕ is the reciprocal of the maximum eigen value of matrix $\mathbf{W}$.

Step 2. Obtain the least square estimates $\hat{\tilde{\alpha}}(\lambda), \hat{\delta}(\lambda), \hat{\boldsymbol{\beta}}(\lambda), \hat{\boldsymbol{\theta}}(\lambda)$ and $\hat{\sigma}^2(\lambda)$ for fix λ.

Step 3. Find the best model minimizing $\mathrm{AIC}(\lambda)$ for variable selection based on following calculation,

$$\mathrm{AIC}(\lambda) \;= NT \log(2\pi e) + NT \log \hat{\sigma}^2(\lambda) + 2 \cdot T \log|\mathbf{I} - \lambda \mathbf{W}| + 2 \cdot d$$

where N is the number of province in Sumatra island, T is number of time period, and d is the number of parameters for $\tilde{\alpha}, \delta, \sigma^2, \boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

The variable selection was conducted by evaluates the model by dropping one explanatory variable at a time, which minimizes the AIC value, until there is no significant drop in AIC value. In this study we used *step* function in R library.

Step 4. Repeat **Steps 1 – 3**.

In this empirical analysis, we apply the VSSR to observe the relationship between industry and regional income, mainly palm oil, in Sumatra, Indonesia. Using annual data from 2011 to 2017 for 10 provinces in Sumatra, the dependent variable is regional income, and the independent variables are 1) GRP growth rate per capita, 2) secondary school enrollment rate, 3) population growth rate, 4) crude palm oil production per labor force, 5) mining production per labor force, and 6) wholesale production per labor force. quantity, and 7) the ratio of children under fifteen years old to population (child labor ratio). The results show that school attendance, palm oil production, and wholesale trade were selected and positively influenced the local income. On the other hand, mining and child labor rates are negatively related after the selection of VSSR. Neighborhood schools, palm oil production, and wholesale trade had a positive impact on each province in Sumatra. Neighborhood income and child labor rates in neighboring provinces have a negative effect. This chapter further explains how to incorporate data between socio-economic variables (e.g., income inequality, child labor participation) and environmental data (e.g., change in forest cover, CO2 emissions, and forest fire emissions), and then examines the actual addition of the standard deviation (degree of topographic unevenness) of the aggregated 1 km square elevation of each province in Sumatra as a variable. The results showed that states with more plains had significantly higher regional income. This was suggested to be related to the fact that the suitable land for palm oil plantations is flat land.

In Chapter 4, the giant grid space adjacency matrix of socioeconomic grid cell data for computationally efficient and sustainable economic analysis was formulated. The advantage of using grid cell data for socio-economic analysis is that it is relatively easy to incorporate satellite data into regional analysis and has an important role to play in observing the relationship between the socio-economy and the environment. This corresponds to the sustainable development goals of harmonizing socio-economic and environmental balances. The creation of a spatial adjacency matrix to project spatial relationships within a region plays an important role in such an analysis. However, no previous studies have provided a practical formulation of spatial adjacency matrices in grid cell data structures that take into account socioeconomic analysis. This problem stems from the existence of NA cells representing non-inhabited areas, such as water bodies, but the shapefiles commonly used in spatial analysis do not contain this information within municipalities. Since these NA cells create a non-rectangular grid, it is important to exclude them in the analysis in order to project the actual information correctly. In this chapter, a method is provided to construct the adjacency matrix using Kronecker products and apply a projection matrix to remove the NA cells to project the real information correctly. The method using the Kronecker product showed very high efficiency compared to the widely used R package called *spdep*. The experimental results show that the computation time is more than 2000 times faster than that of the *spdep* package for huge data with as many as one trillion elements.

This dissertation has established demonstrative original methodology for elucidating the interaction between the economy and the environment in regional analysis.