# Spatial weighting approach in numerical method for disaggregation of MDGs indicators

**S D Permai[1*], U Mukhaiyar[2], N L P Satyaning P P[3], M Soleh[4], Q Aini[5]**

[1]Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
[2]Statistics Research Division, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Bandung 40132, Indonesia
[3]Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya 60111
[4]Department of Statistics, Bogor Agricultural University, Bogor 16680, Indonesia
[5]Regional Balance and Statistical Analysis Sub-Division, Central Statistics Agency, West Nusa Tenggara 84455, Indonesia

[*]Corresponding author: syarifah.permai@binus.ac.id

**Abstract.** Disaggregation use to separate and classify the data based on certain characteristics or on administrative level. Disaggregated data is very important because some indicators not measured on all characteristics. Detailed disaggregation for development indicators is important to ensure that everyone benefits from development and support better development-related policymaking. This paper aims to explore different methods to disaggregate national employment-to-population ratio indicator to province- and city-level. Numerical approach applied to overcome the problem of disaggregation unavailability by constructing several spatial weight matrices based on the neighbourhood, Euclidean distance and correlation. These methods can potentially be used and further developed to disaggregate development indicators into lower spatial level even by several demographic characteristics.

## 1. Introduction

The 17 Sustainable Development Goals (SDGs) build on the successes of the previous eight Millennium Development Goals (MDGs), while including new areas such as climate change, economic inequality, innovation, sustainable consumption, peace and justice, among other priorities. Despite substantial progress has been made on many of MDGs, the progress has been uneven across regions and countries [1]. Millions of people are being left behind, especially the poorest and the vulnerable groups because of their gender, age, disability, ethnicity or geographic location.

Learning from MDGs, one of the highlights of SDGs is “leaving no one behind”. It can be seen that the SDGs targets itself requires more disaggregated data by several demographic characteristics as mentioned above. Since disaggregation are not available for MDGs indicator, there will be a limitation to analyze both SDGs and MDGs data together for monitoring and research purpose. It is indeed important to have disaggregation as detail as possible for development indicators in order to (i) ensure that the benefit of the development reach everyone and (ii) assist the formulation of better policy to

achieve the goals and targets. However, the important point is that disaggregation according to these dimensions would not be relevant for all indicators [2].

This study focus on estimating development indicator at the local level. The local level is the geographical level at which data are requested with a view to planning sub-regional policies or evaluating the results of policy [3]. Spatial disaggregation methods are based on area interpolation techniques. The spatial relationship imposed in disaggregation data process. This way used in all such techniques to decrease error value [4]. Several methods are proposed and piloted to spatially disaggregate one of important indicators in development goals, which is employment-to-population ratio. Employment-to-population ratio is one of indicators for the second target of Goal 1 Eradicate poverty and hunger: achieve full and productive employment and decent work for all, including women and young people. The national-to-province and province-to-city disaggregation has been done using 2011 data.

## 2. Simple Proportion

One of disaggregation method is weighted method. Weighting method using proportion is the simplest approach for disaggregating data. This method assumes that target variable ($Y_i$) is uniformly distributed in each area. The target variable can be estimated as [5],

$$Y_i = \frac{A_i}{\bar{A}} Y \tag{1}$$

Where,
$i = 1, 2, \ldots, n$,
$Y_i$ : value of indicator for unit $i$
$Y$ : value of MDGs indicator in higher level
$A_i$ : value of non-MDGs indicator for unit $i$
$\bar{A}$ : average value of non MDGs indicator in higher level

Note that $A_i$ is a variable that highly correlated or has similar pattern with respective MDGs indicator. For this study, proportion of working population to the total population is used as $A_i$.

## 3. Numerical Method Approach

In this approach, numerical method principle is applied. There are two categories in numerical methods, direct methods and iterative methods. Direct methods give exact solution of problem without rounding error. Iterative methods find solution from a sequence of approximation solutions. This method using starting point $Y^{(0)}$ and generate sequence of approximate solutions $Y^{(k)}$. The latest approximations to the components of Y are used in the update of subsequent components [6]. The simple proportion defined in previous subsection can be considered as the initial value $Y^{(0)}$.

In this paper, numerical method used is iterative method. Iterative methods generate a sequence of approximations to the desired solution, often referred as successive approximation or trial and error method. This method is start with a function, which maps one approximation into another better. In this way a sequence of possible solutions to the problem is generated. The approximation obtained acceptably accurate when the solution is convergent. [7] The sequence is said to converge to the limit if $\left|Y - Y^{(m)}\right| < \varepsilon$. Iterative methods to find a sequence of approximation solutions following

$$\vec{Y}^{(k+1)} = \boldsymbol{W} * \vec{Y}^{(k)} \tag{2}$$

$$\vec{Y}^{(k)} = \left(Y_1{}^{(k)} \quad Y_2{}^{(k)} \quad \ldots \quad Y_n{}^{(k)}\right)' \tag{3}$$

$$Y_i{}^{(0)} = \frac{A_i}{\bar{A}} Y \tag{4}$$

$$\bar{A} = \frac{\sum_{i=1}^{n} A_i}{n} \tag{5}$$

$W$ is a spatial weight matrix $\{w_{ij}\}$ and $\vec{Y}^{(k)}$ is $k$-th iteration value of indicator in $i$-th area.

Stopping rule is defined as if $\left|\vec{Y}^{(k)} - \vec{Y}^{(actual)}\right| < \varepsilon$, for any small real values $\varepsilon > 0$. This approach is used by knowing the real values so that how good this approach can be identified. From this result then data disaggregation to smaller areas can be executed using the same approaches.

The most important thing in numerical method approach for data disaggregation is determining the spatial weight matrix. The spatial weights matrix is an integral part of spatial modeling and defined as the formal expression of spatial dependence between observations [8]. There are several methods can be used to construct the spatial weight matrix. Based on Tobler's first law said that everything is related to everything else, but near things are more related than distant things [9]. Therefore, in this paper several methods of constructing spatial weight matrix using geographical proximity between areas are experimented.

### *3.1. Neighbourhood Based*

Nearest neighbor method uses the simplest way to determine the weight. This method uses the determination of spatial unit share a boundary or not. The next step is to create a matrix $\boldsymbol{M}$ which contains the coding between the units that have shared a boundary or not. This method also called as rook contiguity [10].

$$\boldsymbol{M} = \{m_{ij}\} = \begin{bmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{nn} \end{bmatrix} \tag{6}$$

$$m_{ij} = \begin{cases} 1 & \text{location } i^{th} \text{ and } j^{th} \text{ share the same borderline} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

In most cases it is convenient to normalize spatial weights to remove dependence on extraneous scale factors. This produces row normalization matrix called matrix $\boldsymbol{W}$.

$$T = diag\left(\frac{1}{\sum_{j=1}^{n} m_{1j}}, \frac{1}{\sum_{j=1}^{n} m_{2j}}, \ldots, \frac{1}{\sum_{j=1}^{n} m_{nj}}\right) \tag{8}$$

$$\boldsymbol{W} = T * \boldsymbol{M} \tag{9}$$

### *3.2. Euclidean Based*

Geographical proximity can be measured using distance. The most common distance, Euclidean distance, is applied in this paper. Given x and y is longitude and latitude coordinate, respectively, below is the formula for calculating the distance between the two units [11].

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \tag{10}$$

The problem is there is no maximum limit value of the distance, so that the distance values must be normalized to obtain a spatial weight matrix as follow.

$$w_{ij} = \begin{cases} 0 & if\ i = j \\ \dfrac{\left(\frac{1}{1 + d_{ij}}\right)}{\sum_{i=1}^{n}\left(\frac{1}{1 + d_{ij}}\right)} & if\ i \neq j \end{cases} \tag{11}$$

### *3.3. Correlation Based*

Methods based on correlation are desirable if the relationships among original distances do not follow a mathematically predictable pattern or are thought to be non-linear. The correlations do not change when distances are transformed [12]. Define correlation matrix of intended units based on data history and construct a distance matrix D as follow:

$$d_{ij} = \sqrt{2 * \left(1 - r_{ij}\right)^2} \tag{12}$$

Where $r_{ij}$ is sample correlation between location *i*-th and *j*-th. For obtaining this correlation, the history data (previous observations) of respective MDG's indicator are needed. Two units which have higher correlation means the distance between two units are nearer. So, spatial weight matrix is improved by the correlation among neighbors who shared a boundary.

$$w_{ij} = \frac{m_{ij}\, d_{ij}{}^{-1}}{\sum_{\substack{k=1 \\ k\neq i}}^{n} m_{ik}\, d_{ik}{}^{-1}} \tag{13}$$

## 4. Data and Methods

In this research, the MDGs indicator data used are the employment to population ratio index in 2011 for the national level and province level. The provincial level used are DKI Jakarta and West Java. The data will be disaggregated from national to province and from province to city. Non-MDGs indicator data that used are proportion of working population to the total population in 2011 and administrative map of Indonesia, DKI Jakarta and West Java. The data disaggregation is performed using simple proportion and spatial weighting approach.

## 5. Results and Discussion

The focus of this section is discussing the results and evaluating the method to conclude the best method so far. Aggregation from national to province level has firstly been done using simple proportion and numerical approach with three methods of weight matrix construction explained above. The disaggregation models developed show different results for each province, as shown in Figure 1. There are little differences of estimated pattern among big islands in Indonesia.

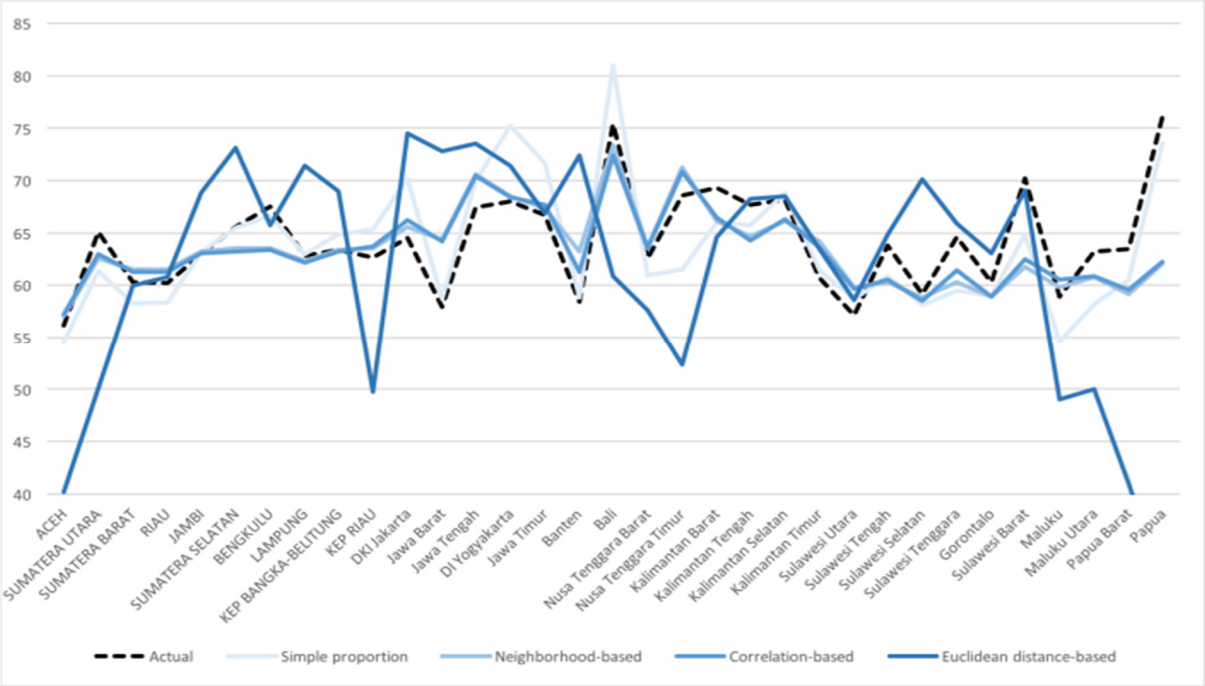


**Figure 1.** National to province disaggregation results and the actual data.

Based on Figure 1, most of the models underestimate the employment to population ratios for Aceh, Papua and Papua Barat, and provinces situated in Sulawesi island. The highest deviation found in estimating the employment to population ratio of Papua Barat. Besides the underestimation, overestimation can be found in provinces situated in Java islands. Employment to population ratio of provinces in Sumatera island, Bali, and Nusa Tenggara are closely estimated, where in Bali and Nusa

Tenggara the variance is smaller while in Sumatera island it is higher. One of the reasons of getting either overestimation or underestimation is that the non-MDGs official statistics used for initial proportion weighting of the respective provinces have different pattern with the employment-to-population ratio. Some provinces are not really affected by their neighborhood, while some are influenced a lot by them. This can also lead to the higher deviation in estimating an indicator in lower spatial level.

**Table 1.** National-to-province model evaluation

| Method | MAE | MAPE | MSE |
|---|---|---|---|
| Simple proportion | 2.689 | 4.1 | 11.323 |
| Numerical Method Approach: | | | |
| Neighborhood-based | 2.726 | 4.2 | 14.710 |
| Euclidean distance-based | 8.347 | 12.9 | 147.018 |
| Correlation-based | 2.615 | 4.0 | 13.319 |

In order to evaluate the models, several statistics to measure the goodness of models have been calculated as shown in Table 1. The lower the value of mean average error (MAE), mean average percentage error (MAPE), and mean squared error (MSE), the better the model. From these three criterion, correlation-based numerical approach have better estimated the employment-to-population ratio of province-level with the lowest MAE and MAPE, 2.615 and 4 respectively. It is also noted that Euclidean distance-based method gives the worst estimation (highest value for the three criterion) results among proposed methods, which indicates that the closer distance does not lead to the higher dependency among locations. That neighborhood-based model is better that the Euclidean distance one indicates that the locations which share same administrative borderline have a bigger chance to influence each other's.

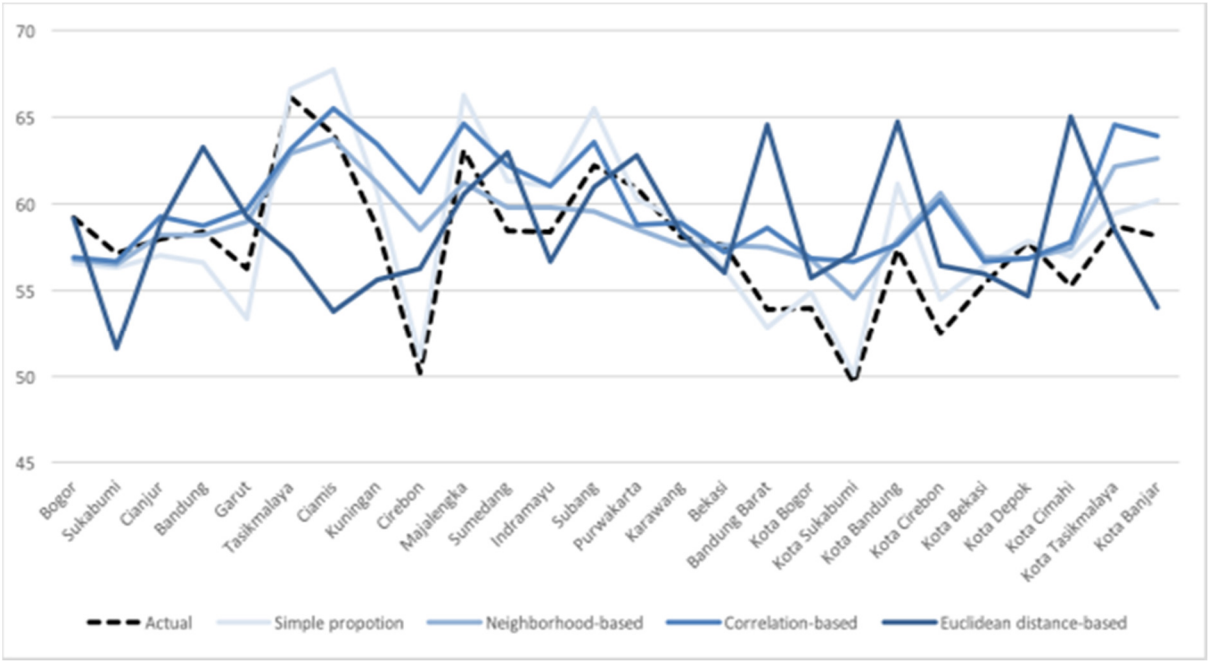


**Figure 2.** Province to city disaggregation results and the actual data for West Java

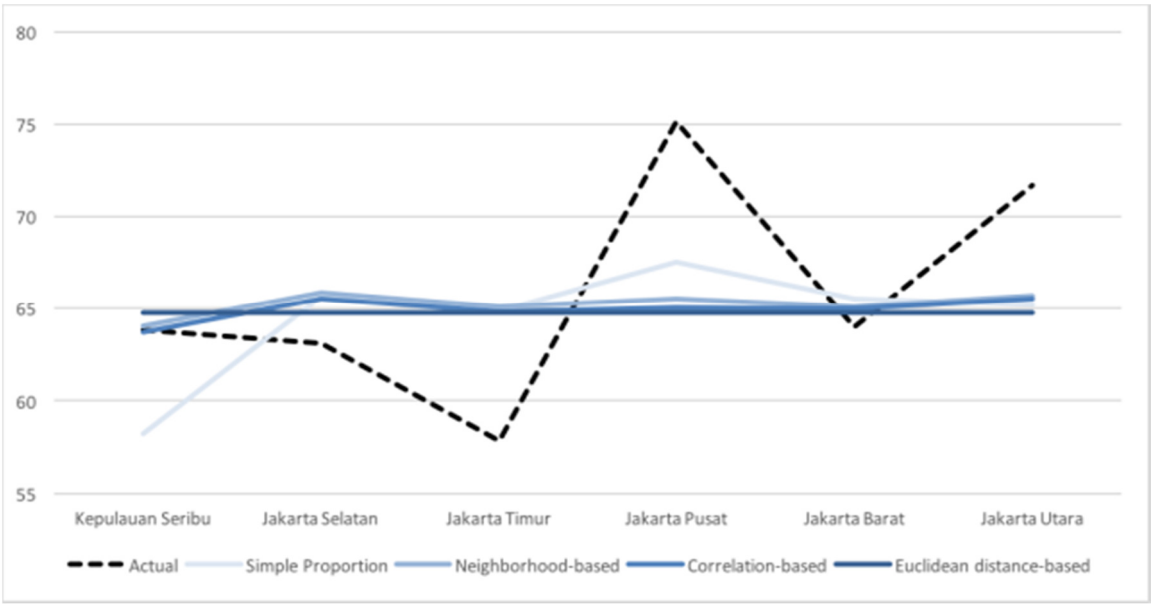


**Figure 3.** Province to city disaggregation results and the actual data for DKI Jakarta

Province to city disaggregation are also done for DKI Jakarta and West Java using the same methods, showed in Figure 2 and Figure 3. Although there are some underestimations (e.g. Ciamis, Tasikmalaya) and a lot of overestimations, the models for West Java disaggregation can closely estimate several cities, for instance Majalengka, Sumedang, Indramayu, Subang, Purwakarta and Karawang. Based on the evaluation criterion in Table 2, simple proportion with the lowest MAE, MAPE and MSE (1.739, 3 and 4.174 respectively) is the better method to disaggregate province level data into city level data.

**Table 2.** Province to city model evaluation

| Province | Method | MAE | MAPE | MSE |
|---|---|---|---|---|
| West Java | Simple proportion | 1.739 | 3 | 4.174 |
| | Numerical Method Approach: | | | |
| | Neighborhood-based | 2.445 | 4.4 | 10.591 |
| | Euclidean distance-based | 4.061 | 7.1 | 26.938 |
| | Correlation-based | 3.062 | 5.5 | 15.834 |
| DKI Jakarta | Simple proportion | 5.137 | 7.8 | 31.606 |
| | Numerical Method Approach: | | | |
| | Neighborhood-based | 4.490 | 6.7 | 31.592 |
| | Euclidean distance-based | 4.604 | 6.8 | 34.711 |
| | Correlation-based | 4.468 | 6.6 | 32.653 |

However, the models for DKI Jakarta are not well estimating the employment to population ratio of its cities and the value for all cities are almost the same towards one number. One crucial aspects that affecting this result is that the weight spatial matrix developed does not suit DKI Jakarta. It can be that the characteristic of five cities are very similar as well as the cities share almost the same borderlines and almost all cities become the neighbor of others. It is obvious that the best model is the simple proportion one, since the weight matrix does not work well for DKI Jakarta.

## 6. Conclusion

Spatial weighting using numerical approach can be potentially used as to estimate development indicators at lower spatial level. Further improvement needed in order to get the most suitable spatial

weight matrix, since it is indeed the most crucial part in numerical method disaggregation. Another important thing is finding the non-MDGs official statistics that highly correlated or have similar pattern with the respective MDGs indicator to construct initial proportion weight. This paper contributes well in proposing the methodologies of data disaggregation to monitor the achievement of development indicators at local level, and therefore, to make sure that no one left behind.

**References**

[1] United Nations 2015 *The millennium development goals report*
[2] Espey J and Karoubi E M 2015 *A Global Initiative For The United Nations* **26**
[3] Pratesi M, Petrucci A and Salvati N 2015 *Global Strategy* Technical Report Series GO-07-2015
[4] Global Strategy 2015 *Technical Report Series GO-07-2015*
[5] Flowerdew R and Green M 1994 *Areal Interpolation and Types of Data* (London: Taylor and Francis)
[6] Kubicek M, Janovska D and Dubcova M 2005 *Numerical Methods and Algorithm* (Praha: Vysokáškolachemicko-technologická v Praze)
[7] McDonough J M 2007 *Lectures in Basic Computational Numerical Analysis* (USA : University of Kentucky)
[8] Getis A and Aldstadt J 2003 *Geographical Analysis* **36** No 2 p 90 – 104
[9] Drukker D M, Peng H, Prucha I R and Raciborski R 2013 *The Stata Journal* **13** No 2 p 242 – 286
[10] Tang B and He H 2015 *IEEE Computational Intelligence Magazine* 1556-603x/15©2015IEEE p 52 – 60
[11] Krislock N and Wolkowicz H 2011 *Handbook of Semidefinite, Cone and Polynomial Optimization: Theory, Algorithms, Software and Applications*
[12] Amerise I L and Tarsitano A 2012 *Working Paper* vol 09 (Italy: Dipartimento di Economia e Statistica, Universita Della Calabria)