Technical Report

The Third Research Dive on Statistics for Sustainable Development Goals

May 2017

Australian Aid

PULSE LAB JAKARTA

Executive Summary

In September 2015, world leaders adopted the Sustainable Development Goals (SDGs)[1] as the successor of the Millennium Development Goals (MDGs)[2]. The SDGs lay out a universal holistic framework to help set the world on a path towards sustainable development. Compared with the eight goals of the MDGs, the SDGs set more comprehensive goals, 17 in total, to address the symptoms of poverty and the issues of peace, stability, human rights and good governance.

In the SDGs era, data have become more essential not only to better measure the achievements of the 17 goals but also to better implement them, specifically given that the SDGs advise countries to collect 230 indicators (even more when country-specific indicators are included) for monitoring 169 targets. Moreover, the SDGs require more disaggregated data, *e.g.*, by gender, by age, and at sub-national levels. The Government of Indonesia is faced with a huge data gap around the 230 indicators, yet there are also opportunities to explore new data sources (*e.g.*, administrative data and data from the private sector).

The Government of Indonesia has conveyed that 36 percent of the SDG global indicators are not yet tracked in Indonesia[3]. Several initiatives have been undertaken by National Bureau of Statistics (BPS), for instance, modifying existing surveys, developing new surveys, and exploring the possibility of using big data and other data sources. However, since some indicators remain uncovered by conventional methods, BPS is looking for more sophisticated approaches, statistical models and analysis to fill the gap of data availability, accessibility, and reliability.

In order to support the Government of Indonesia to utilize existing data collection and monitoring frameworks in Indonesia, the third Research Dive addressed the topic of Statistics for the SDGs and invited 20 participants from academia, statistics researchers and practitioners from BPS as well as five advisors from universities and BPS. During the research days, the participants analyzed 15-year MDG indicator data at the (sub-) national level. The Research Dive included four topics: (1) correlation and causality, (2) proxy indicators, (3) quality of data, and (4) data disaggregation. This outcomes report from the Research Dive is split into six extended abstracts.

The first paper explains the role of statistics for supporting the SDGs. This paper also explores the lessons learned from the achieved and unachieved MDG targets for the implementation of the SDGs. The second to the sixth paper are the outcomes of research by participants. The second paper explores the correlations between MDG indicators, and also clustered the provinces based on performance. For example, the group discovered a positive correlation between extreme poverty and hunger and the incidence of tuberculosis. The third paper explores the role of education and health on poverty alleviation. It showed that literacy rates and sustainable access to basic sanitation have the greatest statistical impact on poverty. The fourth paper explains proxies for currently unavailable SDG indicators, for instance, measuring the proportion of the population using safely managed drinking water services by using measurements including the proportion of the population below the poverty line, the proportion of the population consuming clean water, and the proportion of the population with access to improved sanitation. The fifth paper proposes a framework to ensure the quality of data, by using proxies and validating outliers. The last paper examines the disaggregation of national level data to the provincial level, by applying a set of numerical methods, including simple proportion, neighborhood-based, and correlation-based methods.

Pulse Lab Jakarta is grateful for the cooperation of Binus University, Universitas Brawijaya, Universitas Diponegoro, Universitas Gajah Mada, Universitas Hasanuddin, Universitas Islam Indonesia, Institut Teknologi Bandung, Institut Teknologi Surabaya, Institut Pertanian Bogor (IPB), Institute of Resource Governance Social Change (IRGSC) Kupang, Sekolah Tinggi Ilmu Statistik (STIS), National Bureau of Statistics, BPS Semarang, BPS Kalimantan Timur, BPS Papua Barat, and BPS Nusa Tenggara Barat.

1 http://www.un.org/sustainabledevelopment/

2 http://www.un.org/millenniumgoals/

3 UNDP Indonesia. 2016. Indicators and Data Mapping to Measure Sustainable Development Goals (SDGs) Targets: Case of Indonesia 2015. http://www.id.undp.org/content/dam/indonesia/2016/doc/SDGs%20Indicators%20and%20Data%20Mapping%20in%20Indonesia_UNDP%20UNEP%202015.pdf?download

Advisor Note

Filling the Data Gap with Statistical Modelling

One of the important aspects in the preparation stage of SDGs implementation is mapping data availability. Compared to the MDGs, the SDGs, covering 17 goals, 169 targets and 241 indicators, are more complex because there are many cross-cutting issues among the goals and targets. The latest mapping of data availability for the SDG indicators shows that about 39 percent of the indicators, with their concepts and definitions being matched with the concepts and definitions of the global indicators, are available in the country, while about 30 percent are still in the form of proxy indicators, i.e. the concepts and definitions of the indicators are not matched with global concepts. The remaining 31 percent of the SDG indicators are not available in Indonesia. Another problem that has emerged in the SDGs implementation is about data disaggregation to address the "no one left behind" principle.

Given the fact that there is still a relatively large gap in data availability for the SDG indicators and the data disaggregation problem, several approaches should be taken. One of the approaches that can be taken into account is through statistical modeling. An initiative taken by the Research Dive to invite people from academia with a strong background in statistical theory and those from statistical offices with experience in data production is a brilliant idea. I am very fortunate to join this Research Dive because I learnt a lot from every group about statistical modeling to develop proxy indicators and data disaggregation. In the future such a modeling approach can be an important tool to solve the data availability problem faced in the SDGs implementation.

Dr. Ali Said, M.A
Domain Expert for SDGs Statistical Indicator

Dr. Ali Said, M.A is the Head of Sub-directorate for Statistical Indicators, BPS-Statistics Indonesia. He holds a Ph.D degree from Flinders University, Australia. He is currently working in BPS-Statistics Indonesia in the area of development of statistical indicators. He is also actively involved in the process of SDGs implementation in Indonesia and working closely with the National SDGs' Secretariat.

Enhancing Discussion Among Statisticians

The Research Dive is a good platform and it is important to conduct these events regularly for academia. On this occasion, academia had the opportunity to work with BPS to exchange and verify their research concepts and methods and whether there is potential or not for tackling the real problems. In addition, by applying their research methods, the participants were encouraged to find solutions within a limited time period.

In my opinion, the discussions and joint research among statisticians must be enhanced. By knowing other perspectives and sharing ideas, it may improve their ability to interpret the statistical models or quantitative methods. Particularly in the era of big data, interaction among statisticians and other specialists is needed for government policy making and private sector engagement In the future, I hope there will be more activities like the Research Dive that continue to strengthen the relationship with academia.

Dr. Suhartono
Advisor for Statistics

Dr. Suhartono is currently working as a Senior Statistics Lecturer as well as serving as Head of the Statistics Department in the Institut Teknologi Sepuluh Nopember (ITS). He has taught time series analysis, exploratory data analysis, multivariate analysis and research methodology for the past 19 years. He is the author of "Analisis Data Statistik dengan R" published by Graha Ilmu. He has a B.Sc. in Statistics from ITS Indonesia. He obtained an M.Sc. in Statistical Analysis and Stochastic Systems from the University of Manchester Institute of Science and Technology (UMIS), UK, and received an M.S. Bartlett Price in this Masters Program. Suhartono has a Doctoral degree in Statistics from Universitas Gadjah Mada, Indonesia, under supervision of Prof. Subanar. He pursued a Postdoctoral fellowship at Universiti Teknologi Malaysia (UTM).

Advisor Note

Promoting Creative and Problem-Solving Statisticians

When PLJ called me to join the Research Dive #3, I was very glad because I expected that it would be an excellent event to meet and share with great people. I supposed that there would be statisticians and data scientists with many ideas to deal with the statistical problems offered by the committee. My three-day participation confirmed that this was the case.

I noticed that the Research Dive #3 was a unique event for four reasons. First, it focused on a quite fresh and hot topic, which was statistics for the SDGs. Second, it gave participants an opportunity to express their ideas on designing optimal methodologies that best fit their research problems. Third, it was a really great chance for participants to build strong collaboration with other participants, advisors, as well as the invited audience. Lastly, this Research Dive encouraged the participants to show their best capability to deal with challenging problems.

To sum up my comments, I would agree that PLJ succeeded in creating an environment that was competitive but conducive to collaboration for statisticians to be creative and sensitive to problems in the applied areas.

Dr. Bagus Sartono
Advisor for Statistics

Bagus Sartono currently works as a lecturer in the Department of Statistics - Bogor Agricultural University in the subjects of statistical learning and data science. Apart from teaching, he also supervises undergraduate and postgraduate students' theses in statistics, economics, and business. He has extensive experience in applied research in a wide-range of topics for business and governmental agencies. Together with colleagues, he has written methodological books and many journal papers.

An important step to achieve SDGs

I noticed that the tasks during the three days of the Research Dive event were very challenging. The participants of the Research Dive were digging into the MDG data, which is the predecessor of the SDGs. Although the SDGs data are, more or less, structured data in terms of unit, variable definition and with a reasonable size; the tasks were quite general and methodological, in such a way that the five groups had to explore many substantial matters as well as statistical techniques.

The collaboration of participants with diverse competencies, including statistics, official statistics, social science and computing expertise in each group was one important factor for accomplishing the tasks. I was impressed with their final presentations at the event and very much enjoyed the discussions during this part of the event. Given the presentations, the final technical reports seem promising and I look forward to reading these.

It was a great pleasure and experience to be involved in the discussions and activities in the Research Dive: Statistics for the SDGs. To know the existing situation by analyzing the data related to the SDGs is the first important step to achieve the 17 Sustainable Development Goals.

Drs. Danardono, M.P.H., Ph.D
Advisor for Statistics

Danardono graduated from the Statistics undergraduate Program, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada (UGM) in 1992. He then worked in his alma mater; and also worked as a part time data analyst in the Community Health Nutrition Research Laboratories (CHN-RL), Faculty of Medicine UGM. He received a Master of Public Health in Biostatistics (MPH) from the Department of Biostatistics and Demography, Faculty of Public Health, Khon Kaen University, Thailand. The field of epidemiology and medicine motivated him to do methodological research in the area as his doctoral study In 2005 he earned a Ph.D in Statistics from Umeå University, Sweden. Since then, he has been working in teaching, consulting and research in biostatistics, epidemiology, demography, mortality models and computing. Lately, his interest in mortality models led him to conduct research in actuarial science and he has been involved in many research projects in this area.

Advisor Note

The Research Dive is a Good Platform for Sharing

It was a great honor for me to be an advisor for the Research Dive: Statistics for the SDGs. This special event allowed me to share my long experience as a data analyst and my knowledge as a lecturer for participants' discussion.

I enjoyed all of the participants' enthusiasm in utilizing statistics for developing SDG proxy indicators. Within the Research Dive, PLJ conducted a mini workshop and invited PLJ partners from government and research institutes. Through this opportunity, I was also able to share my research findings during my work with BPS.

I really hope that each of the participants can use this valuable experience to strengthen their capacity in doing further research in their own disciplines. I hope that this event will continue in the near future and be used to capture the real development process of the country. Although the result findings of the five research teams were excellent, I considered that the 'data disaggregation team' has made a promising effort in producing disaggregated data. I believe that when this team has an opportunity to further develop their current result they may be able to address the data gaps for the SDGs.

Dr. Tiodora Hadumaon Siagian, M.Pop.Hum.Res
Advisor for Statistics

Dr. Tiodora obtained her PhD from the Statistics Department of Institut Teknologi Sepuluh Nopember (ITS). She completed her master's degree on Population and Human Resource, the Applied Population Studies Programme, School of the Environment, Flinders University of South Australia. Beforehand, she obtained a Diploma IV in Diploma IV, Institute of Statistics (STIS) Jakarta. She has served as a lecturer in STIS since 2005. Aside from teaching, she also has various experience in Indonesia's National Statistics Agency, from 1991 to 2015. She served as the head for several sections in BPS National: the Head of Preparation of Statistics Activity Section in the Environment Statistics Division (2006-2008); Head of Poverty Statistics Section in the Social Vulnerability Statistics Division (March-July 2008); and Head of Social Environment Statistics Section in the Environment Statistics Division (2014-2015).

Research Dive

Advisors

Ali Said	Badan Pusat Statistik (BPS)
Dr. Bagus Sartono	Bogor Agricultural University
Drs. Danardono, M.P.H., Ph.D	Gajah Mada University
Dr. Suhartono	Sepuluh Nopember Institute of Technology
Dr. Tiodora H. Siagian, M.Pop.Hum.Res	Sekolah Tinggi Ilmu Statistik

Researchers

Group 1 – Identifying correlation structures within and between MDGs Goals

Dr. Adji Achmad Rinaldo Fernandes, S.Si, M.Sc	Brawijaya University
Diaz Fitra Aksioma, M.Si	Sepuluh Nopember Institute of Technology
Lilis Anisah, S.ST, M.Si	BPS of Semarang City
Dr. techn. Rohmatul Fajriyah, S.Si., M.Si	Islamic University of Indonesia
Imaduddin Amin	Pulse Lab Jakarta

Group 2 – The Impact of Health and Education on Poverty Reduction: A Causal Analysis

Achmad Efendi, PhD	Brawijaya University
Dr. Rr. Kurnia Novita Sari	Institut Teknologi Bandung
Rina Indriani, S.ST	BPS Sub-directorate Tourism Statistics
Dr. Yusniar Juliana Nababan, SSi., MIDEC	BPS of West Kalimantan Province
George Hodge	Pulse Lab Jakarta

Group 3 – Safe and Affordable Drinking Water for All:
A Development of a SDGs Proxy Indicator from MDGs Indicators

Dedi Cahyono, SE, MA, MSE	BPS of West Papua Province
Dr. Dedy Dwi Prastyo	Sepuluh Nopember Institute of Technology
Imam Safawi, S.Si., M.Si	Sepuluh Nopember Institute of Technology
Dr. Nanang Susyanto, S.Si, M.Sc	Gajah Mada University
Muhammad Rheza	Pulse Lab Jakarta

Group 4 – Ensuring the Quality of Data: No Accessibility to Raw Data

Dr. Budi Warsito, S.Si, M.Si	Diponegoro University
Hertina Yusnissa, S.ST, MM	BPS of West Nusa Tenggara Province
Marselinus Ulu F., S.Si, M.Sc	Institute of Resource Governance and Social Change
Sri Astuti Thamrin, S.Si, M.Stats, PhD	Hasanuddin University
Muhammad Subair	Pulse Lab Jakarta

Group 5 – Spatial Disaggregation of MDGs Indicator with Numerical Method Approach

Dr. Agus Mohamad Soleh, S.Si, MT	Bogor Agricultural University
Qurratul Aini, S.ST, M.Sc	BPS of West Nusa Tenggara Province
Syarifah Diana Permai, S.Si., M.Si	Bina Nusantara University
Dr. Utriweni Mukhaiyar	Institut Teknologi Bandung
Ni Luh P. Satyaning Paramita	Pulse Lab Jakarta

Table of Contents

From MDGs to SDGs: The Data Challenges

Dikara Alkarisya, Ni Luh Putu Satyaning P. P, Zakiya Aryana Pramestri
Pulse Lab Jakarta
{dikara.alkarisya, ni.paramita, zakiya.pramestri} @un.or.id

1 INTRODUCTION

Continuing the success of the Millennium Development Goals (MDGs), in 2015, world leaders adopted a set of goals to end poverty, protect the planet, and ensure prosperity for all as part of a new sustainable development agenda: the Sustainable Development Goals (SDGs). The SDGs contain 17 goals to be achieved by 2030, which are more ambitious and comprehensive compared to the 8 MDGs. These goals include 169 targets and 230 indicators, with additional country-specific indicators.

One of the key features of the SDGs is the commitment to " leave no one behind" based on the experience with the MDGs. Although substantial progress has been made on many of the MDGs, the progress has been uneven across regions and countries. Millions of people are being left behind, especially the poorest and most vulnerable groups because of their gender, age, disability, ethnicity or geographic location. Therefore, the SDGs require more disaggregated data by those demographic characteristics to monitor the achievement of the indicators for all groups.

Measuring the SDG indicators will be very challenging for some countries, especially for developing countries where rural areas may be difficult to access, including Indonesia. Some indicators are not easy to measure without adequate tools and technology. The important message from the MDGs is that the lack of reliable data can undermine the government's ability to set goals, optimize investment decisions and measure progress[1].

In this paper, we describe the MDGs, SDGs, and the challenges regarding data and monitoring systems to support implementation of the SDGs. We also describe the MDG datasets provided to Research Dive participants, and the characteristics.

2 MDG AND SDG

2.1 Millennium Development Goals

The study has been implemented on 3 data sets (year 2007, 2011 and 2014) of 22 MDGs indicators based on 33 provinces in Indonesia. In 2000, world leaders gathered at the Millennium Summit and committed their nations to a new global partnership in order to alleviating poverty and set out a series of time-bound targets to be achieved by 2015. These are known as the Millennium Development Goals (MDGs).

Referring to Indonesia's MDGs Report 2014 [1], the country achieved the following MDG indicators: halve the proportion of the population below one dollar per day (Goal 1), balance the ratio of girls and boys in primary, secondary and tertiary education, as well as literacy rates (Goal 3), reduce Tuberculosis prevalence (Goal 6), increase the ratio of actual forest cover to total land area and increase the proportion of households with sustainable access to basic sanitation in urban and rural areas (Goal 7), and increase the proportion of population with cellular phones (Goal 8).

Goal 1.	Eradicate extreme poverty and hunger
Goal 2.	Achieve universal primary education
Goal 3.	Promote gender equality and empower women
Goal 4.	Reduce child mortality
Goal 5.	Improve maternal health
Goal 6.	Combat HIV/AIDS, malaria, and other diseases
Goal 7.	Ensure environmental sustainability
Goal 8.	Develop a global partnership for development

On the other hand, there are several indicators still require special attention, such as reducing the number of people living under the national poverty line and below minimum level of dietary energy consumption (Goal 1), infant mortality rate (Goal 4) and maternal mortality rate (Goal 5), ensuring comprehensive knowledge on HIV/AIDs (Goal 6), reducing carbon dioxide (CO2) emission, increasing proportion of households with sustainable access to improved water source, basic sanitation (Goal 7), personal computers, internet access and increase the ratio of exports and imports (Goal 8).

2.2 Sustainable Development Goals

Building on the MDGs, world leaders at Rio+20 (the 2012 UN Conference on Sustainable Development) agreed to put forward a new vision of eradicating extreme poverty by 2030[2]. The UN Secretary General in his synthesis report for the post-2015 sustainable development agenda proposed an integrated set of six essential elements: dignity, people, prosperity, the planet, justice and partnerships [2].

The new agenda aims to balance the environmental, social, and economic dimensions of sustainable development [3]. It is also driven by five major transformations: 1) leave no one behind; 2) put sustainable development at the core; 3) transform economies for jobs and inclusive growth; 4) build peace and effective, open and accountable institutions for all; 5) forge a new global partnership [4]. Incorporating these principal and essential elements in the framework, the SDGs officially commenced on 1 January 2016 and include 17 goals and 169 targets.

SDGs implementation requires a rigorous monitoring system. Increased access to detailed information is needed to ensure that no group is left behind. To have such information, data gathered will need to be disaggregated by gender, geography, income, disability, and other categories. The Secretary-General's High-Level Panel Of Eminent Persons On The Post-2015 Development Agenda also declared the need for a data revolution, to improve the quality of statistics and information available to people and governments [4].

The translation of targets, indicators and data to implement the SDGs has been challenging for countries, including Indonesia. The SDGs advise countries to collect 230 indicators

[1] http://unsdsn.org/wp-content/uploads/2015/04/Data-For-Development-An-Action-Plan-July-2015.pdf

[2] http://www.un.org/sustainabledevelopment/

Goal 1. End poverty in all its forms everywhere	Goal 10. Reduce inequality within and among countries
Goal 2. End hunger, achieve food security and improved nutrition and promote sustainable agriculture	Goal 11. Make cities and human settlements inclusive, safe, resilient and sustainable
Goal 3. Ensure healthy lives and promote well-being for all at all ages	Goal 12. Ensure sustainable consumption and production patterns
Goal 4. Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all	Goal 13. Take urgent action to combat climate change and its impacts
Goal 5. Achieve gender equality and empower all women and girls	Goal 14. Conserve and sustainably use the oceans, seas and marine resources for sustainable development
Goal 6. Ensure availability and sustainable management of water and sanitation for all	Goal 15. Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss
Goal 7. Ensure access to affordable, reliable, sustainable and modern energy for all	Goal 16. Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all level
Goal 8. Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all	Goal 17. Strengthen the means of implementation and revitalize the global partnership for sustainable development
Goal 9. Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation	

(even more when country-specific indicators are included, for instance, Indonesia has a total of 241 indicators).

The Government of Indonesia undertook preliminary data mapping and found that for national indicators, 67.8% are most ready, 26.93% are ready, and 5.26% are not ready[3]. Meanwhile, for the global indicators, 32.27% are most ready, 26.36% are ready, and 36.36% are not ready [5]. Some of the key challenges are related to data availability, quality, and validity. The government also has to deal with incomplete data in terms of inadequate historical data series and disaggregated data.

3 RESULTS AND DISCUSSION

The MDG dataset covers a 14-year series, from 2001-2014. In Indonesia, the MDG data have been collected by the National Bureau of Statistics (BPS), with the support of the ministries and other agencies. The data cover the eight MDG goals, which are broken down into 82 indicators.

The MDG data are provided for different administrative levels, which include country level, province level, and district level. The completeness of data is different for each level.

For country-level data, 21% of indicators have a complete 14-year set of data, while 51% of indicators only have data for less than seven years. Some indicator data are only available in certain years when the surveys were conducted.

For province-level data, data for only 40 indicators are available out of 82 indicators. Most of the unavailable indicators are for Goal 4. Reduce child mortality, Goal 5. Improve maternal health, and Goal 6. Combat HIV/AIDS, malaria and other diseases. Indicators for Goal 8. Develop a global partnership for development are available from 2006.

Aside from incompleteness in terms of historical series of data, there is also missing data for provinces due to regional expansion or formation of new provinces and disaster disruption. Data for Kalimantan Utara were not available for the MDGs since the province was newly established in 2012.

Kepulauan Riau was established in 2002, while the data are available from 2005. Sulawesi Barat and Papua Barat were established in 2004 and 2006 respectively, and the data are available from 2006. Also, there is almost no data for Aceh in 2005 because surveys were not conducted due to the tsunami disaster at the end of 2004.

The district-level data were only provided for 25 indicators, from 2011 to 2013. There is no data for the indicators of Goal 6. Combat HIV/AIDS, Malaria, and other disease.

4 CONCLUSIONS

The implementation of the SDGs in Indonesia brings challenges, especially in terms of the monitoring system. One third of the global indicators are not yet ready for measurement, which confirms the need to meet SDG data requirements to produce reliable information.

During the Research Dive, we addressed these challenges by creating opportunities for academia to explore the publicly shared MDG datasets to support the the monitoring system of the SDGs in Indonesia. The participants were tasked with developing statistical methodologies and analysis, for instance, to investigate the correlation and causation, to develop proxy indicators, to improve quality of data, and to develop data disaggregation methods.

REFERENCES

[1] Ministry of National Development Planning/National Development Planning Agency (BAPPENAS). 2014. Report on Achievement of The Millennium Development Goals in Indonesia 2013

[2] UN (United Nations). 2014. The Road to Dignity by 2030: Ending Poverty, Transforming All Lives and Protecting the Planet. http://www.un.org/ga/search/view_doc.asp?symbol=A/69/700&%3bamp%3bLang=E

[3] United Nations Environment Programme (UNEP). 2013. Embedding the Environment in Sustainable Development Goals. https://sustainabledevelopment.un.org/index.php?page=view&type=400&nr=972&menu=35

[4] UN (United Nations). 2013. A New Global Partnership: Eradicate Poverty and Transform Economics Through Sustainable Development. United Nations Publication, New York. https://sustainabledevelopment.un.org/content/documents/8932013-05%20-%20HLP%20Report%20-%20A%20New%20Global%20Partnership.pdf

[5] UNDP Indonesia. 2016. Indicators and Data Mapping to Measure Sustainable Development Goals (SDGs) Targets: Case of Indonesia 2015. http://www.id.undp.org/content/dam/indonesia/2016/doc/SDGs%20Indicators%20and%20Data%20Mapping%20in%20Indonesia_UNDP%20UNEP%202015.pdf?download

[3] Most ready: data is available in good quality, Ready: some data are available but still need adjustment, or data is not well integrated, or only available in national level, Not ready: data is not available

Identifying correlation structures within and between MDGs Goals

A. A. R. Fernandes
Brawijaya University
Malang, Indonesia
fernandes@ub.ac.id

R. Fajriyah
Universitas Islam Indonesia
Yogyakarta, Indonesia
rfajriyah@uii.ac.id

D. F. Aksioma
Institute Teknologi Sepuluh November
Surabaya, Indonesia
diaz_fa@statistika.its.ac.id

L. Anisah
BPS Semarang City
Semarang, Indonesia
lilis.anisah@bps.go.id

I. Amin
Pulse Lab Jakarta
Jakarta, Indonesia
imaduddin.amin@un.or.id

ABSTRACT

Millenium Development Goals (MDGs) has established more than a decade ago. Evaluating the achievements of the targets means evaluating its indicators as well. The useful evaluation could be based on the exploration of the correlation structures among the indicators. This can be measured with computing their correlation values and implementing the factor and cluster analysis. The impact of the available correlation structures can be revealed through the bi- plot of principal component analysis. Using the MDGs data of Indonesia in 2007, 2011 and 2014 at province level, we discovered that there is a correlation among the MDGs indicators and Goals. Most of the structures are inline to the target type of the MDGs indicators. Further results show that the MDGs achievements across provinces are varies through the years. But Papua, Papua Barat and Nusa Tenggara Timur are remained as the least achiever than other provinces.

KEYWORDS

MDGs indicators, Correlation, Factor Analysis, Clustering Analysis, Principal Component Analysis

1 INTRODUCTION

The Millennium Development Goals (MDGs) addressed the extreme poverty and the basic human right of each person on earth. Eradicating poverty means someone needs to work on many aspects. Because poverty is not a standalone problem, it is a multidi- mensional one. The MDGs contains 8 goals which is measured by different indicators. In total there are 48 indicators to measure the progress of MDGs. Among these MDGs indicators it is very useful to discover the statistically meaningful correlation among them.

The correlation will lead us to the correlation structures on them and measure the effect of multicollinearity toward the results of the analysis. Knowing the correlation structures on poverty and human basic rights aspects will lead to the statistically meaningful information to measure the progress and achievement of the MDGs target. For instance, comparing the MDGs target's achievement among provinces with cluster analysis (CA) and or principal componenet analysis (PCA). In CA and PCA, the variables have to be uncorrelated. Otherwise, there are some consequences with it. The effects of multicollinearity in clustering analysis have been demonstrated by Sambandam [7].

This paper will be divided into three parts. The first section is Introduction, the second one is the Research Methodology, the third one Result and Discussion and followed by the Conclusion and Remarks.

2 RESEARCH METHODOLOGY

2.1 Data

The study has been implemented on 3 data sets (year 2007, 2011 and 2014) of 22 MDGs indicators based on 33 provinces in Indonesia.

2.2 Data Analysis Methods

For a very complex data set, data matrix, usually is chosen to represent multiple items or variables. Some statistical methods, such as factor, principal component and cluster analysis, could be implemented to reduce the complexity of the observed data. Those methods will lead to a discovery of the correlation structures among variables interest. n this paper we will use factor analysis (FA), CA and PCA. First, the correlation values is needed to be computed.

Correlation is defined as a measured of a linear relationship degree between two variables. Collinearity is a high level of correlation and when it is more than two variables then it is called multicollinearity [7]. There are different types of correlation, in this paper we focus on the Pearson's correlation.

Suppose we have two observed variables X_i and $Y_i, i = 1, 2, \ldots, n$. The linear correlation between X and Y can be defined as follows

$$Cor(X,Y) = \frac{\left(\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)\right)}{\sqrt{\left(\sum_{i=1}^{n} {X_i}^2 - (\sum_{i=1}^{n} X_i)^2\right)\left(\sum_{i=1}^{n} {Y_i}^2 - (\sum_{i=1}^{n} Y_i)^2\right)}}$$

Let we have a large set of variables. The number of correlations among variables is definitely many, very complex and it is not an easy task to describe the correlation's pattern. In this case we will need a small underlying factors to account for the main source of variation or the pattern of correlation [4].

Some literatures ([1], [5], [3] and [4]) agree that one of the methods to model a large set of variables (multivariate observation) is factor analysis. It is a method to model the observed variables based on the latent factors. These latent factors are unobserved and have a smaller number than the observed variables.

To determined how small the number of factors is, the 3 aspects below could be used as a consideration

1. The cumulative proportion is at least 0.8
2. Eigenvalues at least 1
3. The elbow of the scree plot

The PCA is one of the oldest dimensionality reduction techniques in case of the large observed variables are highly correlated (multicollinearity). It aims to produce a linear combinations of the observed variables which are uncorrelated each other ([7],[1], [5], [3] and [4]). Suppose we have k variables X with n observations from each. Applying the PCA means the new variable Y is based on the linear combination of $X_1, X_2, \ldots, X_k$.

Once the Principal Components has been formed than the biplot could be used to map the correlation structures of the variables related to the observations. Greenacre [2] defines biplot as the generalization of the scatterplot of observations on two vari- ables. Biplot is a useful tool to explore the correlation pattern among variables or the similarities among observations. The smaller the deviation between the arrow line the higher correlation between these two variables. The object positions on the below (left side) of zero implies that it still need to have an improvement on some specific variable.

Cluster analysis is a technique used for combining observations into groups or clusters such that:

1. Each group or cluster is homogeneous or compact with respect to certain char- acteristics. That is, observations in each group are similar to each other
2. Each group should be different from other groups with respect to the same characteristics. That is, observations of one group should be different from the observations of other groups.

The definition of similarity or homogeneity varies from analysis to analysis, and depends on the objectives of the study. Cluster analysis groups observations such that the observations in each group are similar with respect to the clustering variables. It is also possible to cluster variables such that the variables in each group are similar with respect to the clustering observations. Geometrically, this is equivalent to representing data in an n-dimensional observation space, and identifying clusters of variables. This objective of cluster analysis appears to be similar to that of factor analysis. Recall that in factor analysis we attempt to identify clusters of variables such that the variables in each cluster have something in common; i.e., they appear to measure the same latent factor. It is therefore possible to use factor analysis to cluster observations, and to use cluster analysis to cluster variables ([8] and [6]).

Algorithms designed to perform cluster analysis are usually divided into two broad classes called hierarchical (agglomerative and diversive) and non-hierarchical cluster- ing methods. The agglomerative hierarchical procedures fall into three broad categories: Linkage (single, complete, average and farthest), Centroid, and Error Variance methods. Among these procedures, only linkage algorithms may be used to cluster either objects (items) or variables. The other two methods can be used to cluster only objects. Non-hierarchical methods may only be used to cluster items ([9] and [6]).

3 RESULTS AND DISCUSSION

3.1 Correlation

The initial work, computing the correlation among MDGs indicators, needs to be performed before further analysis is carried out. We summary the results from the Pearson correlation computation in the subsequent graphs as in Figure 3.1.

Based on the computation of the Pearson correlation among variables of the MDGs indicators, we can see that there is a correlation among them. Some of the correlation degree are positively and negatively high which is indicated by the thick dark green and magenta colours respectively, some of them are not which is indicated by the thin/lighter dark green and magenta colours.

From Figure 3.1 - 3.3 we also can see that the correlation is happened within and between goals of MDGs. It can be seen from the variable labels on the graphs. This is only the first investigation and to find the correlation structures among them precisely we need to continue the analysis by implementing factor and clustering analyses.

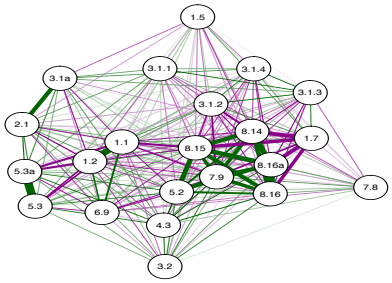


Figure 1a. Correlation structures among MDGs indicators in 2007 data set

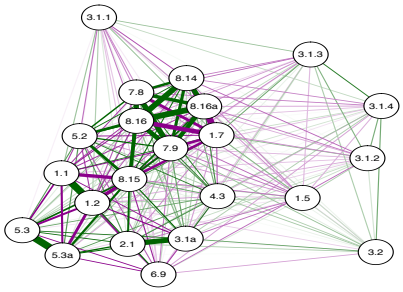


Figure 1b. Correlation structures among MDGs indicators in 2011 data set

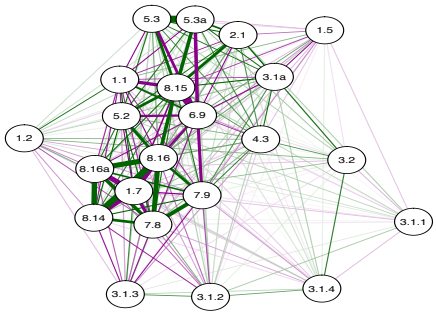


Figure 1c. Correlation structures among MDGs indicators in 2014 data set

3.2 FA Results

The preliminary studies with factor analysis suggest that the optimum factors for MDGs Indonesia data at province level for year 2007, 2011 and 2014 are 6, 5 and 6 respectively. Although the psych package with the fa.parallel function suggests 2 factors.

The underlying structures or construction among the MDGs indicators based on factor analysis are as in Figure 2.

The MDGs target types have been defined as decrease, increase or halted. Table 1 shows that mostly the correlation structures are inline with the MDGs target type, except the halted one. Some indicators, in some years, have different indication of target type. One reason about this is related into the data collection and their completeness. For instance, the indicator about the proportion of population using an improved drinking water source, it is an increase target type. When we used the old method on how to compute the target, the result is the decrease target type. Once we applied the new method, it becomes clear as an increase target type. Therefore, it might depend on how accurate the method measured the indicator.

The implementation of hierarchical factor analysis (omega) shows in Figure 3.5. This results show that it is possible there is

Table 1. The indicators behavior based on factor analysis

MDGs indicator	Indicator name	Target type	2007	2011	2014
1.1a	Proportion of population below national poverty line	Decrease	Decrease	Decrease	Decrease
1.2	Poverty gap ratio	Decrease	Decrease	Decrease	Increase
1.5	Employment-to-population ratio	Increase	Decrease	Decrease	Decrease
1.7	Proportion of own-account and contributing family workers in total employment	Decrease	Decrease	Decrease	Decrease
2.1	Net enrolment ratio in primary education	Increase	Increase	Increase	Increase
3.1.1	Ratio of girls to boys in primary schools	Increase	Decrease	Decrease	Increase
3.1.2	Ratio of girls to boys in Junior high school	Increase	Decrease	Increase	Decrease
3.1.3	Ratio of girls to boys in Senior high school	Increase	Decrease	Increase	Decrease
3.1.4	Ratio of girls to boys in higher education	Increase	Increase	Increase	Increase
3.1a	Literacy ratio of women to men in the 15-24 age group	Increase	Increase	Increase	Increase
3.2	Share of women in wage employment in the non-agricultural sector	Increase	Increase	Increase	Increase
4.3	Proportion of 1 year-old children immunised against measles	Increase	Increase	Increase	Increase
5.2	Proportion of births attended by skilled health personnel	Increase	Increase	Increase	Increase
5.3	Contraceptive prevalence rate	Increase	Increase	Increase	Increase
5.3a	Current contraceptive use among married women 15-49 years old, modern method	Increase	Increase	Increase	Increase
6.9a	Incidence rates associated with Tuberculosis (all cases per 100,000 people per year)	Halted	Decrease	Decrease	Decrease
7.8	Proportion of population using an improved drinking water source	Increase	Increase	Increase	Increase
7.9	Proportion of population using an improved sanitation facility	Increase	Increase	Increase	Increase
8.14	Proportion of the population with fixed-line telephones (teledensity in population)	Increase	Increase	Increase	Increase
8.15	Proportion of population with cellular phones	Increase	Increase	Increase	Increase
8.16	Proportion of households with access to internet	Increase	Increase	Increase	Increase
8.16a	Proportion of households with personal computers	Increase	Increase	Increase	Increase

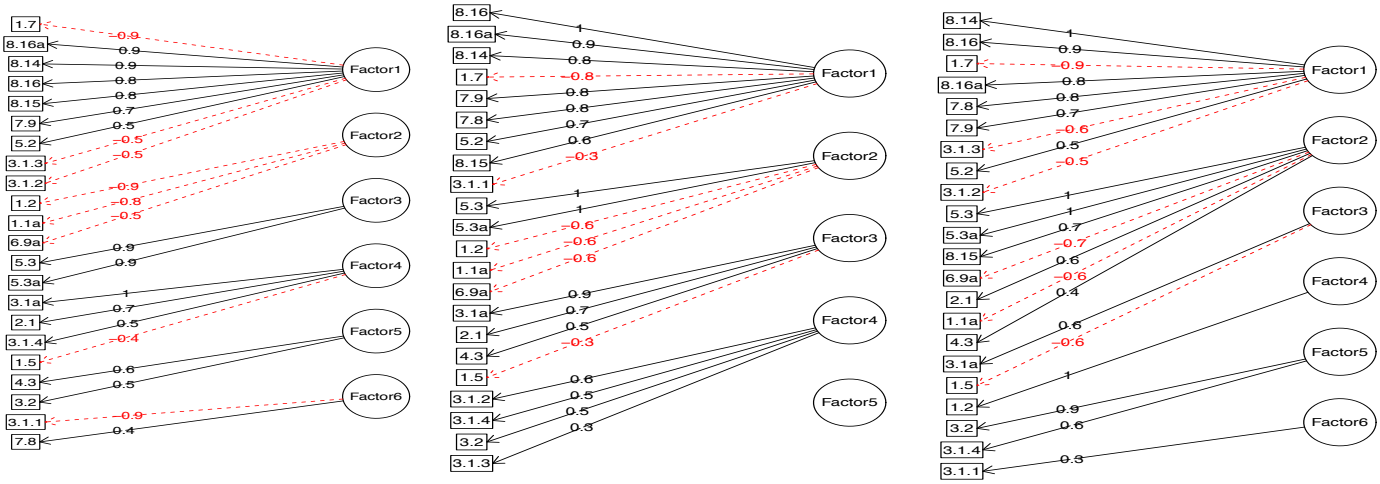


Figure 2. Underlying structures among MDGs indicators in 2007, 2011, and 2014 data sets

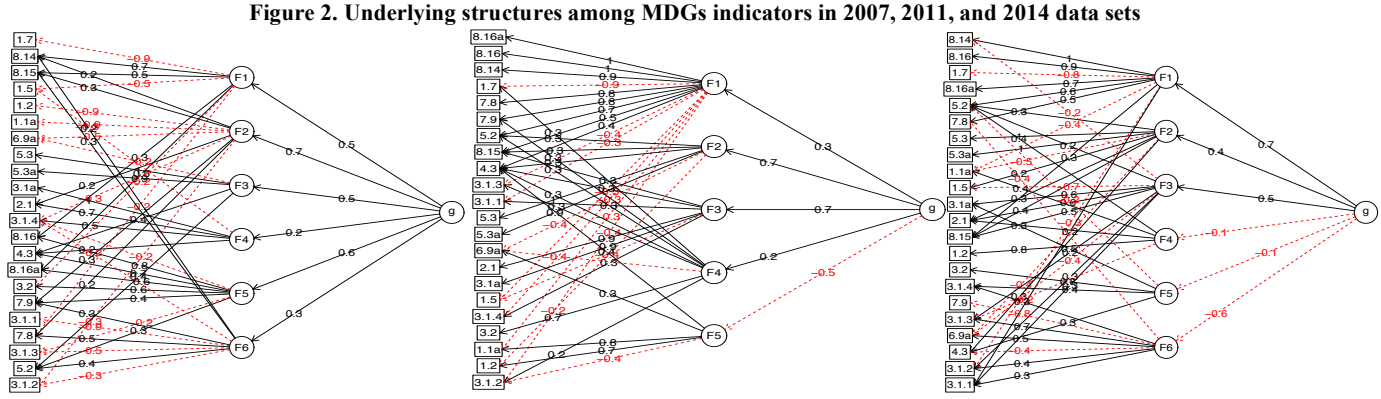


Figure 3. Hierarchical factor analysis of MDGs indicators in 2007, 2011, and 2014 data sets

available the general factor among the MDGs indicators and goals. These results and the results from factor analysis can be further investigated by the structural equation modelling (SEM).

3.3 Clustering Result

Another method to discover the correlation structures among variables is clustering analysis, i.e., item cluster analysis, iclust. In this case, instead of clustering the objects, it clusters the variables. This method will reduce the complexity of data.

The results of iclust implementation can be seen in Figures 4. In this method, we do not specify how many clusters have to be provided rather than follow the algorithm to provide the optimum cluster. The same as factor analysis at psych package and fa.parallel function, the iclust provides two clusters of indicators as well.

The results from iclust are similar to the factor analysis. But, the percentage similar- ity to the MDGs target type are greather than factor analysis. Some indicators which have to be the decrease (increase) type become the increase (decrease) type and in the structure is followed by the decrease (increase) one. For instances the indicators 1.1a nd 1.2 in 2011, indicator 1.5 in 2014, indicators 6.9a and 7.8.

Table 2. The indicators behavior based on *iclust*

MDGs indicator	Indicator name	Target type	2007	2011	2014
1.1a	Proportion of population below national poverty line	Decrease	Decrease	Increase/Decrease	Decrease
1.2	Poverty gap ratio	Decrease	Decrease	Increase/Decrease	Increase
1.5	Employment-to-population ratio	Increase	Decrease	Increase	Decrease/Increase
1.7	Proportion of own-account and contributing family workers in total employment	Decrease	Decrease/Increase	Decrease	Decrease
2.1	Net enrolment ratio in primary education	Increase	Increase	Increase	Increase
3.1.1	Ratio of girls to boys in primary schools	Increase	Increase	Increase	Decrease
3.1.2	Ratio of girls to boys in Junior high school	Increase	Increase	Decrease/Increase	Increase,Decrease
3.1.3	Ratio of girls to boys in Senior high school	Increase	Increase	Increase	Increase,Decrease
3.1.4	Ratio of girls to boys in higher education	Increase	Increase	Increase	Increase
3.1a	Literacy ratio of women to men in the 15-24 age group	Increase	Increase	Increase	Increase
3.2	Share of women in wage employment in the non-agricultural sector	Increase	Increase	Increase	Increase
4.3	Proportion of 1 year old children immunised against measles	Increase	Increase	Increase	Increase
5.2	Proportion of births attended by skilled health personnel	Increase	Increase	Increase	Increase
5.3	Contraceptive prevalence rate	Increase	Increase/Decrease	Increase	Increase
5.3a	Current contraceptive use among married women 15-49 years old, modern method	Increase	Increase,Decrease	Increase	Increase
6.9a	Incidence rates associated with Tuberculo- sis (all cases per 100,000 people per year)	Halted	Halted	Increase	Decrease/Increase
7.8	Proportion of population using an improved drinking water source	Increase	Decrease/Increase	Increase	Increase
7.9	Proportion of population using an improved sanitation facility	Increase	Increase	Increase	Increase
8.14	Proportion of the population with fixed-line telephones (teledensity in population)	Increase	Increase	Increase	Increase
8.15	Proportion of population with cellular phones	Increase	Increase	Increase	Increase
8.16	Proportion of households with access to internet	Increase	Increase	Increase	Increase
8.16a	Proportion of households with personal computers	Increase	Increase	Increase	Increase

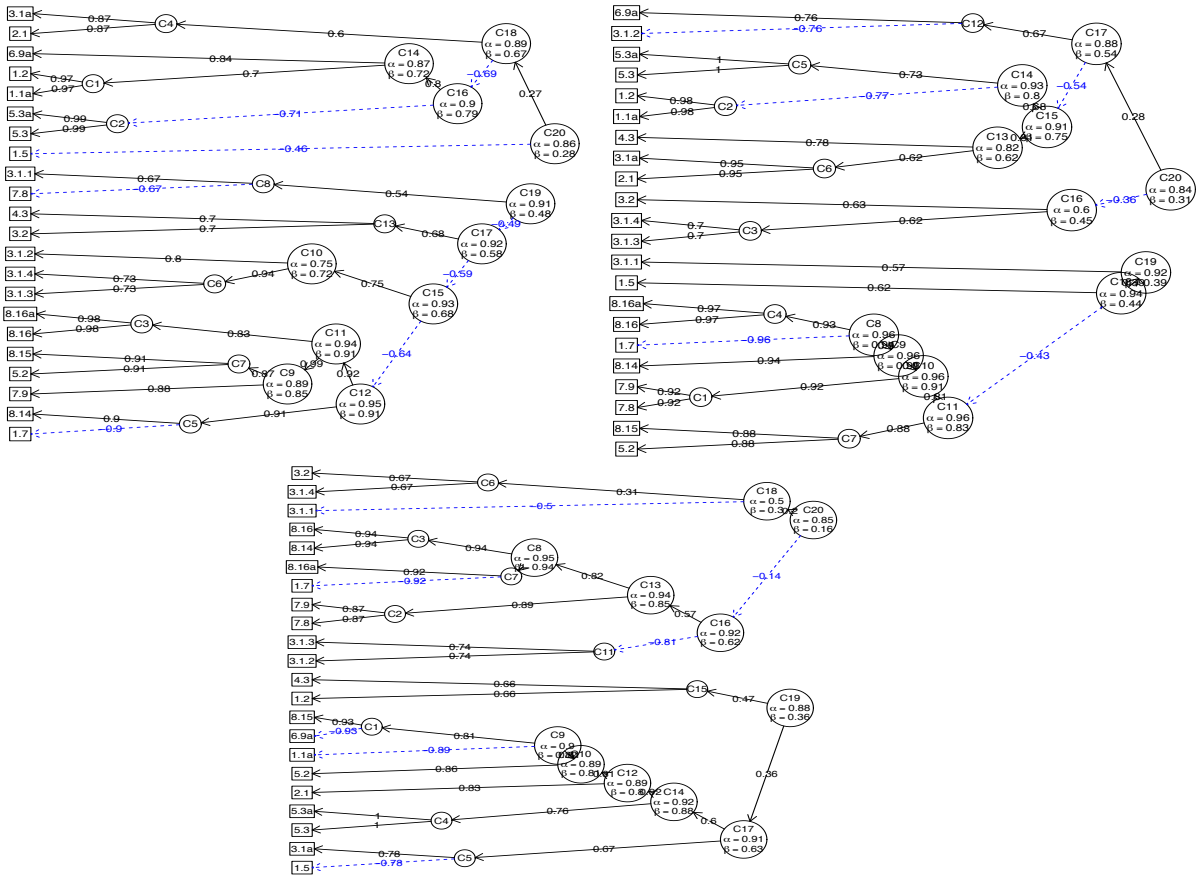


Figure 4. Cluster analysis of MDGs indicators in 2007, 2011, and 2014 data sets

Some indicators, such as 1.5 in 2011, 3.11 in 2011 and 6.9a in 2011, have a totally different target type. One has find a time to take a closer look on how the data is gathered. The summary of the indicators behaviour based on the iclust analysis can be seen at Table 2.

3.3 Biplot Principal Component Analysis

Previous subsections 3.2 and 3.3 suggest that there are two underlying factors among MDGs indicators, with the possibility maximum 6, 5 and 6 respectively for year 2007, 2011 and 2014. This result can be used to perform the biplot principal component analysis to assess the temporal analisis of the MDGs indicators. We show the 3 and 2 dimensions of the biplot Principal Components (PCs) from the PCA in Figures 5.

Based on Figures 5 we can infer that

1. In general, there are 3-4 groups of provinces in the target's achievement. The best target's achievement provinces are DKI, DIY, Kepulauan Riau, Bali and Banten. Banten joins the group since 2011. The least target's achievement proinces are NTB, Maluku, NTT, Papua Barat and Papua. Since 2011 NTB has moved on to the group with a better targets achievement
2. The indicators 8, 7, 6, 5 and 1 tend to have a very high variation across years. Indicator 3 has high variation in 2007 only
3. Throughout the years, provinces at the least target's achievement group are as- sociated with variables 1.2, 6.9a, 1.1a, 1.5 and 1.7. On the other hand provinces at the best target's achievement group are associated more to the variables 7.8, 7.9, 8.14, 8.16 and 8.16a. Although in year 2007 and 2011, variables 5.2, and 8.15 are also associated with this group

4 CONCLUSIONS

Statistical methods, such as factor and clustering analysis can be used to discover the correlation structures among variables with complexity. We have investigated the correlation structures among the MDGs indicators and discover some structures on them. Mostly, the structures are inline with the MDGs indicators target type.

Biplot PCA shows that in general there are no provinces in Indonesia have exceeded the MDGs target. But there are 5 provinces which can be categorized as best MDGs target's achievement, namely: DKI, DIY, Kepulauan Riau, Kaltim and Bali since 2007. And there are prominent provinces which can be categorized as least MDGs targets achievement, namely NTT, Maluku, Papua and Papua Barat.

Further study with different statistical analysis methods, such as SEM, path and longitudinal analysis can be conducted to find a comprehensive functional relationship within and between MDGs goals.

REFERENCES

[1] Timothy A. Brown. 2006. *Confirmatory Factor Analysis for Applied Research. Methodology in the Social Sciences.* The Guilford Press
[2] Michael Greenacre.2010. *Biplots in* Practice. Fundación BBVA,Barcelona.
[3] Wolfgang Karl Härdle and Léopold Simar. 2012. *Applied Multivariate Statistical Analysis, 3 edition*. Springer.
[4] Keenan A. Pituch and James P. Stevens. 2016. *Applied multivariate Statistics for the Social Sciences, Analyses with SAS and IBMs SPSS, 6 edition*. Routledge.
[5] Alvin C. Rencher. 2002. *Methods of Multivariate Analysis*. Wiley.
[6] William Revelle. 1979. Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioural Research.* 14:57-74.
[7] Rajan Sambandam. 2003. Cluster Analysis gets complicated. *Marketing Research,* !5(1).
[8] Subhash Sharma. 1996. *Applied Multivariate Techniques*. Wiley
[9] Neil H. Timm. 2002. *Applied Multivariate Analysis*. Springer

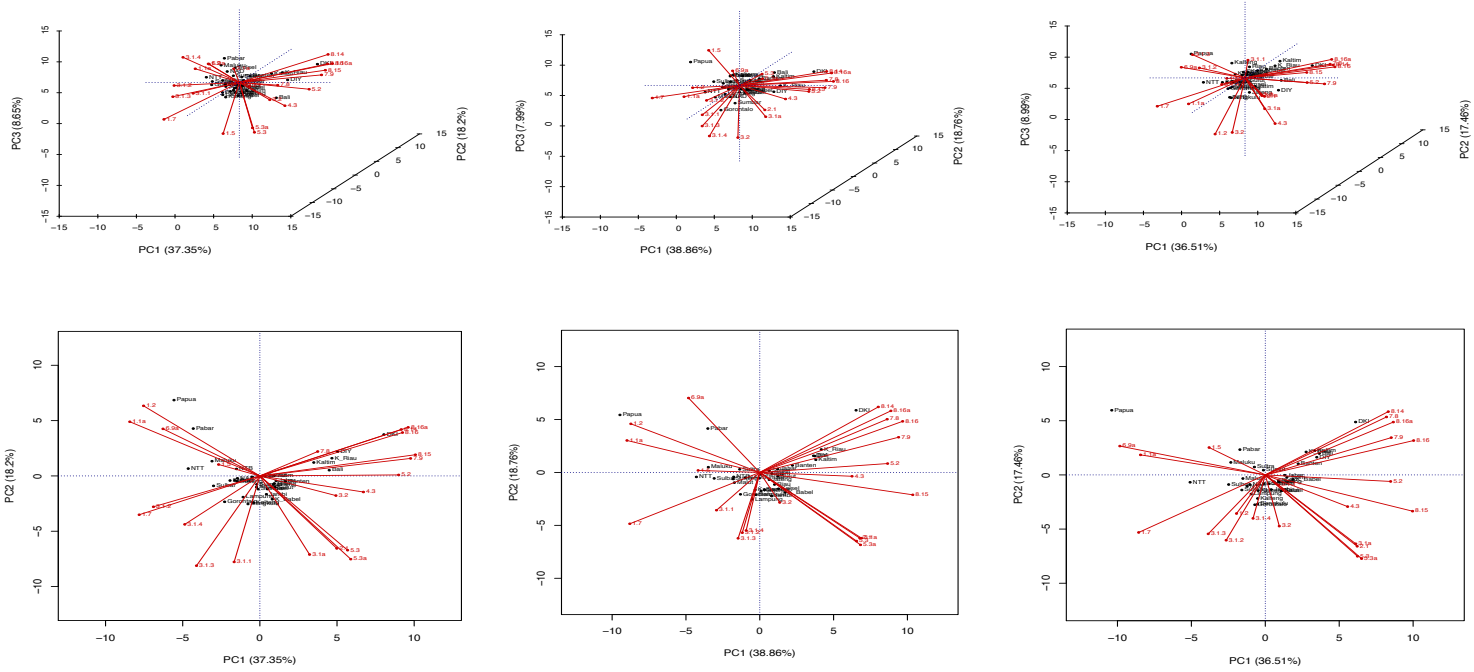


Figure 5. Biplot of MDGs indicators in 2007, 2011, and 2014 data sets

The Impact of Health and Education on Poverty Reduction: A Causal Analysis

A. Efendi
Brawijaya University
Malang, East Java, Indonesia
a_efendi@staff.ub.ac.id

Y. Nababan
BPS Kalimantan Timur Province
West Kalimantan, Indonesia
yusniar@bps.go.id

K. N. Sari
Institute of Technology Bandung
Bandung, West Java, Indonesia
kurnia@math.itb.ac.id

R. Indriani
BPS-Statistics Indonesia
Jakarta, Indonesia
rina.indriani@bps.go.id

G. Hodge
Pulse Lab Jakarta
Jakarta, Indonesia
george.hodge@un.or.id

ABSTRACT

The complexity of the relationships between the variables of interest and in some circumstances access to improved drinking water sources, access to improved sanitation facilities, enrollment in primary education, and improved literacy can impact upon income poverty, although at different timescales. It is evidence of this causal link in the context of Indonesia that the authors of this paper propose to investigate. Firstly, this research aims to confirm whether education and health give impact on poverty reduction. Secondly, its objective is to determine which indicators have the greater influence on poverty reduction. In addition, this research also identifies whether the impact of education and health change over time and space in reducing poverty

There are impacts of education and health on poverty reduction. Education indicator has more impact than health indicator. Estimated models suggest that reducing poverty can be carried out by implementing policy in education health sectors. Moreover, spatial effect is statistically significant. It indicates that local specific policy would be needed given that there are differences on the infrastructure of health and education between regions. The magnitude of estimates show that more effort should be put in regencies, compared to cities, in order to reduce poverty in the regions.

KEYWORDS

Poverty, health, education, causation, regression

1 INTRODUCTION

The theoretical relationships between safe water, sanitation, education and poverty are documented by Asselin and exemplified through case studies in the same volume.[1] Specific to Indonesia, Teguh and Nurkholis find that important factors of poverty dynamics include educational attainment and health shocks, among others. Furthermore, concerning spatial dynamics, the same authors find that households located in Java and Bali are more vulnerable to negative shocks than other areas due to the levels of employment in sectors other than agriculture and the lower average size of agricultural land owned by households.[2]

Specific to the impact of an absence of improved drinking water sources and access to improved sanitation facilities on income poverty, the Department for International Development of the UK Government highlights diarrheal disease as representing circa 90 percent of the avoidable disease burden prevented by good water supply, and that improved access to water facilitates hygiene and greatly facilitates the use of sanitation.[3] Furthermore, Prüss et al. estimate the disease burden from inadequate water, sanitation, and hygiene to be four percent of all deaths and almost six percent of the total disease burden in disability-adjusted life years occurring worldwide.[4] Concerning the link between incidences of disease and income poverty, the OECD and WHO find a strong link between health and livelihoods. In particular, they find that incidences of illness among poor or socially vulnerable persons can trap the entire household in a downward spiral of lost income and high healthcare costs. Connected to the interrelationship between the variables, the OECD and WHO also find that poor people are more vulnerable to this downward spiral as they are more prone to disease and have more limited access to health care and social insurance.[5]

Regarding the impact of the net enrollment ratio in primary education and the literacy rate of 15 to 24 years old on income poverty, the work of Tilak is instructive. The author highlights the interrelationship between poverty of education, including non-participation or low rates of participation of children in schooling and low rates of achievement, and income poverty,

[1] Asselin, L-M. (2009): *Analysis of Multidimensional Poverty: Theory and Case Studies*. Ottawa: Springer.
[2] Teguh, D., and Nurkholis (2013): Finding out of the Determinants of Poverty Dynamics in Indonesia: Evidence from Panel Data, *Bulletin of Indonesian Economic Studies*. [Online] available at: <http://www.tandfonline.com/doi/full/10.1080/00074918.2013.772939> [Accessed 13 March 2017].
[3] Department for International Development (2013): *Water, Sanitation and Hygiene: Evidence Paper*. Glasgow: DFID.
[4] Prüss et al. (2002): Estimating the Burden of Disease from Water, Sanitation, and Hygiene at a Global Level, *Environmental Health Perspectives*, 110:5. [Online] available at: <http://www.who.int/quantifying_ehimpacts/global/en/ArticleEHP052002.pdf> [Accessed 13 March 2017].
[5] OECD, WHO (2003): *DAC Guidelines and Reference Series: Poverty and Health*. Paris: OECD.

finding that poverty of education is a principal factor responsible for income poverty.[6] Machin and McNally also highlight this relationship finding that if government policies increase the probability of attaining appropriate educational qualifications and the employment gains translate into higher household income, then one can expect reductions in the measure of child poverty.[7] The authors referenced above also acknowledge that the link between educational enrollment and poverty reduction is not always evident. Bonal details the main failings that underlie the absence of this relationship in Latin America and attributes many of them to an underestimation of the inverse relationship, namely the effects that poverty has on education.

In sum, the literature confirms the complexity of the relationships between the variables of interest, and that in some circumstances access to improved drinking water sources, access to improved sanitation facilities, enrollment in primary education, and improved literacy can impact upon income poverty, although at different timescales. It is evidence of this causal link in the context of Indonesia that the authors of this paper propose to investigate.

There are three objectives in this research. Firstly, this research aims to confirm whether education and health give impact on poverty reduction. Secondly, its objective is to determine which indicators have the greater influence on poverty reduction. In addition, this research also identifies whether the impact of education and health change over time and space in reducing poverty.

2 MODEL AND DATA SOURCE

The relationships between access to potable water, sanitation, education enrollment, literacy levels, and income are complex, involving both vicious and virtuous circles depending on the direction of the trend. Other variables form part of the transmission mechanisms between the health and education variables, and income poverty, depending on the direction of causation. This paper uses data connected to the Millennium Development Goals (MDGs) in examining the impact of access to improved drinking water sources (MDG 7.8), access to improved sanitation facilities (MDG 7.9), net enrollment ratio in primary education (MDG 2.1), and the literacy rate of 15 to 24 years old (MDG 2.3), on the proportion of population below national poverty line[8] (MDG 1.1a) and on the poverty gap ratio (MDG 1.2) in Indonesia. During 3 years (2011-2013), the poverty ratio for 497 district in Indonesia is showed by boxplot below.

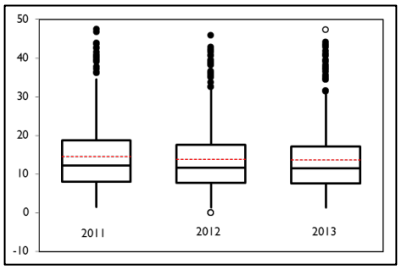


Figure 1. Boxplot for Poverty Level in 3 years (2011-2013)

From Figure 1, the poverty (POV) on district level less than 34%. The average and variance of POV are decreasing. But in every year, there are less than 5% districts become outlier e.g some district in Papua (1 district have POV more than 47%), 2 district i.e Sabu Raijua (NTT) and Lombok Utara (NTB).

There are two models that are employed to identify causal relationship, Simple Regression Linear and Multiple Regression Model. The first model is applied to recognize whether education and health give impacts on poverty reduction respectively, whereas, the second model is used to find causal effect of education and health on poverty reduction simultaneously. The models are formulated as the following.

a. Simple Linear Regression Model

The simple linear regression model is modelled the relationship between one independent (predictor) and dependent (respon) variable, with common function:

$$\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (1)$$

By using ordinary least square for estimation of parameter, the estimator for $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained below:

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XY}} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}} \quad (2)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (3)$$

Note:

Y : Dependent variable (poverty indicator)
X : Independent variable (education or health indicator)
SS : Sum of Square
$\hat{\beta}_0$: estimator of β_0 (intercept)
$\hat{\beta}_1$: estimator of β_1 (slope or coefficient of variable X)

In this model, every independent variables have each regression model. Every models have each estimators in confident interval for each variables.

b. Multiple Linear Regression Model

In multiple linear regression model, the dependent variable is poverty indicator, while the independent variables are education and health indicators. This common model is:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} \quad (4)$$

or

$$POV_i = \hat{\beta}_0 + \hat{\beta}_1 NER_i + \hat{\beta}_2 LIT_i + \hat{\beta}_3 IWA_i + \hat{\beta}_4 BS \quad (5)$$

Note:

POV : dependent variable (percentage of people living under the national poverty line)
NER : Net Enrolment Rate (NER) in primary education
LIT : Literacy Rate for people aged 15-24 years
IWA : proportion of households with sustainable access to an improved water source
BAS : proportion of households with sustainable access to basic sanitation
$\hat{\beta}_j$: estimator of parameters

[6] Tilak, J., Education and Poverty, in Melin, M., (2002): *Education – a Way out of Poverty?* Stockholm: Elanders Novum AB.
[7] Machin, S., and McNally, S., (2006): *Education and child poverty: A literature review*. York: Joseph Rowntree Foundation.
[8] World Bank (2016): *Indonesia Overview*. [Online] available at: <http://www.worldbank.org/en/country/indonesia/overview> [Accessed 13 March 2017].

Poverty is measured by the percentage of people living under the national poverty line. Education is represented by indicator of Net Enrolment Rate (NER) in primary education and Literacy Rate (LIT); while health status of households is indicated by the proportion of households with sustainable access to an improved water source (IWA) and the proportion of households with sustainable access to basic sanitation (BAS). All data are available in district and province level. This paper also modeled the impact of cities because of the different location give difference impact for poverty (POV). Cities is become dummy variable in multiple regression model.

3 RESULTS

Simple Linear Regression is employed in determining the causal effect of health and education indicators to the poverty reduction. In addition, to see the time effect, model is also estimated for each year. Models are calculated by using data of district level. R^2 of each model represents the magnitude of impact that is given by health or education in reducing poverty in a particular year. The results are summarized in Figure 2.

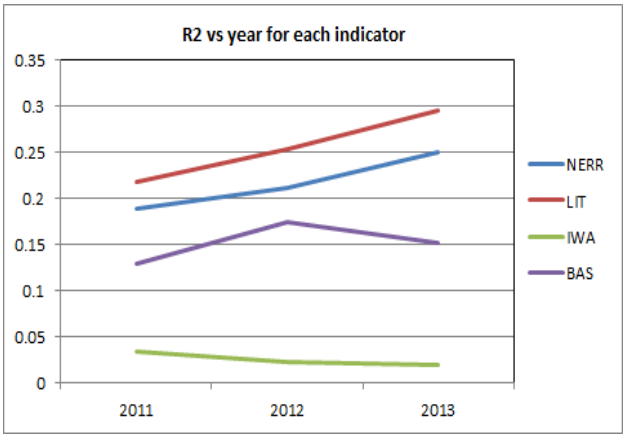


Figure 2. R^2 of Simple Linear Regression model for each indicator (NER, LIT, IWA, and BAS) to reduce poverty in a particular year (2011-2013)

For that models, R^2 are enough small that show the variance of each indicator give small impact to variance of reduce poverty. The highest R^2 only reach 30 percent in 2013. In Figure 2 reveals that in terms of education, an increase in literacy rate will give a higher impact, compared to an increase in net enrollment, in reducing poverty. The impact magnitude that is given by literacy rate is constantly higher over time. The figure also shows that the effect that is given by an increase in basic sanitation to reduce the poverty is higher than the impact from water source improvement. Similarly, the impact magnitude is constantly greater over time, although there is a slightly decrease of magnitude in 2013. This figure is also to confirm that education and health give impact on poverty reduction respectively.

Moreover, confidence interval of each estimator is calculated. By plotting them into a graph, it can be determined whether models that are estimated for respective years are significantly different in estimating the causal relationship between health or education and poverty. If the interval is intersecting, then one can conclude that the models are no significantly different in estimating the causal relationship. The confidence interval for each estimator is presented in Figure 3a to 3d.

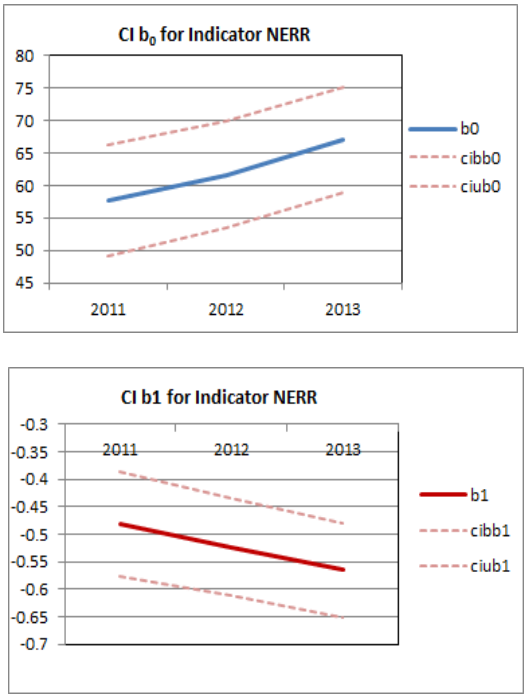


Figure 3a. The Confidence Interval fo Net Enrollment Rate (NER) Indicator

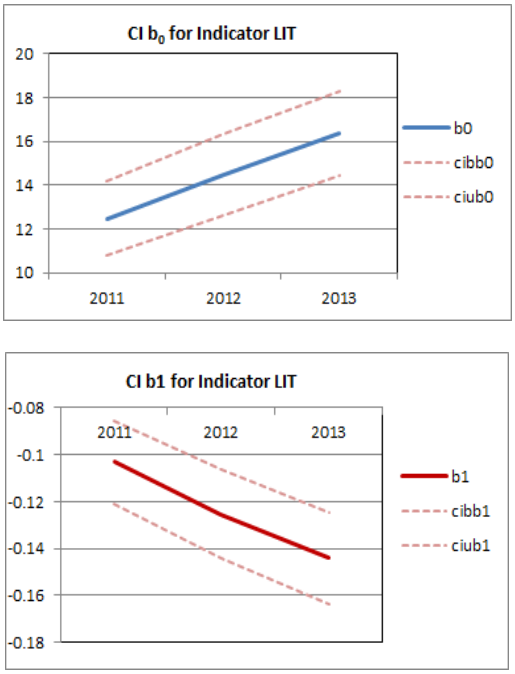
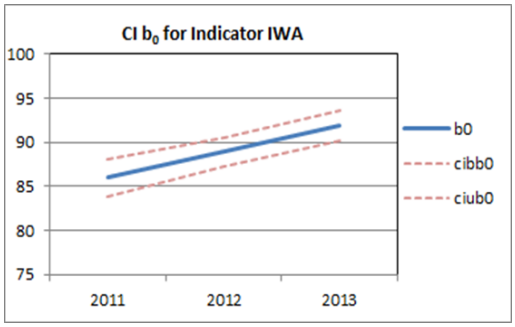


Figure 3b. The Confidence Interval for Literacy Rate (LIT) Indicator



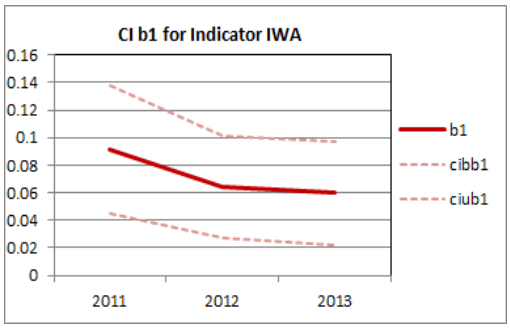
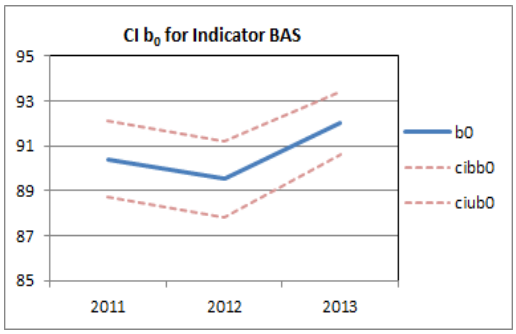


Figure 3c. The Confidence Interval fo Improved Water Source (IWA) Indicator



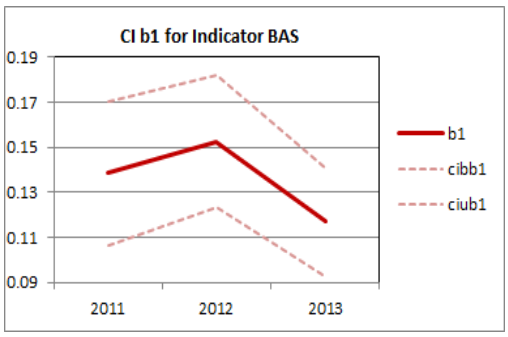


Figure 3d. The Confidence Interval fo Basic Sanitation (BAS) Indicator

The figures above show that by comparing confidence interval for respective year and identifying intersection area, the difference amongst estimators can be determined. Intersection area is indicated by drawing a horizontal line from upper (ciub) and lower limit (cibb) of confidence interval. If the line crosses other area of estimator confidence interval of other year, then it may be concluded that models are no significantly different in estimating the causal relationship between health or education and poverty reduction. In other words, the figures suggest that one may choose to use of any particular year of simple linear regression to estimate the impact of education or health on poverty reduction. From that estimator, NER and LIT give negative value but IWA and BAS give positive value for POV

Multiple Regression model is calculated to find the effect of health and education on poverty reduction simultaneously. The model is also estimated for respective year. the results is given in Table 1.

Table 1. The Result of Estimator Parameter of Multiple Regression Model with R^2 and F value

Year		B	seB	cilB	ciuB	p-value	R2	F
2011	Constant	60.300	3.551	53.321	67.279	.000	.384	130.573
	LIT	**-.419**	.039	-.495	-.343	.000		
	BAS	**-.114**	.016	-.145	-.083	.000		
2012	Constant	59.644	3.704	52.366	66.923	.000	.385	152.773
	BAS	**-.155**	.013	-.181	-.129	.000		
	NER	**-.413**	.041	-.494	-.332	.000		
2013	Constant	68.884	4.025	60.976	76.792	.000	.413	173.591
	LIT	**-.499**	.043	-.584	-.415	.000		
	BAS	**-.115**	.013	-.140	-.090	.000		

The result shows that for respective year, both education and health give impact on poverty reduction. This finding is also confirmed by negative values of b1 estimates. It implies that the better health or education condition, poverty will decline. It suggests that reducing poverty can be carried out by implementing policy in education health sectors. Moreover, given by the magnitude of the coefficient, the impact of education indicator is higher than health indicator. Furthermore, both education and health indicator have impact on poverty reduction over time. However, there is a slight difference in terms of education indicator that gives effect on poverty reduction. In 2012, education indicator included in model is different from indicator used in 2011 and 2013. It indicates that there may be a different education policy program to be applied in order to reduce poverty.

Multiple Regression model with Dummy Variable is also employed to find the spatial effect of health and education on poverty reduction simultaneously. The model is estimated for respective year. Dummy variable is use to distinguish between city and regency. Value of 1 will be assigned for city, while value of 0 is for regency. By assuming that there are prominent differences in health and education infrastructure between regions, thus one may expect that there is a different impact of health and education policy to the poverty reduction in city and regency.

Table 2. Multiple Regression Model with Cities as Dummy Variable

Year		B	seB	cilB	ciuB	p-value	R2	F
2011	Constant	60.345	3.530	53.406	67.283	.000	.396	90.049
	LIT	-.425	.039	-.501	-.349	.000		
	BAS	-.092	.018	-.128	-.055	.000		
	Cities	-2.613	1.076	-4.728	-.499	.016		
2012	Constant	60.669	3.693	53.413	67.925	.000	.385	106.256
	BAS	-.131	.016	-.161	-.100	.000		
	NER	-.433	.042	-.514	-.351	.000		
	Cities	-2.603	.892	-4.355	-.851	.004		
2013	Constant	71.502	4.048	63.549	79.455	.000	.436	75.903
	LIT	-.372	.070	-.510	-.234	.000		
	BAS	-.101	.015	-.132	-.071	.000		
	IWA	.042	.017	.008	.075	.014		
	NER	-.182	.066	-.312	-.053	.006		
	Cities	-1.850	.882	-3.583	-.117	.036		

Note: B show coefficient of each indicator, p-value give information that each coefficient are significant for that model.

The result shows that education and health policy give impact on poverty reduction over time. However, not all indicators of health and education are included into the model. In 2011, LIT and BAS are considered to give a significant impact on poverty reduction. In 2012, BAS and NER determine poverty reduction; whereas, in 2013, all indicator of health and education status are included into the model.

Individual test of dummy variable is statistically significant. This indicates that location also contributes in determining poverty reduction. Given negative values of dummy estimates, it shows that more effort should be put in regencies compared to cities if similar poverty reduction policy would be applied, so the expected result could be achieved. It also suggests that to reduce poverty, local specific policy would be needed. Moreover, the magnitudes of coefficients show that the impact of education is higher than health on poverty reduction. In 2013, more indicators are used as predictors. There are two education indicators, two health indicators, and a dummy variable.

4 CONCLUSIONS

There are impacts of education and health on poverty reduction. Education indicator has more impact than health indicator. Estimated models suggest that reducing poverty can be carried out by implementing policy in education health sectors.

There is a difference on education indicator that is included into the model in 2012. It implies that there may be a different education policy program needed to be applied in order to reduce the poverty. Moreover, spatial effect is statistically significant. It indicates that local specific policy would be needed given that there are differences on the infrastructure of health and education between regions. The magnitude of estimates show that more effort should be put in regencies, compared to cities, in order to reduce poverty in the regions.

REFERENCES

[1] Asselin, L., M. 2009. Analysis of Multidimensional Poverty: Theory and Case Studies. Ottawa: Springer.

[2] Gujarati, Damodar N. 2003. Basic Econometrics, Fourth Edition. New York: McGraw-Hill.

[3] Teguh and Nurkholis 2013: Finding Out of The Determinants of Poverty Dynamics in Indonesia: Evidence From Panel Data, *Bulletin of Indonesian Economic Studies*.

Safe and Affordable Drinking Water for All: A Development of a SDGs Proxy Indicator from MDGs Indicators

D. D. Prastyo
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
dedy-dp@statistika.its.ac.id

D. Cahyono
BPS Papua Barat Province
Makokwari, Indonesia
dedicah@bps.go.id

N. Susyanto
Universitas Gadjah Mada
Yogyakarta, Indonesia
nanang_susyanto@ugm.ac.id

I. S. Ahmad
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
safawi@statistika.its.ac.id

M. Rheza
Pulse Lab Jakarta
Jakarta, Indonesia
muhammad.rheza@un.or.id

ABSTRACT

The SDGs as the successor of MDGs program designs a universal holistic framework toward sustainable development. There are many new indicators in SDGs that are not supported by the available data as measured in MDGs. Some of these new indicators are very expensive to be collected. Therefore, the proxy model for this SDGs indicator with predictors from MDGs indicator is required. The indicator as a focus in this research is safely managed drinking water services. Due to the very small sample size of the available data, this work employed bootstrap on M-estimate regression. As the result, this paper gives the prediction of SDGs indicator, the percentage of safely managed drinking water services, for each district and city in Yogyakarta province.

KEYWORDS

SDGs indicator, safe water, MDGs indicator, clean water, small sample, bootstrap, M-estimate regression

1 INTRODUCTION

United Nations (UN) and its agencies have led and funded the international development since the late 1940s. Up to 1990s, the development approach was initiated by its specialized agencies at various World Summits and Conference to address three issues related to economic, social, and environmental. The Millennium Development Goals (MDGs) integrate the development agenda of United Nations Development Program (UNDP), United Nations Environment Program (UNEP), World Health Organization (WHO), United Nations Children's Fund (UNICEF), United Nations Educational, Scientific and Cultural Organization (UNESCO), and other development agencies [1].

In 2015, the successor of MDGs program so-called Sustainable Development Goals (SDGs) was adopted by the world leaders. The SDGs design a universal holistic framework to help set the world on a path towards sustainable development on economic, social inclusion, environmental sustainability, and good governance. Compared to MDGs with 8 Goals, the SDGs set more comprehensive goals (17 Goals with 230 indicators, even more by including country specific indicators, for monitoring 169 targets) to address the symptoms of poverty and the issues of peace, stability, human rights, and good governance.

The data related to MDGs indicator in Indonesia are regularly collected by government agencies. But, the data for many new indicators are not available yet. Some of them are very expensive to be collected. This issue becomes very important because without data the goals achievement is not measurable even the progress during the period is out of control. If the relationship between new indicators and existing indicators is known, for example via a model as a proxy, then the unavailable data related to certain indicators can be estimated from the available data.

This paper describes the development of a model as a proxy for SDGs indicators with observable MDGs indicators as determinant. The work focuses on one indicator related to water where the SDGs target is to provide safe and affordable drinking water for all. This issue becomes essential because data from Indonesian Demographic and Health Survey (IDHS) 2012 showed that an Infant Mortality Rate (IMR) of approximately 43 deaths per 1000 births in Indonesia, where 40 percent of infant death are caused by diarrhea and pneumonia [2] and in general about 88 percent of cases diarrheal disease can be linked to poor WASH (water, sanitation, and hygiene) service provision [3]. Due to the very small sample size available which is employed to build the model, the bootstrap regression provides alternative solution to develop the proxy model.

2 DATA AND METHOD

Goal 6 of SDGs aims to ensure the availability and sound management of water and sustainable sanitation for all by 2030. More specifically, the indicator 6.1.1 measures the proportion of population using safely managed drinking water services. In September 2015, due to the lack of data about the level of safe water in Indonesia, Badan Pusat Statistik (BPS) or Statistics Indonesia collaborated with the Ministry of Health, Badan Perencanaan Pembangunan Nasional (Bappenas) or National Development Planning Agency, and UNICEF implemented water quality survey (or Survei Kualitas Air (SKA) in

Indonesian) to obtain a detailed overview of water quality, sanitation, and hygiene at household level. This survey was expected to give baseline estimate to both the development targets of Indonesia as well as the SDGs. This experience is used to advocate to stakeholders to replicate the SKA in the future all over Indonesia.

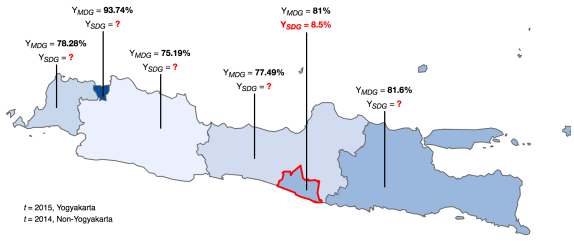


Figure 1. **Map of provinces in Java Island with its percentage of improved drinking water source (Y_{MDG}) and safe drinkable water (Y_{SDG}).**

In the SKA, the sample size was 940 households, taken from all the districts and cities in Yogyakarta. Data and sample collection was conducted by the BPS Yogyakarta team while the water analysis was conducted by the Ministry of Health's Centre for Environmental Health Engineering and Control of Diseases (BBTKLPP) Yogyakarta. The water quality parameters tested were the existence of E.Coli for microbiological indication of faecal contamination as well as nitrate and chloride to detect anthropogenic impact.

2.1 Proxy at Province Level

The survey, as per the definition of the National MDGs indicator 7.8, showed that 81.0% of households in Yogyakarta have access to improved drinking water source (Y_{MDG}), where 89% of which are contaminated with E.Coli despite the high level of access to an improved water source. This is indicating that improved drinking water sources are not always free of faecal contamination. In contrast to E.Coli contamination, the percentage of samples from the SKA 2015 with elevated level of nitrate and chloride was found to be much lower. Only 6.3% of household drinking water samples contain more than 50 mg/L of nitrate (i.e. exceeding the Ministry of Health's national water quality standards). No Chloride result exceeded the national water quality standards of 250 mg/L for chloride, as set by the Ministry of Health. The proportion of households with access to safely managed drinking water and sanitation facilities, as per the definition of the SDGs indicator 6.1.1, were estimated respectively to be 8.5% (Y_{SDG}). This shows a substantial difference from the (National) MDG estimates.

Fig. 1 displays the proportion of safely managed drinking water service (Y_{SDG}) compared to proportion of improved water drinking source (Y_{MDG}) in Yogyakarta. The other provinces in Java Island only have data about MDG indicator. If one is able to develop the proxy model for SDG indicator (Y_{SDG}) with predictors from MDG indicator (Y_{MDG}) for Yogyakarta, then the SDG indicator for safely managed drinking water source in other provinces in Java island can be estimated from that proxy model with certain adjustment. We propose equation (1) as a proxy model at province level as follows:

$$Y_{SDG} = Y_{MDG} \times StC\ ratio \times Adjustment\ Scale, \qquad (1)$$

where $StC\ ratio$ is Safe-to-Clean ratio calculated from SKA survey in Yogyakarta, i.e. 8.5 / 81.0 = 10.49% and $Adjustment\ Scale$ is used to customize the diversity of characteristics of other provinces compared to Yogyakarta using equation (2) as follows:

$$\frac{Y_{MDG}\ (\text{other province})}{Y_{MDG}\ (\text{Yogyakarta})} = f(X_1, X_2, \dots, X_p) + \varepsilon, \qquad (2)$$

with X_j, where $j = 1,2, \dots, p$, is a ratio of proportion of water source type consumed by population in corresponding province compared to Yogyakarta's, and ε is error. By assuming the linear dependence between response and predictors, one can employ linear regression as the form of $f(X_1, X_2, \dots, X_p)$. This formulation seems rationale, but it will end up with validation problem where the Y_{SDG} at province level only available for Yogyakarta. It means that there is only single datum used for validation which can be misleading.

2.2 Proxy at District Level

The development of proxy model for Y_{SDG} at province level has a problem with validation scheme. Therefore, we need an alternative that the Y_{SDG} observations used for validation are available more than single datum. This is possible for district level in Yogyakarta where there are five districts with Y_{SDG} data are available as shown by Fig. 2.

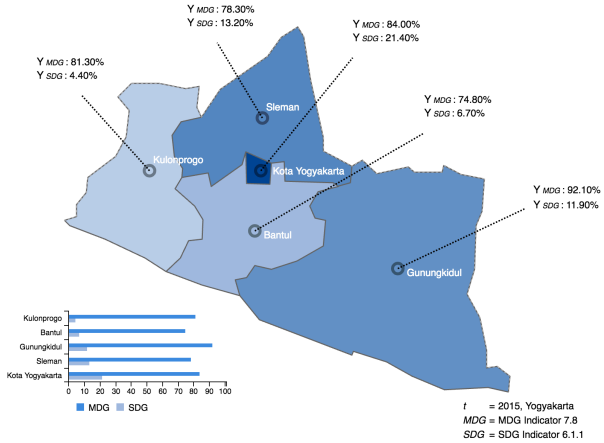


Figure 2. Map of districts in Yogyakarta province with its percentage of clean drinkable water (Y_{MDG}) and safe drinkable water (Y_{SDG}).

It is necessary to note that the Y_{SDG} values displayed in Fig. 2 are "proxy of safe water" (percentage of improved drinking water source not contaminated by E.Coli) rather than based on definition of SDG indicator 6.1.1. Therefore the ratio "proxy of safe water" at district level to safely managed drinking water at province level named as *AdjScale* is formulated in (3) as follows:

$$AdjScale = \frac{(1 - 0.89)}{0.085} = 1.29 \qquad (3)$$

At district level, the (1) is modified becomes (4) and (5) as follows:

$$\text{Estimate of } Y_{SDG} = \text{Proxy of } Y_{SDG} \times \frac{1}{AdjScale}, \quad (4)$$

$$\text{Proxy of } Y_{SDG} = f(Y_{MDG}) + \varepsilon, \quad (5)$$

where the Y_{MDG} chosen as determinant are MDG indicator 7.8 (percentage of population consuming clean water), MDG indicator 7.9 (percentage of population have access to improved sanitation), and MDG indicator 1.1 (percentage of population below poverty line). There are only five data available as Proxy of Y_{SDG} corresponding to five districts and city as written in Table 1.

2.3 Bootstrap Regression

The available data to build model is quite small, i.e. only five observations as written in Table 1. The question about how good the model obtained might arises. In order to deal with such issue, this work employs bootstrap regression [4] to resampling the observations.

Table 1. Data for Proxy Model Building (in percentage)

District	Proxy of Y_{SDG} (No-E.Coli)	Predictor (Y_{MDG})		
		Ind 7.8	Ind 7.9	Ind 1.1
Kulon Progo	4.4	81.3	76.5	21.40
Bantul	6.7	74.8	93.7	16.33
Gunung Kidul	11.9	92.1	67.4	21.73
Sleman	13.2	78.3	91.9	9.46
Yogyakarta	21.4	84.0	92.8	8.75

The Bootstrap introduced by [5] is a general approach to statistical inference based on building a sampling distribution for a statistic by resampling from the data at hand. Applying bootstrap in regression, there are two general ways: one can treat the predictors as random, i.e. potentially changing from sample to sample, or as fixed. The random predictors resampling is also called case resampling, and fixed predictors resampling is also called residual resampling. This work employs the residual resampling with a slight modification by resampling scaled and centered residual as suggested by [6], see Algorithm 6.3. The source code for entire analysis in this work was written in **R** open source statistical software. We refit the bootstrap regression using an *M*-estimator with the Huber weight function that can be invoked by the **rlm** function in the **MASS** package, which is available when **car** package, as the support for [4] released after 2012, is loaded. The **Boot** with a capital "B" is a function in the **car** package provides a simplified front-end to the **boot** package [6] (also has function **boot**) that requires some programming.

```
> summary(NonEcoli.boot.res)
                   R   original    bootBias   bootSE    boo
(Intercept)     1000 -120.92183 -0.19307761 26.53576 -121.5
CleanWater      1000    1.19633  0.00086219  0.15352    1.1
ImprovedSanitation 1000    0.49110  0.00069129  0.14134    0.4
ProcentPoverty  1000   -0.46703  0.00407454  0.17806   -0.4
```

Figure 3. Summary output of Bootstrap regression

3 RESULTS AND DISCUSSION

The Fig. 3 shows the summary of Bootstrap regression. The column <original> gives the M-estimates for each coefficient of regression obtained from original observation. The column <bootBias> displays the bootstrap estimates of bias, i.e. the difference $\mathbf{Bias} = (\bar{\boldsymbol{\beta}}^* - \boldsymbol{\beta})$ between the average bootstrapped value of the statistic $(\bar{\boldsymbol{\beta}}^*)$ and its original-sample value $(\boldsymbol{\beta})$, with $\boldsymbol{\beta}$ is vector of coefficient corresponding to predictor in Table 1.

Therefore, the estimates of Bootstrap on M-estimate regression are calculated by $\bar{\boldsymbol{\beta}}^* = \boldsymbol{\beta} + \mathbf{Bias}$, i.e. the $\bar{\beta}_0^* = -120.926$ (intercept), $\bar{\beta}_1^* = 1.196$ (Clean Water), $\bar{\beta}_2^* = 0.491$ (Improved Sanitation), and $\bar{\beta}_3^* = -0.466$ (Poverty). In addition, the bootstrap estimates of the standard error $\left[\widehat{SE}^*(\boldsymbol{\beta}^*)\right]$ are computed as the standard deviation of the bootstrap replicates.

The confidence interval for the each coefficient estimator is calculated by means of two approaches. First, it uses normal theory with the bootstrap standard errors. Second approach uses the percentile method that gives quantiles for a number of intervals simultaneously. The confidence interval computation is summarized in Fig. 4. For two side hypothesis, with Type-I error five and ten percent, the two approaches result in the same sign (positive or negative) for each bootstrap estimates.

```
> confint(NonEcoli.boot.res, parm=(2:4), level=c(0.90, 0.95), type="norm")
Bootstrap quantiles, type =  normal

                        2.5 %        5 %       95 %     97.5 %
CleanWater          0.8945603  0.9429376  1.4479894  1.4963667
ImprovedSanitation  0.2133969  0.2579332  0.7228854  0.7674216
ProcentPoverty     -0.8200993 -0.7639904 -0.1782209 -0.1221119
> confint(NonEcoli.boot.res, parm=(2:4), level=c(0.90, 0.95), type="perc")
Bootstrap quantiles, type =  percent

                        2.5 %        5 %       95 %     97.5 %
CleanWater          0.8712603  0.9450355  1.4476159  1.4476159
ImprovedSanitation  0.2111462  0.2233554  0.7588458  0.7588458
ProcentPoverty     -0.7828240 -0.7412499 -0.1954660 -0.1928122
```

Figure 4. Confidence Interval

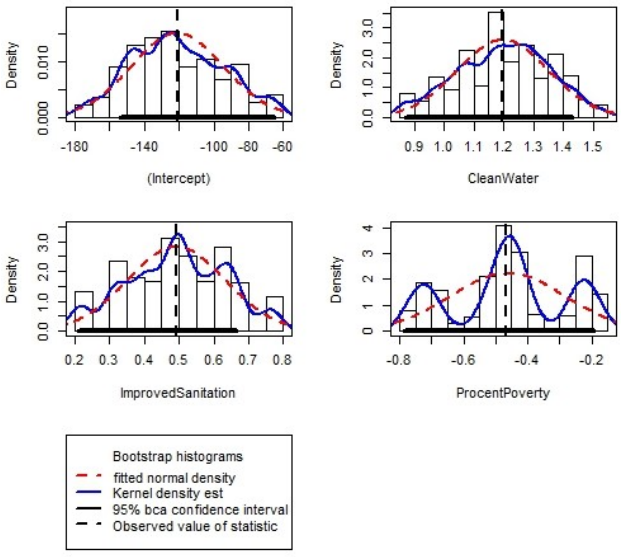


Figure 5. Empirical Density for Each Coefficient's Estimator

Each coefficient estimated from bootstrap replication is displayed as histogram in Fig. 5 with addition of the kernel density estimates and the normal density based on the bootstrap mean and standard deviation. The vertical dashed line represents the original point-estimate whereas the thick horizontal line gives a confidence interval calculated from bootstrap. The normal approximation is poor for coefficient that corresponding poverty indicator that seems has mixture density.

For other coefficient, the confidence intervals are not close to symmetric about the original values. This informs that inference from the bootstrap is different from the asymptotic theory. The bootstrap is likely to be more accurate in this small sample.

Fig. 6 examines the joint distribution of the bootstrapped pair of coefficient. The scatterplot of bootstrap replication for each pair with concentration ellipse was superimposed at level 50, 90, and 99% using a robust estimate of the covariance matrix of the coefficients.

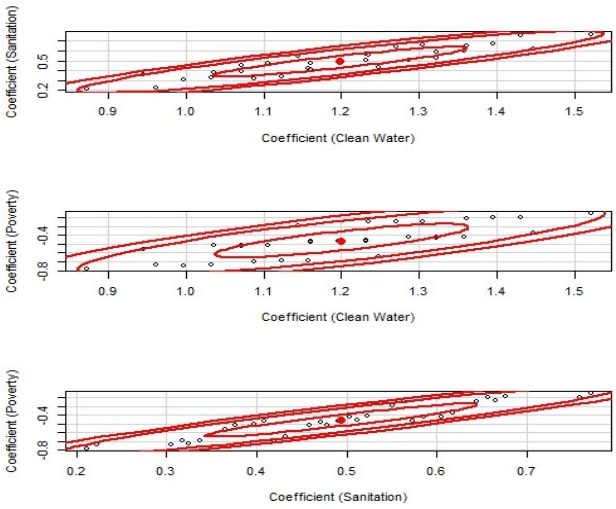


Figure 6. **Scatterplot of bootstrap replication for Each Pair of Coefficient's Estimator.**

Table 2 gives the information about the estimation of the percentage of improved drinking water source that are not contaminated by E.Coli as a proxy of safely managed drinking water service. The predicted values obtained from bootstrap regression produces mean squared error 13.60%. This means that the proposed model has prediction error about that value which is quite reasonable for such very small sample size.

Table 2. Estimates of the Proxy of Y_{SDG} (Percentage)

District	No-E.Coli (observation)	No-E.Coli (estimation)	Squared of residual
Kulon Progo	4.4	3.9	0.25
Bantul	6.7	7.0	0.09
Gunung Kidul	11.9	12.2	0.09
Sleman	13.2	13.5	0.09
Yogyakarta	21.4	21.0	0.16
MSE			**0.1360**

Table 3. Estimates of the Y_{SDG} (Indicator 6.1.1)

District	Proxy of Y_{SDG} (observation)	Proxy of Y_{SDG} (estimation)	Estimates of Y_{SDG}
Kulon Progo	0.044	0.039	**0.0303**
Bantul	0.067	0.070	**0.0537**
Gunung Kidul	0.119	0.122	**0.0944**
Sleman	0.132	0.135	**0.1040**
Yogyakarta	0.214	0.210	**0.1626**

The results showed in Table 2 are used to calculate the SDGs indicator as written in Table 3. The estimate of percentage of the safe drinking water source in Kulon Progo district is only 3.03%, the lowest among other districts and city. Bantul, Gunung Kidul, and Sleman districts have percentage respectively 5.37%, 9.44%, and 10.40%. The Yogyakarta city has the highest percentage among others with 16.26%. This empirical results show that the proposed methods gives reasonable output although the available data is very small.

4 CONCLUSIONS

The SDGs designs a universal holistic framework toward sustainable development as the successor of MDGs program. There are many new indicators in SDGs that are not supported by the available data as measured in MDGs. Some of these new indicators are very expensive to be collected, for example the safely managed drinking water source as the focus of this research. Therefore, the proxy model for this SDGs indicator with predictors from MDGs indicator is required. The pilot project about the safe drinking water was done in Yogyakarta in 2015. There are 940 household surveyed as smallest unit sampling. But, there are only five data corresponding to each district and city that comply with the indicator requested. Such small sample size becomes a problem to obtain valid estimation of the proxy model. This work employed bootstrap on M-estimate regression in order to obtain the more accurate estimate based on replication of sample at hand. As the result, this paper gives the prediction of SDGs indicator, i.e. percentage of safely managed drinking water services, for each district and city in Yogyakarta province. The proposed method resulted in about 13.60% of error in prediction which quite reasonable compared with the small sample size. The method described in this manuscript can be improved further in two ways, by improving the statistical methods used or by doing more intensive research on predictor selection as inputs in proxy model.

ACKNOWLEDGMENTS

This work was fully supported by Pulse Lab Jakarta under the third Research Dive program with theme Statistics for Sustainable Development Goals held in Jakarta.

REFERENCES

[1] Robert Sanjiv Kumar, Neeta Kumar, and Saxena Vivekadhish. 2016. Millennium development goals (MDGS) to sustainable development goals (SDGS): Addressing unfinished agenda and strengthening sustainable development and partnership. *Indian J Community Med.* 41(1), 1–4. DOI: 10.4103/09700218.170955

[2] Badan Pusat Statistik. 2016. Mewujudkan Aksesibilitas Air Minum dan Sanitasi yang Aman dan Berkelanjutan Bagi Semua: Hasil Survei Kualitas Air di Daerah Istimewa Yogyakarta Tahun 2015. Badan Pusat Statistik, Jakarta.

[3] L. Fewtrell, A. Prüss-Üstün, R. Bos, F. Gore, J. Bartram. 2007. Water, Sanitation and Hygiene: Quantifying the Health Impact at National and Local Levels In Countries with Incomplete Water Supply and Sanitation Coverage. WHO: Geneva, Switzerland.

[4] J. Fox and S. Weisberg. (2011). *An R Companion to Applied Regression, 2nd Ed.* Sage, Thousand Oaks, CA.

[5] B. Efron. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1-26.

[6] A. C. Davison and D. V. Hinkley. 1997. *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Ensuring the Quality of Data: No Accessibility to Raw Data

S. A. Thamrin
Hasanuddin University
Makassar, Indonesia
tuti@unhas.ac.id

H. Yusnissa
BPS NTB Province
West Nusa Tenggara, Indonesia
hertinay@gmail.com

M. U. Fahik
Institute of Resource Governance and Social Change, Kupang
East Nusa Tengara, Indonesia
selusfahik@gmail.com

B. Warsito
Diponegoro University
Semarang, Indonesia
budiwrst2@gmail.com

M. Subair
Pulse Lab Jakarta
Jakarta, Indonesia
muhammad.subair@un.or.id

ABSTRACT

In this paper, we proposed the method to ensure the data quality when we are not accessibility to the raw data. This method includes checking the completeness the data and does the data validation. In validation data step, the more challenging way is not only to use the internal data but also use the external/ proxy data to ensure the data quality. The methodology was implemented use the MDGs dataset, Indonesia. The result indicated that there is a positive indication we can ensure the quality of the data although we do not have access to raw data. Because of the assumption and the heterogeneity of characteristics of the data, every approach only conducts for the specific indicators in specific provinces. In validation the data using external/proxy data by grouping the provinces in the same islands show that the results was not 'consistent'. There is the data in one provinces always become outliers in the islands.

KEYWORDS

Correlation, external data, missing data, outlier, proxy, validation

1 INTRODUCTION

The main purpose of ensuring data quality is to present reliable information. It means that data including survey process such as data collection and statistical accuracy meet the need of user and less of error (misinformation). Discovering whether data are of acceptable quality is very important. Data quality is composed of eight distinct aspects: relevance, objectivity, validity, reliability, integrity, completeness, generalizability, and utility.

Sometimes we find that specific facts varying from source to source. We need to research which data/source are most accurate. Currently, the Statistics Centre Bureau of Indonesia, called BPS, have done the several steps to ensure the data quality if the raw data is accessible. These steps are ensure the completeness, do the cross tabulation within the internal dataset, test the validation and reliability of the data and calculate Relative Standard Error (RSE). However, sometimes we do not have access to raw data and we need to make sure the quality of data. In this situation, we need to use the other approach to improve the quality of 'collected' data. For instance to statistically identify inconsistent data points using other data. Utilizing external data is only going to benefit if it is correct.

Evaluating the quality of data has been studied widely by many researchers. Strong et al. [6] discussed the data quality in context of organization and their research can adopts a data-consumer perspective. Pipino et al. [5] described principles that can help organizations to develop usable data quality metrics. Batini et al. [1] proposed comprehensive methodologies for data quality assessment and improvement.

The literature on incorporating the internal data and the external/ proxy data into the methodology in ensuring the data quality is sparse and limited to small-scale studies [3-4]. Therefore, in this paper, we consider to recommend a 'framework' to ensure the quality of data when we do not have access to raw data.

The paper is organized as follows. In Section 2, we describe the proposed method if no accessibility to raw data. In Section 3, we apply the proposed method using a Millennium Development Goals (MDGs) dataset. The results are discussed further in Section 4. The conclusion and the future work are presented further in the last section.

2 DATA QUALITY FRAMEWORK

In this section, we explain the proposed method if no accessibility to raw data. The proposed method of ensuring the data quality is shown in Fig. 1. To ensure the data quality, there are several steps that we propose to do as the following.

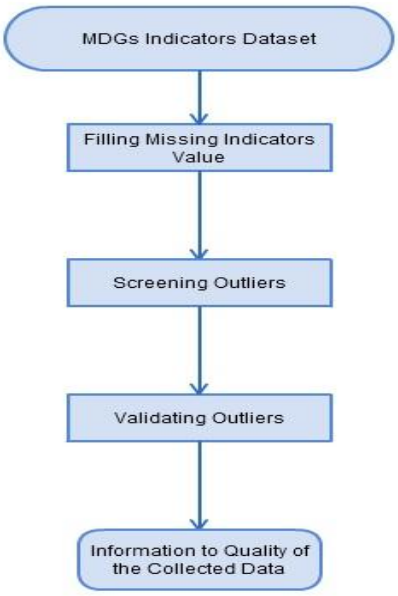


Figure 1. Data quality framework

2.1 Ensure the Completeness

The completeness of the data needs to be checked first to see whether there are the missing data or not. The reason of the missing data needs to be considered carefully, whether the data is missing at random or not.

Before understanding the reasons why data are missing, it is important to correctly handle the remaining data. If there is no real pattern for missing values, the missing values are mostly random. Meanwhile, if the values are missing completely at random, the data sample is likely still representative of the population. However, if the values are missing systematically, the analysis will be biased.

In statistics, missing data, or missing values, occur when the data value is not available for the indicator. Missing data is a common occurrence that can occur either because of nonresponse or the lack of the data were not publish. As the impact, missing data can have a significant effect on the conclusions that be drawn from the data.

There are many approaches to deal with the missing data. In this paper, before filling the missing data, we plot the data to see the pattern of data; linear trend, non-linear trend or locally linear trend.

The data are missing because we were not able to find full data in the dataset. There is no real pattern for missing values, apart from some periods are missing, the missing values are mostly random and linear. So we can use linear trend to fill the data. Otherwise if we find the data do not have the linear trend or locally linear trend, we can use the nearest neighbourhood technique to fill the data. The diagram for filling the missing data can be seen in Fig. 2.

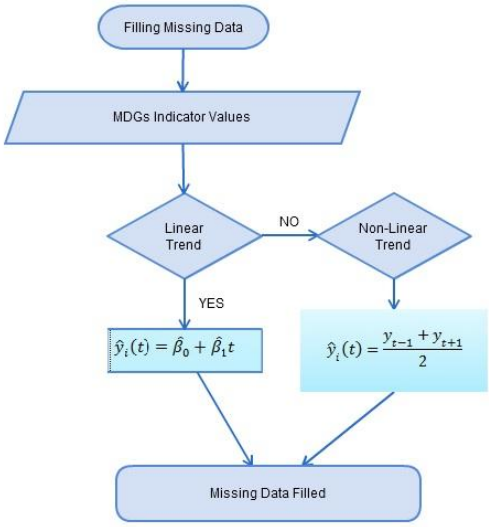


Figure 2. The flow of filling the missing data

Suppose we have the linear model and the estimated linear model, respectively as following:

$$y_i(t) = \hat{\beta}_0 + \hat{\beta}_1 t \,, \qquad (1)$$

where *i*= 1, 2, …n, y is the dependent variable, t is the independent variable, β_0 is the coefficient intercept and β_1 is the coefficient parameter. By using the linear trend test, we can see if we have $\hat{\beta}_1 \neq 0$, then it means that our model have the linear trend. As the consequence, the missing data is then estimated using the equation (2). Meanwhile, if $\hat{\beta}_1 = 0$, then our model have the non-linear trend or locally linear trend. The missing data is then estimated using:

$$\hat{y}_i(t) = \frac{y_i(t-1)+y_i(t+1)}{2} \qquad (2)$$

2.2 Data Validation

We can validate the data by finding the outliers data either with the internal data or the external/ proxy data. In this paper, we use the absolute standardized method to determine the outliers for the internal data. If the number of absolute standardized of the data is more than 2 then we categorized the data as an outlier.

Furthermore, to validate the outliers using the external data or proxy, we use the following steps:

a) From the output list of outliers from the internal data, we prepare all proxy indicators based on the same area, i.e. province.
b) For all proxy indicators, we calculate the correlation either with complete data including outliers or excluding outliers.
c) We compare the result of the both correlations from part (b) using the Pearson product moment formula (*r*) as follow:

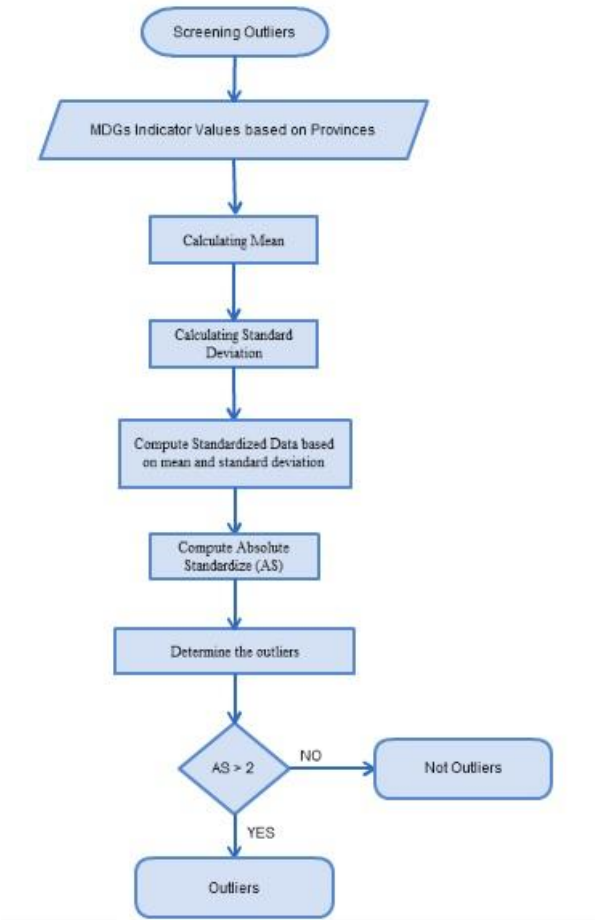


Figure 3. The diagram for screening the outliers' data.

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

The flow diagram for screening the outliers and validating data are presented in Fig. 3 and 4.

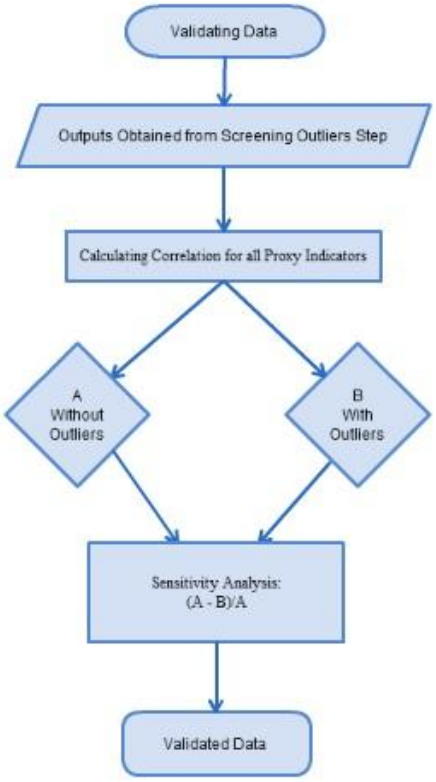


Figure 4. The diagram for validating data.

3 APPLICATION TO MDGS DATASET

We applied the proposed method of ensuring the data quality by using the MDGs dataset from 2001-2014. This dataset contains 34 provinces in Indonesia and there are 82 indicators for each province. To describe first the completeness of the data, in this paper, we used the MDGs indicator 1.1a (poverty) for MDGs data of Daerah Istimewa (DI) Yogyakarta, DKI Jakarta and West Nusa Tenggara provinces, respectively.

From the Fig. 5, we can see the MDGs data of DI Yogyakarta for all of the indicators. Apart from some periods as the one illustrated in the Figure 5, it is not the real pattern for missing values. In 2004, there is the missing data (the blank dot). These missing values are mostly random.

To find the data are linear on not, we use a scatter plot to show the relationship between independent variable and time sequence of the observation. With regression analysis, we use a scatter plot to visually whether Y and t are linearly related. Figure 6 present the scatter plot between year of the data and the MDGs indicator 1.1a (poverty) for West Nusa Tenggara Province. Each point on the graph represents a single (t, Y) pair. In the Figure 6, the graph is a straight line; the relationship between year and the MDGs indicator 1.1a for West Nusa Tenggara Province is linear. The linear model for Fig. 6 can be written as:

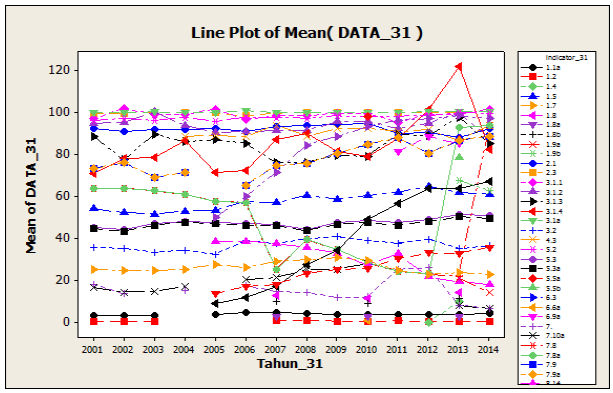


Figure 5. Plot of the MDGs data for all of indicators in Daerah Istimewa (DI) Yogyakarta province.

$y(t) = 41.1879 + 0.8081\ t,$

with the probability of value (p-value) is 0.004373. This p-value is less than 0.05. It means that the coefficient of trend is significant. In this case, t indicates the time sequence of observations. Therefore the trend model can be used to estimate the missing value. The MDGs Indicator 1.1a of West Nusa Tenggara Province is following a linear trend, so that the missing value in 2004 can be estimated by using the linear trend.

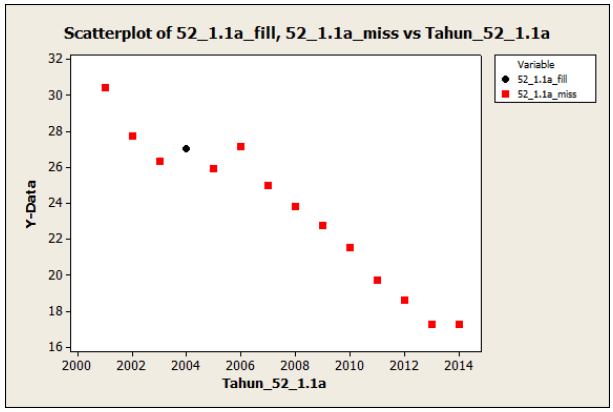


Figure 6. The Scatter Plot of MDGs Indicator 1.1a of West Nusa Tenggara Province.

To show the non-linear trend, we make a scatter plot between year of the data and the MDGs indicator 1.1a (poverty) for DKI Jakarta province. From the Fig. 7, we can see that the graph is not a straight line, the relationship between year and the MDGs indicator 1.1a (poverty) for DKI Jakarta province is non-linear. The linear model for Figure 7 can be written as:

$y(t) = 3.5741 + 0.0275\ t$

with the probability of value (p-value) is 0.3931 that is bigger than 0.05. It means that the coefficient of trend is not significant. Therefore the trend model cannot be used to estimate the missing value. The MDGs Indicator 1.1a of DKI Jakarta Province is following the non-linear trend. The missing value in 2004 can be estimated by using the nearest neighbourhood technique.

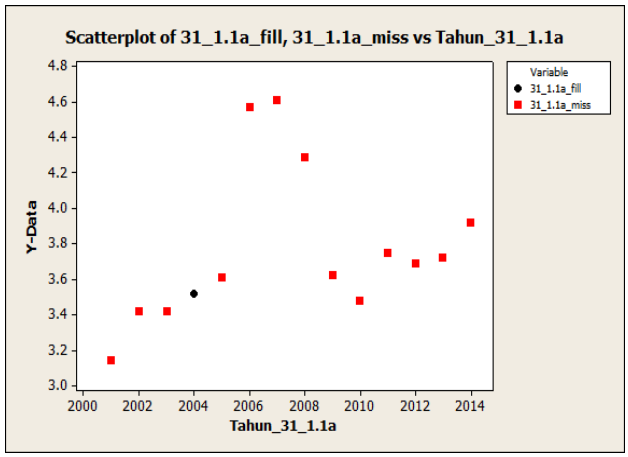


Figure 7. The Scatter Plot of MDGs Indicator 1.1a of DKI Jakarta Province.

The next step to ensure the data quality is validation data. We used the indicator 1.1a (poverty) and indicator 5.3 (Contraceptive prevalence rate) of all provinces from 2001-2014 as the internal data to find the outliers'.

By following the screening outliers flow, we find the outliers; extreme values for this indicator. These values are those that lie outside of the statistical model being used to describe the data as indicated in Fig. 3. The outliers for these indicators can be seen in Tables 1 and 2.

From Tab. 1, we can see there are five provinces in 2001, which have the indicator values, indicate as the outliers for indicator 1.1a; poverty.

Table 1. The outliers for indicator 1.1a for all provinces in Indonesia from 2001-2014.

Province ID	Province	Year	Value
13	West Sumatera	2001	15.2
34	DI Yogyakarta	2001	24.5
36	Banten	2001	17.2
61	West Kalimantan	2001	19.2
63	South Kalimantan	2001	11.9

Meanwhile, for indicator 5.3 (Contraceptive prevalence rate), almost all of provinces (70.5%) in 2014 have the outliers, except Bali province in 2001 (especially in the 4th comparison) show there is significant change in correlation if the data is removed from set.

Table 2. The outliers for indicator 5.3 for all provinces in Indonesia from 2001-2014.

Province ID	Province	Year	Value
11	Nangroe Aceh Darussalam	2014	53.59
12	North Sumatera	2014	49.9
13	West Sumatera	2014	51.07
14	Riau	2014	56.63
15	Jambi	2014	67.46
16	South Sumatera	2014	65.03
17	Bengkulu	2014	69.27
18	Lampung	2014	67.8
32	West Java	2014	66.29

33	Central Java	2014	63.17
34	DI Yokyakarta	2014	59.59
36	Banten	2014	62.23
51	Bali	2001	30.97
52	West Nusa Tenggara	2014	58.2
53	East Nusa Tenggara	2014	43.43
61	West Kalimantan	2014	69.66
62	Central Kalimantan	2014	71.86
63	South Kalimantan	2014	67.51
71	North Sulawesi	2014	67.8
73	South Sulawesi	2014	52.79
74	Southeast Sulawesi	2014	53.51
75	Gorontalo	2014	65.24

To describe the validation data with external/ proxy data, we used the indicator ratio of girls to boys in primary schools, ratio of girls to boys in Junior high school, ratio of girls to boys in senior high school and ratio of girls to boys in higher education [2]. These proxies have the correlated indicator with the proportion of population below national poverty line for MDGs indicator 1.1a.

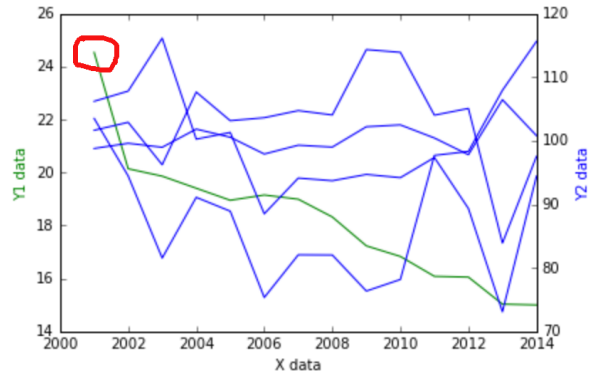


Figure 8. Bivariate correlation before removing the outliers for MDGs indicator 1.1a in DI Yogyakarta province from 2001-2014.

Fig. 8 shows the sample bivariate correlation before removing the outliers for MDGs indicator 1.1a in DI Yogyakarta province from 2001-2014. From Fig. 8, we can see that there is a significant change in correlation if the outlier data is removed from original MDGs indicator 1.1a in DI Yogyakarta.

Table 3. The correlation before and after removing outlier with proxy indicator for validating data.

Proxy Indicator	Correlation before	Correlation after
Ratio of girls to boys in primary school	-0.508	-0.545
Ratio of girls to boys in junior school	0.116	0.217
Ratio of girls to boys in senior high school	0.639	0.662
Ratio of girls to boys in higher education	-0.209	-0.654

Based on the Figure 4, we find the sensitivity of the data by counting the correlation between MDGs indicator 1.1a and external/ proxy data. We used the indicator ratio of girls to boys in primary schools, ratio of girls to boys in Junior high school, ratio of girls to boys in senior high school and ratio of girls to boys in higher education before, we called it as r_before. Next, we count the correlation between MDGs indicator 1.1a and external/ proxy data ratio of girls to boys in primary schools, ratio of girls to boys in Junior high school, ratio of girls to boys in senior high school and ratio of girls to boys in higher education after we removed the outlier, we called it as r_after. The detail can be seen in Table 3. Table 3 shows that there is a significant change in correlation if the outlier is removed from dataset (especially in the 4th comparison).

4 CONCLUSIONS

In this paper, we already presented the proposed 'framework' to ensure the quality of data when we do not have access to raw data. Overall, the results of this study indicated that there is a positive indication we can ensure the quality of the data although we do not have access to raw data. Because of the assumption and the heterogeneity of characteristics of the data, every approach only conducts for the specific indicators in specific provinces. Our example to validate the data using external/proxy data by grouping the provinces in the same islands show that the results was not 'consistent'. There is the data in one provinces always become outliers in the islands. We have not found the fix method for automatic decision making for the validation of the data using external/proxy data. We still need to find the method to model threshold/confident level, especially if the dataset has minimum cases/record. Furthermore, we also need to find the method for decision making for data validation. For instance, if the data size is significant, we could consider using t-test, interval comparison, etc. The other thing, we should need to do more testing to improve the framework and methodology with more samples and 'qualified' data.

REFERENCES

[1] C. Batini, C. Cappiello, C. Francalanci and A. Maurino. 2009. Methodologies for Data Quality Assessment and Improvement. *ACM Computing Survey 4, 3* (July. 2009), 1–52. DOI: http://dx.doi.org/10.1145/1541880.1541883.

[2] M. C. Lo Bue, and S. Klasen. 2013. Identifying Synergies and Complementarities Between MDGs: Results from Cluster Analysis. *Social Indicator Researches* 13, 2 (September 2013), 647–670. DOI: 10.1007/s11205-013-0294-y.

[3] Y. N. Marinez, C. H. McMahan, G. N. Barnwell and H. S. Wigodsky. 1984. Ensuring Data Quality in Medical Research Through an Integrated Data Management System. *Statistics in Medicine* 3: 101-111. DOI: 10.1002/sim.4780030204.

[4] D. E. Mulyana. 2010. Critical Topics in Ensuring Data Quality in Bio-Analytical LC-MS Method Development. *Bioanalysis*, 2, 3 (June 2010): 1051-1072. DOI: 10.4155/bio.10.60.

[5] L. L. Pipino., Y. W. Lee., and R. Y. Wang. 2002. Data Quality Assessment. *Communications of the ACM, 45, 4* (April 2002): 211-218.

[6] D. M. Strong., Y. W. Lee., and R. Y. Wang. 1997. Data Quality in Context. Commnunications of the ACM, 40, 5 (May 1997): 103-110: DOI: http://dx.doi.org/10.1145/253769.253804.

Spatial Disaggregation of MDGs Indicator with Numerical Method Approach

Agus Mohamad Soleh
Institut Pertanian Bogor
Bogor, West Java, Indonesia
agusms@apps.ipb.ac.id

Qurratul Aini
Indonesia Statistics Agency
West Nusa Tenggara, Indonesia
qurraatulaini@bps.go.id

Syarifah Diana Permai
Bina Nusantara University
Jakarta, Indonesia
syarifah.permail@binus.ac.id

Utriweni Mukhaiyar
Institut Teknologi Bandung
Bandung, West Java, Indonesia
utriweni@math.itb.ac.id

Ni Luh Putu Satyaning P. P
Pulse Lab Jakarta
Jakarta, Indonesia
ni.paramita@un.or.id

ABSTRACT

Detailed disaggregation for development indicators is important to ensure that everyone benefits from development and support better development-related policy making. This paper aims to explore different methods to disaggregate national employment-to-population ratio indicator to province- and city-level. Numerical approach is applied to overcome the problem of disaggregation unavailability by constructing several spatial weight matrices based on the neighborhood, 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 (United Nations, 2015). 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 that 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.

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 (Pratesi, et.al, 2015). 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 METHODOLOGY

2.1 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 (Yi) is uniformly distributed in each area. The target variable can be estimated as (Flowerdew and Green, 1994).

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

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

2.2 Numerical Method Approach

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 (Kubicek, Janovska and Dubcova, 2005).

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. The sequence is said to converge to the limit if $\left|Y - Y^{(m)}\right| < \varepsilon$ (McDonough, 2007). Iterative methods to find a sequence of approximation solutions following

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

where

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

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

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

$\boldsymbol{W}$ is a spatial weight matrix $\{w_{ij}\}$

$\vec{Y}^{(k)}$ is *k*-th iteration value of indicator in *i*-th area

Stopping rule is defined as if $\left|\vec{Y}^{(k)} - Y^{(actual)}\right| < \varepsilon$.

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 (Getis and Aldstadt, 2003). 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 (Drukker, *et.al*, 2013). Therefore, in this paper several methods of constructing spatial weight matrix using geographical proximity between areas are experimented.

2.2.1 Neighborhood-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 (Tang and He, 2015).

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

where

$$m = \begin{cases} 1 & if\ two\ locations\ share\ the\ same\ borderline \\ 0 & otherwise \end{cases}$$

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

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

where

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

2.2.2 Euclidean Distance-based

Geographical proximity can be measure 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 (Krislock and Wolkowicz, 2011).

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

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.

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

2.2.3 Correlation-based

Methods based on correlation are desirable if the relationships between the 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 (Amerise and Tarsitano, 2012). Define correlation matrix of intended units based on data history and construct a distance matrix D as follow.

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

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.

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

where

$$m_{ij} = \begin{cases} 1 & if\ two\ locations\ share\ the\ same\ borderline \\ 0 & otherwise \end{cases}$$

3 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.

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 and 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.

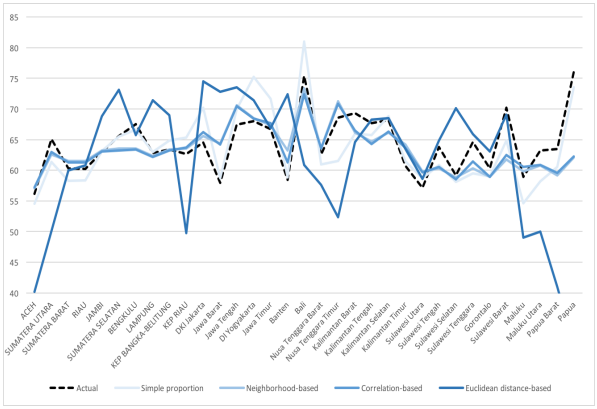


Figure 1. National-to-province disaggregation results and the actual data

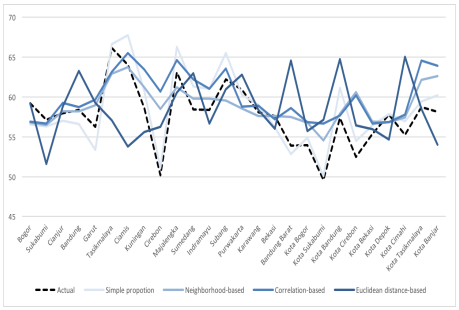
In order to evaluate the models, several statistics to measure the goodness of models have been calculated as shown in Table 1. The lower the value of mean average error (MAE), mean average percentage error (MAPE), and mean squared error (MSE), the better the model. From these three criterias, correlation-based 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.

Table 1. National-to-province model evaluation

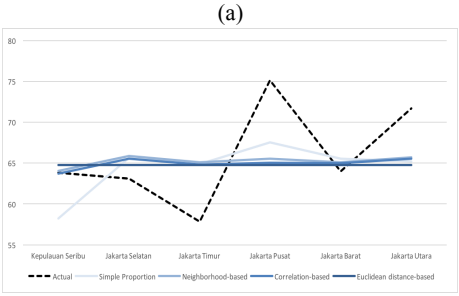
	MAE	MAPE	MSE
Simple proportion	2.689	4.1	11.323
Neighborhood-based	2.726	4.2	14.710
Euclidean distance-based	8.347	12.9	147.018
Correlation-based	2.615	4.0	13.319

Province-to-city disaggregation are also done for DKI Jakarta and Jawa Barat using the same methods. Although there are some underestimations (e.g. Ciamis, Tasikmalaya) and a lot of overestimations, the models for Jawa Barat disaggregation can closely estimate several cities, for instance Majalengka, Sumedang, Indramayu, Subang, Purwakarta and Karawang. Based on the evaluation criterion, 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.

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 also 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.



(a)



(b)

Figure 2. Province-to-city disaggregation results and the actual data for Jawa Barat (a) and DKI Jakarta (b)

Table 2. Province-to-city model evaluation

		MAE	MAPE	MSE
Jawa Barat	**Simple proportion**	1.739	3	4.174
	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
	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

4 CONCLUSIONS

Numerical method 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 thing that is also important 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

Amerise, I.L. and Tarsitano, A. 2012. Weighting Distance Matrices Using Rank Correlations. Working Paper n. 09 – 2012. Italy : Dipartimento di Economia e Statistica, Universita Della Calabria.

Drukker, D.M., Peng, H., Prucha, I.R. and Raciborski, R. 2013. Creating and Managing Spatial Weighting Matrices With The Spmat Command. The Stata Journal, 13, Number 2, pp. 242 – 286.

Flowerdew, R. and Green, M. 1994. Areal Interpolation and Types of Data, in S. Fotheringham and P. Rogerson (Eds), Spatial Analysis and GIS, 121-45, London: Taylor and Francis.

Getis, A. and Aldstadt, J. 2003. Constructing the Spatial Weights Matrix Using a Local Statistics. Geographical Analysis, Vol. 36, No. 2, pp. 90 – 104.

Krislock, N. and Wolkowicz, H. Euclidean distance matrices and applications, in: Miguel Anjos, Jean Lasserre (Eds.), Handbook of Semidefinite, Cone and Polynomial Optimization: Theory, Algorithms, Software and Applications.

Kubicek, M., Janovska, D. and Dubcova, M. 2005. Numerical Methods and Algorithm. Praha: Vysoká škola chemicko-technologická v Praze. ISBN 80-7080-558-7.

McDonough, J.M. 2007. Lectures in Basic Computational Numerical Analysis. USA: University of Kentucky.

Pratesi, M., Petrucci, A., Salvati, N., 2015. Spatial Disaggregation and Small-Area Estimation Methods for Agricultural Surveys: Solutions and Perspectives. Technical Report Series GO-07-2015, Global Strategy.

Tang, B. and He, H. 2015. Extended Nearest Neighbor Method for Pattern Recognition. IEEE Computational Intelligence Magazine. 1556-603x/15©2015IEEE. August 2015, pp. 52 – 60.

United Nations. 2015. The Millennium Development Goals Report.

Zhao, X. 2014. Nonlinear Programming. Encyclopedia of Business Analytics and Optimization. Category: Algorithms and Programming. IGI Global, pp. 1637 – 1646. DOI: 10.4018/987-1-4666-5202-6.ch147

http://rd.pulselabjakarta.id/

Pulse Lab Jakarta is grateful for the generous support from
the Department of Foreign Affairs and Trade of the Government of Australia.