© 2018 IJSRSET | Volume 4 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Assessment Method for Weighting and Aggregation in Constructing Composite Indicators of Mixed Data

Arni Nurwida*, Aji Hamim Wigena, Budi Susetyo
Department of Statistics, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, West Java, Indonesia

ABSTRACT

Composite indicators are often encountered in various studies, especially in the social sector. Composite indicators are constructed from several steps such as weighting and aggregation. The classical weighting method such as weighting based on factor analysis and regression analysis cannot handle a mixture of numeric and categoric variables. Furthermore, using a dependent variable as the estimator in weighting based on regression analysis is sometimes manipulated by respondents. An approach to address this problem uses the weighting method based on factor analysis of mixed data. The classical aggregation method such as linear additive aggregation cannot handle a mixture of compensatory and non-compensatory numeric variables. Therefore, to address this problem, a geometric aggregation was used. The case study constructed the five models of household welfare status of Dramaga village, Bogor regency that used the combination of weighting method based on multiple correspondence analysis and factor analysis of mixed data and linear and geometric aggregation. The five models are compared. The best model was model using the weighting method based on factor analysis of mixed data and the geometric aggregation for the numeric variables and the linear aggregation for the categoric variables.

Keywords: Composite Indicators, Factor Analysis of Mixed Data, Geometric Aggregation, Household Welfare Status

I. INTRODUCTION

The Indonesian household welfare status was constructed by Statistics Indonesia (BPS) in collaboration with the National Team for the Acceleration of Poverty Reduction (TNP2K) in order to classify the Indonesian households based on their welfare levels. It was formed into 10 status levels, from the lowest to the largest level (TNP2K, 2013).

The construction of the welfare status used a number of individual composite indicators which consist of numeric and categoric variables. The composite indicators are constructed using two steps, i.e. weighting and aggregation variables. The weighting calculates the weight of indicators and the aggregation constructs the composite index by combining the individual indicators and its weights (OECD, 2008).

BPS constructed the welfare status using the weighting and the aggregation method based on regression analysis with the household expenditure per capita as the dependent variable and the mixed data as the independent variables. The weights were calculated from the parameters of the independent variables and the aggregation used the linear regression model (TNP2K, 2013).

The main problems in weighting process are (1) the individual indicators constructed from numeric and categoric variables and (2) the dependent variable is

usually manipulated (Sari, 2011). The problem of aggregation process is that the numeric variables are correlated.

There are several alternatives of weighting and aggregation methods, such as the weighting method based on multiple correspondence analysis and factor analysis of mixed data and the linear and geometric aggregation method (OECD, 2008). The weighting method based on multiple correspondence analysis was designed for the categoric variables, while the weighting method based on factor analysis of mixed data for the mixed variables (Pages, 2004).

This weighting method describes that in a set of variables there is a latent structure that explains a certain value and the weights are calculated from the loading factors and the eigenvalues (OECD, 2008 and Asselin, 2009). Based on the research of Castano (2002), the weighting method based on factor analysis of categoric data provided more accurate results than the weighting method based on regression analysis.

The linear aggregation method is a compensatory aggregation that is designed for the low correlated variables, while the geometric aggregation method for the low and high correlated variables (Mazziota & Pareto, 2013). According to OECD (2008), aggregation of numeric variables in a ratio scale is more accurate using the geometric aggregation than the linear aggregation method. A various combination of the weighting and aggregation methods can be formed to construct the composite indicators (OECD, 2008). Therefore, this paper discussed the weighting method based on multiple correspondence analysis and factor analysis of mixed data and the linear and geometric aggregation method to construct the composite indicators of mixed data.

II. METHODS AND MATERIAL

This study used the the social protection program data collection (PPLS) 2011 collected by BPS and TNP2K. This study was focused on the households of Dramaga village, Bogor regency, Indonesia, which consists of 775 households and 3155 household members. The variables consist of 15 numeric variables and 19 categoric variables as presented in Table 1.

The study developed the five models as presented in Table 2. The steps in this study were as follows:

1. Standardize the numeric variables into a Z form.
2. Transform the numeric variables that were correlated negatively with the PPLS 2011 household welfare status into a $1/x_{iq}$ form (Mazziota & Pareto, 2013).
3. Weighting uses multiple correspondence analysis, as below (Asselin, 2009):
   a. Transform the numeric values into the categoric values form using the quantile method (Becker, Chambers & Wilks, 1988).
   b. Perform the multiple correspondence analysis (Pages, 2015).
   c. Choose the factors which give the eigenvalue* larger than 1 (Asselin, 2009).
   d. Choose the loading factor which gives the largest discriminant value of the variable for every modality k_j and categoric variable *j.*
   e. Calculate the weight of the modality k_j and categoric variable *j.*
4. Weighting uses factor analysis of mixed data:
   a. Perform the factor analysis of mixed data (Pages, 2004).
   b. Weighting the numeric variables, as below (OECD, 2008):
      i. Choose the factors which give the cumulative variance larger than 60% and the eigenvalue larger than 1.
      ii. Choose the largest loading factor for every numeric variables *q.*
      iii. Calculate the weight of the numeric variables q (w_q).
      iv. Transform the weight into the value with a range [0,1] and total 1.
   c. Weighting the categoric variables using the same method on points 3.c – 3.e.

Table 1. Variables

Variable	Description	Type	Variable	Description	Type
X_1	Age of household head	Numeric	X_{18}	Educational level of hh head	Categoric
X_2	1/Dependency ratio	Numeric	X_{19}	Working status of household head	Categoric
X_3	Net elementary and middle school enrolment ratio	Numeric	X_{20}	Main occupational sector of household head	Categoric
X_4	1/Gross elementary and middle school enrolment ratio	Numeric	X_{21}	Status of residence mastery	Categoric
X_5	1/Household size	Numeric	X_{22}	Wall material	Categoric
X_6	At least one of the household members graduated from middle school	Numeric	X_{23}	Roof material	Categoric
X_7	At least one of the household members graduated from high school	Numeric	X_{24}	Source of drinking water	Categoric
X_8	At least one of the household members graduated from college	Numeric	X_{25}	Way of getting drinking water	Categoric
X_9	1/Number of school-aged child in elementary school	Numeric	X_{26}	Source of main lighting	Categoric
X_{10}	1/Number of school-aged child in middle school	Numeric	X_{27}	Toilet facility	Categoric
X_{11}	1/Number of sc.-aged child in high sc.	Numeric	X_{28}	Final stool disposal site	Categoric
X_{12}	Number of school-aged child in college	Numeric	X_{29}	Refrigerator ownership	Categoric
X_{13}	Proportion of household members working in the primary sector	Numeric	X_{30}	Motorcycle ownership	Categoric
X_{14}	Proportion of household members working in the secondary sector	Numeric	X_{31}	Main job position of household head	Categoric
X_{15}	Proportion of household members working in the tertiary sector	Numeric	X_{32}	Floor material	Categoric
X_{16}	Sex of household head	Categoric	X_{33}	Sector and main job position of h.h.	Categoric
X_{17}	Marital status of household head	Categoric	X_{34}	Working status of household head and residence mastery status	Categoric

Table 2. Models

Model	Weighting Method based on	Aggregation Method	
		Numeric Variables	Categoric Variables
1	Multiple correspondence analysis	-	Linear
2	Multiple correspondence analysis	-	Geometric
3	Factor analysis of mixed data	Linear	Linear
4	Factor analysis of mixed data	Geometric	Geometric
5	Factor analysis of mixed data	Geometric	Linear

5. Aggregation for the five models to construct the composite index and to determine the welfare status, as below:
   a. Sort the composite index from the smallest to the largest value.
   b. Split the index into four status.
6. Validate the five models using four methods:
   a. The Mann-Whitney test (Daniel, 1990).
   b. The robust analysis ($\bar{R}_s$) (OECD, 2008).
   c. The classification accuracy test (Foody, 2002).
   d. The Area Under the ROC Curve (AUC) (Hand, 2001).
7. Choose the best model of the five models based on the accepting H_0 on the Mann-Whitney test, the smallest of $\bar{R}_s$, the largest of the classification accuracy, and the largest of the AUC.

III. RESULTS AND DISCUSSION

Data Description

The households classified into four status, i.e. status 1 consists of 108 (13.94%), status 2 consists of 224 (28.90%), status 3 consists of 408 (52.65%), and status 4 consists of 35 (4.52%). The low correlation of the

numeric variables is between the variables of X_8 and X_3 (0.33), and the high correlation is between the variables of X_6 and X_7 (0.71).

In the composite indicators, the numeric variables should have a positive correlation with the composite index (Mazziota & Pareto, 2013). The correlation between the numeric variables and the composite index are presented in Table 3 and it showed there are six numeric variables with negative correlation so it should be transformed to $1/x_{iq}$ for a positive correlation.

Table 3. Correlation between the numeric variables and the composite index

	Numeric Variables														
Before Transformation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
	0.15	**-0.21**	0.07	**-0.19**	**-0.64**	0.20	0.34	0.08	**-0.31**	**-0.21**	**-0.04**	0.10	0.49	0.31	0.44
After Transformation	X_1	X_2^*	X_3	X_4^*	X_5^*	X_6	X_7	X_8	X_9^*	X_{10}^*	X_{11}^*	X_{12}	X_{13}	X_{14}	X_{15}
	0.15	**0.23**	0.07	**0.19**	**0.57**	0.20	0.34	0.08	**0.31**	**0.22**	**0.04**	0.10	0.49	0.31	0.44

Weighting Variables

The weighting variables using 2 weighting methods, they were the weighting method based on (1) multiple correspondence analysis, and (2) factor analysis of mixed data.

1. Weighting based on Multiple Correspondence Analysis (MCA)

The weighting based on multiple correspondence analysis can only be used for the categoric variables, so the numeric variables were categorized using the quantile method with the categories 4 or less. The weighting was conducted to all categories of the categoric variables. Based on the category weighting procedure, the first 12 factors with the eiganvalue larger than 1 were selected.

The weighting process used one of 12 loading factors with the largest variable discriminant value. The weights were calculated based on the different between the loading factor and the loading factor of the worst category, then the weights were divided by the square root of eigenvalue, as presented in Table 4 and 6.

Table 4. Weight of the variables of X_1 - X_{15}

Variable	**Categories**	**Weight**	**Variable**	**Categories**	**Weight**	**Variable**	**Categories**	**Weight**
X_1	$X_{1.1}$	0.00		$X_{5.4}$	1.54	X_{11}^*	$X_{11.1}$	0.00
	$X_{1.2}$	0.32	X_6	$X_{6.1}$	0.00		$X_{11.2}$	5.23
	$X_{1.3}$	1.52		$X_{6.2}$	0.87		$X_{11.3}$	5.99
	$X_{1.4}$	3.72		$X_{6.3}$	1.54	X_{12}	$X_{12.1}$	0.00
X_2^*	$X_{2.1}$	0.00		$X_{6.4}$	2.17		$X_{12.2}$	0.55
	$X_{2.2}$	0.55	X_7	$X_{7.1}$	0.00	X_{13}	$X_{13.1}$	0.00
	$X_{2.3}$	3.06		$X_{7.2}$	0.40		$X_{13.2}$	0.65
	$X_{2.4}$	2.78		$X_{7.3}$	1.62		$X_{13.3}$	0.82
X_3	$X_{3.1}$	0.00		$X_{7.4}$	1.29		$X_{13.4}$	1.00
	$X_{3.2}$	1.82	X_8	$X_{8.1}$	0.00	X_{14}	$X_{14.1}$	0.00
	$X_{3.3}$	3.68		$X_{8.2}$	7.02		$X_{14.2}$	0.64
	$X_{3.4}$	4.91	X_9^*	$X_{9.1}$	0.00		$X_{14.3}$	0.93
X_4^*	$X_{4.1}$	0.00		$X_{9.2}$	0.13		$X_{14.4}$	0.46
	$X_{4.2}$	0.52		$X_{9.3}$	0.38	X_{15}	$X_{15.1}$	0.00
	$X_{4.3}$	1.82		$X_{9.4}$	2.56		$X_{15.2}$	0.37
	$X_{4.4}$	2.59	X_{10}^*	$X_{10.1}$	0.00		$X_{15.3}$	0.48
X_5^*	$X_{5.1}$	0.00		$X_{10.2}$	4.23		$X_{15.4}$	0.60
	$X_{5.2}$	0.11		$X_{10.3}$	3.98			
	$X_{5.3}$	5.13		$X_{10.4}$	4.41			

2. Weighting based on Factor Analysis Of Mixed Data (FAMD)

The weighting based on factor analysis of mixed data can handled both the numeric and categoric variables. The first step was weighting the numeric variables. Based on the numeric weighting procedure, the first 17 factors with the cumulative variance larger than 60% and the eigenvalue larger than 1 were selected.

Table 5. Weight of the numeric variables

Variables	Weight	Variables	Weight	Variables	Weight	Variables	Weight	Variables	Weight
X_1	0.192	X_4^*	0.024	X_7	0.070	X_{10}^*	0.038	X_{13}	0.091
X_2^*	0.028	X_5^*	0.140	X_8	0.020	X_{11}^*	0.024	X_{14}	0.082
X_3	0.018	X_6	0.095	X_9^*	0.095	X_{12}	0.005	X_{15}	0.077

Table 6. Weight of the variables of X_{16} - X_{34}

Variables	Categories	Weight		Variables	Categories	Weight	
		MCA	FAMD			MCA	FAMD
X_{16}	Female	0.00	0.00	X_{27}	No toilet	0.00	0.00
	Male	0.92	9.46		Public	1.22	5.15
X_{17}	No married	0.00	0.00		Self-owned	0.48	0.00
	Married	0.44	9.16	X_{28}	Others	0.00	6.55
X_{18}	Elementary school	0.00	0.00		Holes	0.87	10.24
	Middle school	0.16	5.01		River/lake/sea	0.67	0.00
	High school	1.21	7.20		Septic tank	1.08	1.87
	College	1.23	3.96	X_{29}	No	0.00	2.79
X_{19}	No working	0.00	0.00		Yes	0.85	0.00
	Working	0.14	14.31	X_{30}	No	0.00	6.38
X_{20}	No working	0.00	0.00		Yes	1.07	4.20
	Tertiary	0.06	14.11	X_{31}	No working	0.00	6.53
	Secondary	1.45	15.54		Others	0.32	0.00
	Primary	2.56	12.07		Laborer/employee	1.23	2.60
X_{21}	Others	0.00	0.00		Self-employed	1.27	0.00
	Free rental	2.47	65.41	X_{32}	Soil	0.00	0.00
	Contract/lease	0.87	76.34		No soil	0.16	6.07
	Self-owned	1.81	67.81	X_{33}	No working	0.00	0.00
X_{22}	Others	0.00	0.00		Tertiary and others	0.14	13.74
	Wood	0.97	5.86		Tertiary and laborer/ emp.	0.07	17.65
	Wall	1.35	4.67		Tertiary and self-employed	0.79	12.76
X_{23}	Others	0.00	0.00		Secondary and others	0.70	15.29
	Asbestos	6.64	13.18		Secondary and laborer/emp.	0.50	17.31
	Tiles	6.57	13.19		Secondary and self- employed	2.60	12.44
X_{24}	Others	0.00	0.00		Primary and others	0.96	11.79
	Unprotected wells	0.18	1.61		Primary and laborer/emp.	1.36	17.82
	Protected wells	0.34	3.09		Primary and self-employed	2.67	2.67
	Drilling wells	1.00	0.91	X_{34}	No working and free rental	0.00	0.00
	Tap water	0.00	1.84		No working and contract/lease	0.79	0.79
	Bottled water	1.06	3.39		No working and self-owned	0.98	0.98
X_{25}	No buying	0.00	0.00		Working and others	14.64	14.64
	Buying	0.47	0.47		Working and free rental	0.38	0.38
X_{26}	No electricity	0.00	0.00		Working and contract/lease	0.12	0.12
	PLN without electric meter	0.76	0.76		Working and self-owned	0.62	0.62
	PLN with electric meter	0.11	0.11				

The weights were calculated by weighting the largest loading factor with the proportion of its variance, then they were transformed into the value with a range [0,1] and total 1. These transformed values were the weights as presented in Table 5.

The last step was weighting the categoric variables using the same method as the weighting based on multiple correspondence analysis, Based on the category weighting procedure, the first 11 factors with the eiganvalue larger than 1 were selected. The weights presented in Table 6.

Aggregation Variables

The weights of the variables were used in the next step that is aggregation the variables. The aggregation was to construct the composite index and to determine the household welfare status by the five models. The aggregation using the linear and/or geometric aggregation as presented in Table 2.

The composite index is the household welfare in the continuous form. The household welfare status was constructed by sorting the index from the smallest to the largest value and then split it into 4 classes of status (status 1, 2, 3, and 4) using the cuts off that referred to the PPLS 2011 as presented in Table 7. The range of the index for every status and the models are presented in Table 8. The composite index range is used to classify the new households after the aggregation process.

Table 7. Dramaga village household welfare status

Rank	Model 1		Model 2		Model 3		Model 4		Model 5		Status
	ID	Index	ID	Index	ID	Index	ID	Index	ID	Index	
1	5605909	0.86	5626513	1.41	5626513	4.96	5640175	-10.67	5648412	-2.23	1
2	5626513	0.92	5639988	1.47	5640083	5.27	5640030	-8.94	5640175	-1.76	1
3	5648415	0.92	5626430	1.47	5639988	5.48	5648942	-8.88	5640233	-1.54	1
4	5625806	0.92	5640011	1.5	5640095	5.58	5639980	-8.56	5626513	-1.22	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
772	5625819	1.62	5606500	1.88	5605906	12.24	5639696	10.74	5606081	19.94	1
773	5640393	1.66	5639696	1.91	5606238	12.35	5606550	12.46	5639696	20.99	4
774	5606500	1.66	5644149	1.91	5648217	12.38	5640176	12.84	5640176	21.00	4
775	5640176	1.68	5626010	1.91	5639696	12.71	5626010	14.34	5626010	22.70	4

Table 8. Composite index range

	Model 1	Model 2	Model 3	Model 4	Model 5
Status 1	[0.86,1.09)	[1.41,1.62)	[4.96,7.17)	[-10.67,-4.79)	[-2.23,4.17)
Status 2	[1.09,1.20)	[1.62,1.74)	[7.17,10.34)	[-4.79,0.08)	[4.17,8.09)
Status 3	[1.20,1.45)	[1.74,1.85)	[10.34,11.56)	[0.08,8.61)	[8.09,16.44)
Status 4	[1.45,1.68]	[1.85,1.91]	[11.56,12.71]	[8.62,14.34]	[16.44,22.70]

Validation of the five models of the household welfare status to the PPLS 2011 household welfare status using 4 methods, i.e. (1) Mann-Whitney test, (2) robust analysis ($\bar{R}_s$), (3) classification accuracy, and (4) *Area Under the ROC Curve* (AUC), as presented in Table 9.

Table 9. Validation models

	Model 1	Model 2	Model 3	Model 4	Model 5
P-Value of Mann-Whitney	1.00	0.02	1.00	1.00	**1.00**
$\bar{R}_s$	0.50	0.66	0.70	0.46	**0.43**
Accuracy (%)	53.81	42.45	44.00	57.68	**60.13**
AUC	0.77	0.68	0.59	0.78	**0.81**

Based on the four tests, the model 5 is the best model because it accepted H_0 on the Mann-Whitney test which means that all status have different characteristics, the smallest of $\bar{R}_s$ (0.43), the largest of

the classification accuracy (60.13%), and the largest of the AUC (0.81). low and high correlation and at least in a ratio scale.

The weighting method based on factor analysis of mixed data provided more accurate results than the weighting method based on multiple correspondence analysis, with the test statistics 0.33. On the numeric variables, the geometric aggregation method provided more accurate results than the linear aggregation method. While on the categoric variables, the linear aggregation method provided more accurate results than the geometric aggregation method.

IV. CONCLUSION

The best model to construct the Dramaga village household welfare status was the model using the factor analysis of mixed data as the weighting method and the geometric aggregation method for the numeric variables and the linear aggregation method for the categoric variables.

V. REFERENCES

[1]. Asselin, L.M. (2009). Analysis of Multidimensional Poverty: Theory and Case Studies. Economic Studies in Inequality, Social Exclusion and Well-Being, 7(1), 19-51

[2]. Becker, R.A., Cahmbers, J.M. & Wilks, A.R. (1988). The New S Language: A Programming Environment for Data Analysis fnd Graphics, California: Wadsworth & Brooks/Cole

[3]. Castano, E. (2002). Proxy Means Test Index for Targeting Social Programs: Two Methodologies and Empirical Evidence. Lecturas de Economia, 56(1), 135-144

[4]. Daniel, W.W. (1990). Applied Nonparametrics Statistics. Boston: PWS-KENT

[5]. Foody, G.M. (2001). Status Of Land Cover Classification Accuracy Assessment. Remote Sensing of Environment, 80(1), 185-201

[6]. Hand, D.J. (2001). A Simple Generalisation of The Area Under The ROC Curve for Multiple Class Classification Problems. Machine Learning, 45(2), 171-186

[7]. Mazziotta, M., & Pareto, A. (2013). Methods for Constructing Composite Indices: One for All or All for One?. Rivista Italiana di Economia Demografia e Statistica, 67(2), 67-80

[8]. Organisation for Economic Co-Operation and Developement [OECD]. (2008). Handbook on Constructing Composite Indicators Methodology and User Guide, Paris: Author

[9]. Pages, J. (2004). Analyse Factorielle de Donnees Mixtes. Revue de Statistiquee Appliquee, 52(4), 93-111

[10]. Pages, J. (2015). Multiple Factor Analysis by Example Using R. Rennes: CRC Press

[11]. Sari, W.J. (2011). Pembentukan Indikator Sasaran dengan Proxy Means Test Berdasarkan Metode Princals, Depok: Universitas Indonesia

[12]. Tim Nasional Percepatan Penanggulangan Kemiskinan [TNP2K]. (2013). Pembangunan Basis Data Terpadu untuk Mendukung Program Perlindungan Sosial, Jakarta: Author