Optimum scoring scheme to classify villages into urban-rural group

Optimum scoring scheme to classify villages into urban-rural group

R P Manik[1], Herlina[1], A H Wigena[1], S R Surbakti[2] and B Sartono[1*]

[1]Department of Statistics, IPB University, Bogor, 16680, Indonesia
[2]BPS-Statistics Indonesia, Jakarta, 10710, Indonesia

*E-mail: bagusco@ipb.ac.id

Abstract. The scoring method has been used by BPS-Statistics Indonesia since the 1980s to classify urban/rural areas. Currently, the 2010 scoring method uses total score 10 as a threshold to classify villages to urban-rural status. If the total score more than or equal 10, the villages are classified as urban, and rural otherwise. Applying the 2010 scoring method on raw data of *Pendataan Potensi Desa* (PODES) 2008 and 2014 shows 1266 villages change from urban to rural. Therefore, it is necessary to evaluate the determinant of predictors and the criteria of each predictor. The purpose of this research is to show the optimum scoring method from several optimizations that change the predictors and several optimizations that add new predictors. Exploratory Data Analysis (EDA) used to obtain the predictors and scores for each new criterion. In relation to this research problem, optimization is used to get the best results under given constraint. The constraint of the optimization carried out is the assumption that the changes in rural to urban status are increasing, and the changes in urban to rural are not existing. The optimum scoring method obtained from this study is the one excluding cinema (X8), changing the criteria of percentage of households with cable phone (X11) and percentage of households with electricity (X12), replacing predictor hotels (X10) into a starred hotel and adding minimarket as a new predictor. This optimization uses 12 variables with threshold 10. The implication of this study for future research is the use of more advanced statistical methods than EDA to determine the criteria of each predictor.

1. Introduction

Scoring methods have been used by BPS since the 1980s to do urban-rural classifications for the villages in Indonesia. Currently, the 2010 scoring method uses total score 10 as a threshold to classify villages to urban-rural status. If the total score more than or equal 10, the villages are classified as urban, and rural otherwise [1]. The variables that used to get score are population density, percentage of agricultural households and existence/access to reach urban facilities [1]. Accessibility to reach urban facilities uses the criteria of distance from the village to the facilities. The detail of the current scoring method presented in Table 1.

This research was motivated by the results of data exploration by applying the 2010 scoring method on Table 1 to determine the urban-rural status of villages in Indonesia based on *Pendataan Potensi Desa* (PODES) 2008 and 2014 datasets. This exploration concluded that the current scoring method still shows the number of villages with urban status increasing and on the other hand the number of villages with status rural declining. The proportion of urban villages in 2008 was 20.09 percent (13387 villages) and increased to 27.24 percent (18224 villages) in 2014. This is in line with the concept of regional development. However, exploration also shows that there were 1266 villages that changed status from urban in 2008 to rural in 2014 (Table 2). This concludes that a new scoring method is needed.

Research relating to the evaluation of rural-urban classifications in Indonesia is still limited. Research [1] is limited to the determination of the variables that distinguish villages into urban-rural status. There is no research that modifies the right criteria and score in the current scoring method. Modifications should not only be on predictor variables but also criteria and scores on each predictor variable. So that it is in line with the current village development. On the other hand, the current scoring method is not modified too long (since 2010). For example, Scotland evaluates the classification of urban-rural areas every two years [6].

Table 1. Guidelines to urban-rural classification with criteria and scores in 2010

1. Population Density		2. Percentage of Agricultural Household		3. Access to Urban Facility		
Criteria	Score	Criteria	Score	Urban Facility	Criteria	Score
<500	1	>70.00	1	a. Kindergarten	- Have or 2.5 km - > 2.5 km	1 0
500-1,249	2	50.00-69.90	2	b. Junior High School		
1,250-2,499	3	30.00-49.99	3	c. Senior High School		
2,500-3,999	4	20.00-29.99	4	d. Traditional Market	- Have or ≤ 2 km - > 2 km	1 0
4,000-5,999	5	15.00-19.99	5	e. Mall/Shopping Complex		
6,000-7,499	6	10.00-14.99	6	f. Cinema	- Have or ≤ 5 km - > 5 km	1 0
7,500-8,499	7	5.00-9.99	7	g. Hospital		
> 8,500	8	< 5.00	8	h. Hotel/Pub/Beauty shop	- Have - Not Have	1 0
				i. Percentage of household with cable phone	- ≥ 8.00 - < 8.00	1 0
				j. Percentage of household with electricity	- ≥ 90.00 - < 90.00	1 0

Data exploration was also carried out by paying attention to the data distribution (i.e. density, percentage, and distance) and the mean of a score of each variable. Exploration on data distribution shows that the variables of population density, percentage of agricultural households, percentage of households having telephone and the percentage of households with electricity have different data patterns between 2008 and 2014 so that these variables require new criteria.

Table 2. Transition matrix of the urban-rural status of the villages in 2008 and 2014 by applying 2010 scoring method

		2014		Total
		Rural	Urban	
2008	Rural	47412	6103	53515
	Urban	1266	12121	13387
	Total	48678	18224	66902

Further, based on the mean of score of each variable, it is known that the existence of kindergartens, the existence of junior high schools, the existence of cinemas, the percentage of households having telephone and the percentage of households with electricity tend to be homogeneous because they are close to score 0 (minimum) or score 1 (maximum). When data is homogeneous, variables become more difficult to be a differentiator.

There are some researchers that criticized this scoring method by proposing some new variables. [2] recommended the existence of internet rental and the existence of the bank as the new alternative determinant variable. [3] proposed the existence of internet access, the existence of lighting on the village main road, the existence of bank and existence to the mini market. Furthermore, the use of the number of fewer variables was able to predict that were as good as with many variables in urban-rural classifications [2]. Moreover, the addition of new predictor variables can increase the level of classification accuracy [2].

This study aims to propose a new scoring method by (i) changing variable; (ii) changing the categorized data criteria; (iii) introducing new variable; and (iv) optimizing the score threshold.

Material

The dataset of this work was obtained from the raw data on *Pendataan Potensi Desa* (PODES) 2008 and 2014.

Table 3. Variable of urban-rural classification

Variable	Variable Name
Y	Urban-rural status in 2014
X1	Population density
X2	Percentage of agriculture household
X3	Existence/Access of kindergarten
X4	Existence/Access of junior high school
X5	Existence/Access of senior high school
X6	Existence/Access to market
X7	Existence/Access of shopping complex
X8	Existence/Access of cinema
X9	Existence/Access of hospital
X10	Existence/Access of hotel/inn/pub/beauty salon
X11	Percentage of households having a telephone
X12	Percentage of households with electricity
X13	Number of kindergartens (TK)
X14	Number of junior high school (SMP)
X15	Number of senior high school (SMA)
X16	Number of minimarkets (Minimarket)
X17	Existence of street light in the village
X18	Existence of internet
X19	Existence of hotel
X20	Existence of inn
X21	Existence of pub

X1 to X12 are variables that have been used in the current scoring method [1]. Meanwhile, X13 to X21 is the new variables proposed [2],[3]. Step of pre-processing was done to get suitable data on a transition matrix, i.e. the villages on the raw that is available on 2008 and 2014 period.

2. Methods

The data analysis was conducted in several steps as follow:

2.1. Exploratory data analysis (EDA)

One or more data sets were analyzed to obtain the main characteristics of the data, including using visual methods or graphical representation by EDA [4]. The main characteristic of data, i.e. central tendency and distribution, we collect a summary of existing observation values [5]. In this paper, the results of data exploration are useful for obtaining predictor variables and criteria for each predictor variable that is in line with current regional developments. In the end, they can be used to obtain urban-rural classifications that are relevant to current conditions. The objective of EDA in this paper can be summarized as follows: (i) Observe data distribution of each predictor variables in the current scoring method. We take the necessary characteristics of quantitative data for each predictor variable, e.g. minimum, quintile, median and maximum as a summary of the existing observation values; (ii) Observe data distribution of new variable to determine the score criteria. We use boxplot of each new variable; (iii) Then we propose new criteria and variable by justifying the proposed changes.

2.2. Optimization

Optimization can be interpreted as a series of activities to get the best results under the given conditions. Optimization carried out, in this paper, is divided into optimization by changing the variables criteria and optimization by adding new variables. Table 4 summarized all optimization applied.

Table 4. Optimization design

Scenario	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	Number of variables
1	√	√	√	√	√	√	√		√	√	√	√										10
2	√	√	√	√	√	√	√		√	√		√										10
3	√	√			√	√	√		√	√	√	√										9
4	√	√			√	√	√		√	√		√										8
5	√ new	√ new		√	√	√	√		√	√	√ new	√ new										11
6	√	√				√	√		√	√	√	√	+	+	+							11
7	√	√	√	√	√	√	√		√		√	√							+	+		12
8	√	√	√	√	√	√	√		√		√	√							+			11
9	√	√	√	√	√	√	√		√		√	√								+		11
10	√	√	√	√	√	√	√		√		√	√									+	11
11	√	√	√	√	√	√	√		√		√	√						+	+			12
12	√	√	√	√	√	√	√		√		√	√				+			+			12
13	√	√	√	√	√	√	√		√		√	√					+		+			12
14	√ new	√ new	√	√	√	√	√		√		√ new	√ new				+			+			12
15	√ new	√ new	√	√	√	√	√		√		√ new	√ new						+	+			12

Note: √ variable in the 2010 scoring system; √ **new** using the proposed new criteria;
+ new variable

The variables X1 to X12 are variables that have been used in the current scoring method. Meanwhile, the variables X13 to X21 are the new variables proposed. Scenarios 1 to 5 is optimization by changing variable criteria. Scenarios 6 to 15 are optimizations by adding new variables.

Optimization requires a constraint function as an evaluation of the optimization process carried out. The constraint of the optimization in this paper is the increasing number of villages from rural becomes urban, and small (even zero) amounts in urban becomes rural. The application of scoring methods in the 2008 and 2014 data shows that there were 1266 villages that changed their status from the previous urban villages to rural villages (see Table 2). Thus, the optimum scoring method was proposed when the optimization results showed that the number of villages that had changed their status from urban villages to rural villages was smaller than 1266 villages and the proportion of villages with urban status was less than 30 percent (30 percent assumed from the current proportion of 27.24 percent). The optimization process is carried out by combining the old criteria and new criteria simultaneously. Then the total score threshold is tried from 8 to 15.

Technically, the optimization stage is as follows: the 2008 data still uses the current scoring method. Whereas, the proposed changes, both the criteria and the addition of new variables according to the scenario (Table 4), are applied to 2014 data. The results of the classification are summarized in the transition matrix. Then the selection of scenarios is done based on the constraints of optimization.

3. Results and Discussion

3.1. Justification of proposed changes

Determination of new criteria, including the number of criteria, is done by observing a summary of data distribution for each predictor variable. Previous exploration (see Introduction) showed that population density (X1), percentage of agricultural household (X2), percentage of households having a telephone (X11) and percentage of households with electricity (X12) require new criteria.

Table 5. Summary of data distribution in 2014

Variable	Min	1st Quintile	Median	Mean	3rd Quintile	Max
X1	0.09	78.97	347.25	1484.03	1116.55	62642.42
X2	0.00	37.93	71.85	63.14	91.98	100.00
X11	0.00	0.00	0.00	1.60	0.00	100.00
X12	0.00	94.68	100.00	90.83	100.00	100.00

The distribution of X1 in Table 5 shows that there is 75 percent of villages that have a population density of up to 1116.55 people per km^2. So, only 25 percent of villages are more than that. If the data are applied to the 2010 scoring method (Table 6 column of old criteria), then very few villages will get a score between 4 and 8. So, population density in groups of scores 4 to 8 can be combined into 1 group that gets the maximum score, i.e. population density ≥ 2500. While population density <2500, is categorized again. By applying 7 criteria, the range up to 2499 is divided by 7 so we get an interval of 350.

The distribution of X2 shows that 50 percent of villages have a percentage of agricultural households between 70 and 100. So, if the data are applied to the 2010 scoring method, half of the villages have a score of 1 and there are score criteria that have little / no observation frequency. So, X2 data is categorized again following the existing data distribution. Data up to quintile 1 (37.93 sets to 45) is given a score of 4. Then between the 1 to 1 quintile (71.85 set to 65), the quintile is given a score of 3. Data between quintiles 2 to quintile 3 (91.98 sets to 95) given a score of 2. And the rest, the data between quintiles 3 to quintiles 4 (100) is given a score of 1.

The distribution of X11 shows that more villages have a low percentage of households having a telephone. This is shown in the value of quintile 3 of 0.00 percent. This means that 75 percent of the villages do not have a telephone anymore. The thing that should be suspected as a reason is that more people are moving from cable telephone technology to cellular phones. If the X11 data are applied to the 2010 scoring method, almost all villages will not get a score. So, X11 data is categorized again following the distribution of existing data. Villages that have a percentage of telephone households more than or equal to 1.6 (rounded 2) are given a score of 1 and vice versa get a score of 0.

The distribution of X12 (Table 5) shows more villages with a high percentage of households have electricity, even exceeding 90 percent. This is indicated by the value of the quintile 1 of 94.68 percent. This means that 75 percent of villages already have a percentage of households with the electricity of more than 94.68 percent. Of course, this is a good development because more people are getting access to electricity. If the X12 data applies the 2010 scoring method, almost all villages will get a score of 1.

So a 90 percent score criterion cannot make a difference. So, X12 data is categorized again following the distribution of existing data. Villages that have a percentage of households with electricity more than or equal to 94.68 (rounded 95) are given a score of 1 and vice versa get a score of 0. By increasing the score criterion, the percentage of households with electricity can be a differentiator. As well as raising the standard of the urban size of a village. From the exploration results above X1 still uses 8 criteria and X2 uses 4 criteria. X11 and X12 each use 2 criteria (Table 6).

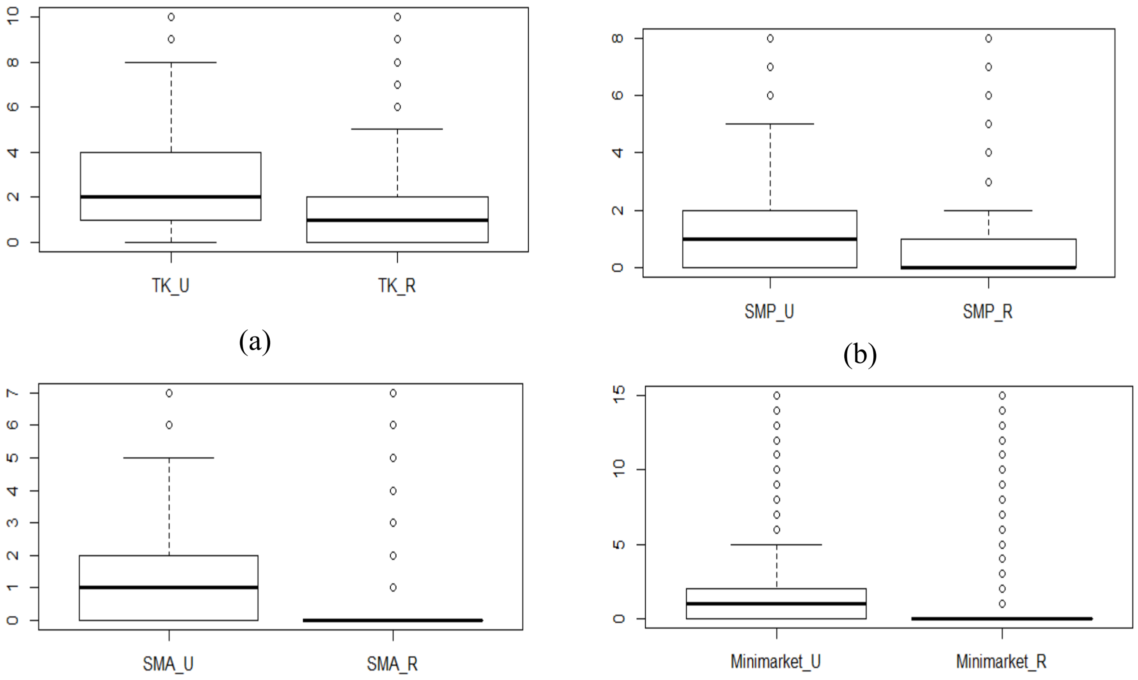
Table 6. The proposed new criteria for variable

	Old criterion	Old score	New Criterion	New score
X1	<500	1	<350	1
	500-1249	2	350-699	2
	1250-2499	3	700-1049	3
	2500-3999	4	1050-1399	4
	4000-5999	5	1400-1749	5
	6000-7499	6	1750-2099	6
	7500-8499	7	2100-2449	7
	≥ 8500	8	≥ 2500	8
X2	≥ 70.00	1	≥ 95.00	1
	50.00-69.99	2	65.00-94.99	2
	30.00-49.99	3	45.00-64.99	3
	20.00-29.99	4	< 45.00	4
	15.00-19.99	5		
	10.00-14.99	6		
	5.00-9.99	7		
	< 5.00	8		
X11	≥ 8.00	1	≥ 2.00	1
	< 8.00	0	< 2.00	0
X12	≥ 90.00	1	≥ 95.00	1
	< 90.00	0	< 95.00	0

The use of new variables requires determining the score criteria. The number of kindergartens, the number of junior high school, the number of senior high school and the number of minimarkets (X13 to X16) are new variable with continuous data. Then data exploration needs to be done to determine the score criteria. Data exploration was carried out by looking at the data distribution of each predictor variable that was differentiated according to urban and rural status (Y) in 2014.

Figures 1 with label '_U' is for distribution data in urban group, then with label 'R' in the rural group. Figures 1.a. shows the number of kindergartens (X13) in a village with an urban status more possible to have more kindergarten (median 2 and 3rd quintile 4) than a village with rural status (median 1 and 3rd quintile 2). The criteria and scores for the variable number of kindergarten units can be set as follows: when there is no kindergarten and the closest kindergarten distance of more than 2.5 km is given a score of 0, when there is no kindergarten and the closest kindergarten distance is less than or equal to 2.5 km given a score of 1, when the number of TK units in villages between 1 and 2 are given a score of 2, and when the number of TK units in the village is at least 3 given a score of 3.

With the same way, the criteria and scores for the number of junior high school units (SMP) can be set as follows: when there is no SMP and the distance of the closest SMP more than 2 km is given a score of 0, when there is no SMP and the distance of the nearest SMP is less than or equal to 2 km or there is a minimum of 1 SMP given a score 1, and when the number of SMP units is at least 2 in the village given a score of 2.



(a)

(b)

(c)

(d)

Figure 1. Boxplot X13, X14, X15, X16 in (a) to (d). Each variable is clustered by current urban-rural status.

Criteria and scores for the number of senior high school units (SMA) can be set as follows: when there is no SMA and the distance of the closest high school of more than 2.5 km is given a score of 0, and when there is no SMA and the distance of the nearest SMA is less than or equal to 2.5 km or there is a minimum of 1 SMA given score 1.

The criteria and score for the number of minimarket units can be set as follows: when there is no minimarket a score of 0 is given, when the number of minimarkets in the village between 1 to 2 is given a score of 1, when the number of minimarkets in the village between 3 to 4 is given a score of 2, and when the number of minimarkets in a village of at least 5 is given a score of 3.

Furthermore, variable hotels, inns, pubs, internet access and street lights (X17 to X21) are explained by the existence so that the score criteria are ‘there’ = 1 and ‘none’ = 0.

3.2. The results of the scenario analysis

The optimization process was carried out in 15 scenarios (table 4) and tried thresholds ranging from 8 to 15. Experiments in the threshold range of 8-15 made many scenario combinations. Table 7 shows only a part of all scenarios that are carried out with an optimal scenario. The constraints are the number of villages that changed their status from the previous urban villages to be smaller than 1.266 and the urban percentage was less than 30 percent.

The ‘RR’ in Table 7 is the number of villages with rural status in 2008 to remain in rural status in 2014. The ‘RU' is the number of villages with rural status in 2008 to become urban in 2014. The ‘UR' is the number of villages with urban status in 2008 to become rural in 2014. And The ‘UU' is the number of villages with urban status in 2008 to remain urban in 2014. ‘Rural' is the number of villages with rural status in 2014 and the ‘Urban' is the number of urban villages in 2014. The total of first 4 columns (RR, RU, UR, and UU), which is 66,902, equal to the total of columns 5 and 6 (Urban and Rural) in each row of a table. Then, optimum rank is the order of UR starts from the smallest UR. The best optimization is on the optimum rank of 1

Table 7. The optimum scenario

Scenario	RR	RU	UR	UU	Rural	Urban	Optimum threshold	Optimum rank
1	46470	7045	1083	12304	47553	19349	10	
2	46508	7007	1099	12288	47607	19295	10	
3	45868	7647	1003	12384	46871	20031	8	5
4	45907	7608	1020	12367	46927	19975	8	
5	46101	7414	967	12420	47068	19834	11	4
6	46258	7257	1138	12249	47396	19506	11	
7	46470	7045	1084	12303	47554	19348	10	
8	46657	6858	1146	12241	47803	19099	10	
9	46559	6956	1112	12275	47671	19231	10	
10	46685	6830	1157	12230	47842	19060	10	
11	45941	7574	971	12416	46912	19990	11	
12	46022	7493	953	12434	46956	19946	10	3
13	47048	6467	1207	12180	48255	18647	11	
14	45983	7532	934	12453	46917	19985	10	1
15	45904	7611	948	12439	46852	20050	11	2

4. Conclusion

The best optimization among several scenarios is the 14th scenario, i.e. excluding cinema variables (X8), changing criteria of percentage of households having a telephone (X11) and percentage of households with electricity (X12), changing the X10 variable into star hotel (X19) and adding a variable of a minimarket. This optimization uses 12 variables with threshold 10.

In the next research, it is better to find new criteria based on more advanced statistical methods.

References

[1] [BPS] Badan Pusat Statistik. 2010. Peraturan Kepala Badan Pusat Statistik Nomor 37 Tahun 2010 tentang Klasifikasi Perkotaan dan Perdesaan di Indonesia. Badan Pusat Statistik Republik Indonesia. Jakarta (ID): BPS.

[2] Surbakti SR, Erfiani and Sartono B 2015 Alternative Determinant Variables in Urban/Rural Village Classification in Indonesia in *International Conference On Research, Implementation, and Education* (pp. 261–270). Yogyakarta, Indonesia. Retrieved from http://eprints.uny.ac.id/23643/

[3] Surbakti SR, Listianingrum T and Arsiani IK 2018 Improved Area Classification, a Fundamental Step to Support Inclusive Economic Statistics in *Asia-Pacific Economics Statistics Week* 2008. Bangkok, Thailand. Retrieved from http://communities.unescap.org/asia-pasific-economic-statistics/apes-2018-featured-papers

[4] Komorowski M, Marshall DC, Salciccioli JD and Crutain Y 2016 Exploratory Data Analysis in *Secondary Analysis of Electronic Health Records* (pp. 185–203). New York (US): Springer Publishing. https://doi.org/10.1007/978-3-319-43742-2_15

[5] Saefuddin A, Notodiputro KA, Alamudi A and Sadik K 2009 *Statistika Dasar* (Jakarta (ID): Grasindo).

[6] Granville S, Mulholland S and Stanisforth J 2009 Use and Understanding of The Scottish Government Urban Rural Classification. *Scottish: Scottish Government. Retreived* from https://www.webarchive.org.uk/wayback/archive/20180515181917mp_/http://www.gov.scot/Resource/Doc/281343/0084923.pdf