RUTCOR
RESEARCH
REPORT

RECONSTRUCTION OF WORLD BANK'S CLASSIFICATION OF COUNTRIES

Nima Mirzaei[a] Béla Vizvári[b]

RRR 08-2011, June 2011

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/∼rrr

[a]Department of Industrial Engineering, Eastern Mediterranean University, Famagusta, Mersin 10, Turkey, bela.vizvari@emu.edu.tr

[b]RUTCOR, Rutgers University, bvizvari@rutcor.rutgers.edu, and Department of Industrial Engineering, Eastern Mediterranean University, Famagusta, Mersin 10, Turkey, bela.vizvari@emu.edu.tr

Rutcor Research Report
RRR 08-2011, June 2011

RECONSTRUCTION OF WORLD BANK'S CLASSIFICATION OF COUNTRIES

Nima Mirzaei Béla Vizvári

Abstract. The main objective of this paper is to analyze if the classification of countries provided by World Bank (WB) can be reconstructed with a linear and/or integer programming model called Multi-Group Hierarchical Discrimination if only data published by WB are used. WB has a public database of countries containing economical-financial and political factors. The parameters of the model have been determined on a collection of 44 countries. The model has been verified on other 39 countries. Only four out of 39 countries were misclassified which shows the power the elaborated model. Logical Analysis of Data (LAD) also has analyzed the problem. The attempt of the reconstruction of the classification uses 19 indicators. An important by-product of the reconstruction is that the methods select the most important indicators. Interestingly, the result proves that the World Bank classification is not based only on GNI per capita. In addition, more criteria that are important have main role in classification of countries.

1 Introduction

Financial risk management has become an important subject for operation researchers at the beginning of 1990s, because it provides important information in the field of financial engineering [6]. Operations researchers, financial investigators, statisticians, and econometricians have proposed many practical approaches to measure and assess the financial risks. Most of the approaches depend on a probabilistic notion of financial risk defined as the variance of the expected return. Furthermore, numerous optimization techniques have projected for this purpose [2].

Financial risk management assists decision makers and financial managers to make effective financial decisions. It allows investors and financial managers to decide where, when and how to invest their funds. Moreover, in the large scale, global companies are able to make decision in which countries locate their new branches or invest their capital.

In this study countries were classified according to World Bank categorization. Subsequently, for operational and analytical purposes, the World Bank's main criterion for classifying countries is Gross National Income (GNI) per capita. In the past, World Bank used the Gross National Product (GNP) instead of GNI to classify countries. Based on GNI per capita, each country is categorized into one of four economic classes, low income (995 USD or less), middle income (which subdivided into two classes, lower middle 996-3,945 USD and upper middle 3,946-12,195 USD), and the last class is high income (12,196 USD or more). In addition to GNI per capita criterion, there are two other criteria that are utilized in classifying countries, which are as follows:

- **Geographic region:** Classifications reported for geographic region are for low income and middle-income countries only. Low income and middle-income countries are sometimes entitled as developing countries. The use of the term is inconvenient, but it does not mean that all the countries in that class are experiencing similar development or that other countries have reached a preferred or final phase of development. It is important to know that the classification by income does not necessarily reflect development status.
- **Lending category:** International Development Association (IDA) countries are those that had a per capita income in 2009 less than 1,165 USD and lack the financial ability to borrow from International Bank for Reconstruction and Development (IBRD). IDA loans are deeply concessional-interest-free loans and grants for economic growth and improve programs aimed at boosting living conditions.

It should be noted that World Bank publishes income classification every year on 1st of July. These official classifications are fixed during the World Bank's fiscal year until the end of next June. Countries remain in the predefined categories in which they are classified irrespective of any revisions of their income data.

In this study, we use two different methods to conclude some result about country risk rating. The first method is Multi-Group Hierarchical Discrimination (MHDIS) method, which is based on Multi Criteria Decision Aid (MCDA). MHDIS was suggested in 2000 by Zopounidis [7]. Multi-Group Hierarchical Discrimination (MHDIS) method classifies set of alternatives to the specified class. A set of additive utility functions is developed by linear and/or mixed integer programming. The alternatives are classified according to the value of the utility functions is above or below a certain threshold. For good summary, see [8].

The second method is a novel technique in risk management that is called Logical Analysis of Data (LAD). Logical Analysis of Data is a qualitative method that uses pair comparison between sets of alternatives.LAD generates sets of Boolean constraints such that the alternative belonging to one class must satisfy completely the constraints and the alternatives of the opposite class must not satisfy them completely. The basic description of LAD is given in [1]. It is application in a similar area is [4].

2 Data and Method

The data that are used in this study are taken from World Bank database website. According to the World Bank the countries under consideration are categorized into four classes by considering their income level:

High-income economies (class C4) are mostly European ones as well as United States, Canada, Australia, New Zealand, Japan, Hong Kong, etc.

Upper-middle income economies (class C3) are countries from Europe (e.g., Poland and Hungary), South and Eastern Asia, and Latin-South America.

Lower-middle income economies (class C2) are Eastern Europe, Asia, Africa, and South-Latin America.

Low-income economies (class C1) are mostly from Africa and Asia.

Criteria are selected according to their importance and effect on the economical and political situation of countries. Countries are chosen by considering the data availability of selected criteria of the alternatives (countries), it means that if an alternative does not have enough information about one or more criteria on data set, then the alternative is eliminated automatically from a sample of data. When there is no data related to some criteria then it is not possible to compare alternatives with each other and classify them in the organized classes. The number of classes depends to predefined ranges and can be vary, but according to the World Bank, countries are classified in to four groups according to their income level. The steps of filtrations are as follows: We selected all countries and factors which available in the World Bank database webpage as raw information. Totally, 241 countries and 43 factors (political and economic) are available in the World Bank website database. For each factor, the average of data starting from 1990 up to 2008 was considered for the analysis. At the end, by considering the best combination of countries and criteria and specified model, the most important factors are recognized which are mentioned in Table 1and 2. In each step of filtration, some countries or some factors were eliminated because of lack of data availability for those countries or factors.

In this study, 44 alternatives (countries) were selected for the analysis, and each country falls in a specified set (a or b) shown in Table 1.

No.	Country	Set /Class	No.	Country	Set/Class
1	Argentina	C3	23	Japan	a/C4
2	Australia	a/C4	24	Korea, Rep.	a/C4
3	Austria	a/C4	25	Luxembourg	a/C4
4	Bolivia	b/C2	26	Mexico	b/C3
5	Belgium	a/C4	27	Netherlands	a/C4
6	Brazil	b/C3	28	New Zealand	a/C4
7	Bulgaria	b/C3	29	Norway	a/C4
8	Canada	a/C4	30	Oman	a/C4
9	China	b/C3	31	Paraguay	b/C2
10	Colombia	b/C3	32	Poland	b/C3
11	Czech Republic	a/C4	33	Portugal	a/C4
12	Dominican Republic	b/C3	34	Russian Federation	b/C3
13	Ecuador	b/C2	35	Slovak Republic	C4
14	Denmark	a/C4	36	Spain	a/C4
15	Finland	a/C4	37	Sweden	a/C4
16	France	a/C4	38	South Africa	b/C3
17	Germany	a/C4	39	Switzerland	a/C4
18	Hungary	a/C4	40	Turkey	b/C3
19	Iceland	a/C4	41	Uruguay	b/C3
20	India	b/C2	42	Venezuela, RB	b/C3
21	Indonesia	b/C2	43	United Kingdom	a/C4
22	Italy	a/C4	44	United States	a/C4

Table 1. Countries with their income classes and sets

This classification constitutes the basis for the development of the appropriate country risk assessment model. The classes was divided in to two sets, the C1, C2 and C3 classes belong to set b and C4 belong to set **a**.

By considering the availability of data in the World Bank database, we have considered 19 indicators (criteria) for this study, including both economic and political factors. Each criterion has three levels in general form, which are low, medium and high. Although, some specified criterion was divided to six levels. The 19 indicators with their levels are shown in the Table 2.

Evaluation Criteria	Indicators	Levels
g_1	Electric power consumption (kWh per capita)	6
g_2	Energy use (kg of oil equivalent per capita)	3
g_3	Exports of goods and services (% of GDP)	3
g_4	Fertility rate, total (births per woman)	3
g_5	GDP (current USD)	3
g_6	GDP growth (annual %)	6
g_7	GNI per capita, Atlas method (current USD)	3
g_8	GNI per capita, PPP (current international dollar)	3
g_9	GNI, Atlas method (current USD)	3
g_10	GNI, PPP (current international dollar)	3
g_11	Gross capital formation (% of GDP) 6	
g_12	Imports of goods and services (% of GDP)	3
g_13	Inflation, GDP deflator (annual %)	6
g_14	Military expenditure (% of GDP)	3
g_15	Mobile cellular subscriptions (per 100 people) 6	
g_16	Net migration	3
g_17	Population growth (annual %)	6
g_18	Population, total	3
g_19	Surface area (sq. km)	6

Table 2. Criteria and indicators with their levels

3 Models of Multi Hierarchical Discrimination

The MHDIS method has been used to develop the model in this study as a non-parametric approach [3]. The problem involves two or more ordered groups of alternatives for comparison, also, this model is based on the regression analysis. The notation and formulas are described in the following pages [3].

f is the objective function of the basic model. It is the total error of the utility function in countries misclassification. It is to be minimized. Variable S is on the right hand side of constraints and is either a nonnegative constant number defining the gap of separation of the two classes or it is the objective function if perfect separation is possible. The results of the model are sensitive on its value. In the following section we will discuss more about the value of S.

Initially, a reference set A consisting of n alternatives $a_1, a_2, ..., a_n$, classified into q ordered classes $C_1, C_2, ..., C_q$ (C_q is preferred to C_{q-1}, C_{q-1} is preferred to C_{q-2}, etc.) is used for model development (i.e., training sample). The alternatives are described (evaluated) along a set of m evaluation criteria $g = ()g_1, g_2, ..., g_m)$. The evaluation of an alternative a on criterion g_i is denoted as $g_i(a)$ which is the level of a at alternative i. The set of criteria may include both criteria of increasing and decreasing preference. For example, high GDP is preferred to low GDP but in the case of inflation rate low one is preferred to high inflation for an alternative.

A criterion g_i is assumed to have p_i different levels, which are rank-ordered from the lower one g_{1i} (the least preferred value) to the higher one $g_i^{p_i}$ (the most preferred value). The number of criterion levels is specified according to the evaluations of the alternatives integrated in the training sample. In this model r_{ai} denotes the position of the evaluation of the alternative a on criteria g_i within the rank ordering of the criterion levels from the lower one g_{1i} to the higher one $g_i^{p_i}$. In general, W_{ij} is award function for the existing levels of the criteria and M_{ij} is penalty function for the missing levels of criteria. Index k indicates class of alternative, index i indicates criteria and index j indicates level of the criteria.

The basic model below is a linear programming model that is used in our study for classification [3]. The quantity of S can be a fixed positive value or it can be considered as a variable.

$$\min f = \sum_{a \in C_k} e(a) + \sum_{b \in \bar{C}_k} e(b) \tag{1}$$

subject to

$$\sum_{i=1}^{m} \sum_{j=1}^{r_{ai}-1} W_{ij} - \sum_{i=1}^{m} \sum_{j=r_{ai}}^{p_i-1} M_i j - e(a) \geq S \quad \forall a \in C_k \tag{2}$$

$$\sum_{i=1}^{m} \sum_{j=r_{bi}}^{p_i-1} M_{ij} - \sum_{i=1}^{m} \sum_{j=1}^{r_{bi}-1} W_{ij} - e(b) \geq S \quad \forall a \in \bar{C}_k \tag{3}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{p_i-1} M_{ij} = 1, \quad \sum_{i=1}^{m} \sum_{j=1}^{p_i-1} W_{ij} = 1 \tag{4}$$

$$S, W_i j, M_i j, e(a), e(b) \geq 0 \tag{5}$$

Note that in the original paper [3], the following clarification is not explained.

Lemma. If S is a variable in problem (1)-(5) then there is an optimal solution with $S = 0$.

Proof: Assume that S is positive, if all $e(a)$ and $e(b)$ is zero then the solution remains feasible if S is decreased to zero, the value of the objective function is not changed, i.e. the solution is still optimal. If S is positive and some errors (es) are positive, as well, then let:

$$\varepsilon = \min\{S, \min\{e(a) \mid e(a) > 0, a \in C_k\}, \min\{e(b) \mid e(b) > 0, b \in \bar{C}_k\}\} > 0.$$

Then all positive es and S can be decreased by ε. The constraints are still satisfied and the objective function is decreased by

$$\varepsilon(|\{a|e(a) > 0, a \in C_k\}| + |\{b|e(b) > 0, b \in \bar{C}_k\}|) \geq \varepsilon > 0.$$

Thus the previous solution was not optimal. □

If the optimal value of problem (1)-(5) is zero then the maximal separation gap for perfect classification can be obtained by the following model:

$$\max S \tag{6}$$

$$subject to \sum_{i=1}^{m} \sum_{j=1}^{r_{ai}-1} W_{ij} - \sum_{i=1}^{m} \sum_{j=r_{ai}}^{p_i-1} M_{ij} \geq S \quad \forall a \in C_k \tag{7}$$

$$\sum_{i=1}^{m} \sum_{j=r_{bi}}^{p_i-1} M_{ij} - \sum_{i=1}^{m} \sum_{j=1}^{r_{bi}-1} W_{ij} \geq S \quad \forall b \in \bar{C}_k \tag{8}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{p_i-1} M_{ij} = 1, \sum_{i=1}^{m} \sum_{j=1}^{p_i-1} W_i j = 1 \tag{9}$$

$$S, W_{ij}, M_{ij} \geq 0 \tag{10}$$

One drawback of models (1)-(5) and (6)-(10) is that although there are several classes in the underlying problem they separate the countries (alternatives) into two classes only. It is possible to classify them into the required classes by a simple model as follows. The model formalized four classes as the classification problem in question has four classes. However, the generalization is straightforward for more number of classes.

$$\min f = \sum_{a \in C_1 \cup C_2 \cup C_3} \sum e_{U(a)} + \sum_{a \in C_2 \cup C_3 \cup C_4} e_{L(a)} \tag{11}$$

subject to

$$L(a) = \sum_{i=1}^{m} \sum_{j=1}^{r_{ai}-1} W_{ij} - \sum_{i=1}^{m} \sum_{j=r_{ai}}^{p_i-1} M_{ij} \quad \forall a \in C_k \tag{12}$$

$$\left.\begin{array}{l} \forall a \in C_1 : U_1 \geq L(a) - e_{U(a)}, \\ \forall a \in C_2 : L_2 \leq L(a) + e_{L(a)}, U_2 \geq L(a) - e_{U(a)}, \\ \forall a \in C_3 : L_3 \leq L(a) + e_{L(a)}, U_3 \geq L(a) - e_{U(a)}, \\ \forall a \in C_4 : L_4 \leq L(a) + e_{L(a)} \end{array}\right\} \tag{13}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{p_i-1} M_i j = 1, \quad \sum_{i=1}^{m} \sum_{j=1}^{p_i-1} W_{ij} = 1 \tag{14}$$

$$U_1 + S \leq L_2, \;\; U_2 + S \leq L_3, \;\; U_3 + S \leq L_4 \tag{15}$$

$$\forall a, S, e_{U(a)}, e_{L(a)} \geq 0 \tag{16}$$

$$U_1, U_2, U_3, L_2, L_3 \;\; \text{are unrestricted} \tag{17}$$

4 Application to the training set

In the first step models (1)-(5) has been solved on the training set. The model contained 44 countries and S was fixed to zero. 19 criteria each with 3 levels were used in this model. All models in this study were solved by Lingo 12.0. As result of first computation, country 1(Argentina) and 39(Switzerland) were misclassified, it means that country 1 belonged to upper class or set a and country 39 belonged to the lower class or set b. Important criteria were 6, 8, 11, 13, 15, and 17 which are GDP growth, GNI per capita (PPP), Gross capital formation, Inflation (GDP deflator), Mobile cellular subscriptions, and Population Growth respectively, and total error is 0.666.

In the next analysis S was raised from 0 to 0.01 and number of criteria and their levels have been the same. As result, Lingo solution demonstrates that the total error increased to 0.7333 and countries 1, 4, 13, 39 which are Argentina, Bolivia, Ecuador, and Switzerland were misclassified. In addition, important criteria were 1, 6, 11, 13, 15, and 17 which are Electric power consumption, GDP growth, Gross capital formation, Inflation (GDP deflator), Mobile cellular subscriptions, and Population Growth respectively. It is clear from the study that the result of the analysis depends on S. However, the number of important criteria does not differ much when the value of S is increased or decreased.

If the system of levels of criteria is refined, that equality of the classification can be improved. In the second analysis six levels were used at certain criteria instead of three levels (see Table 2), and criterion 19 (Surface area) was employed in the analysis as one of the important criteria. Dividing criteria into more levels eliminate the gap that may occur between countries and declare the influence of criteria in classification of countries. Subsequently, we have run the model after those changes (S=0) and new result have shown that there was no misclassification and only two criteria 13, and 6 are important. Once more we have run the model with S=0.01, there was no misclassification but the number of important criteria were increased. As results, the important criteria are criteria 4, 7, 11, 13, 15 and 19, which are Fertility rate, GNI per capita, Gross capital formation, Inflation, Mobile cellular subscriptions, Surface area respectively. In addition, penalty and reward values are as follows:

$$\begin{array}{ll} M_{4,2} = 0.02 & W_{15,1} = 0.51 \\ M_{7,1} = 0.02 & W_{19,1} = 0.49 \\ M_{11,2} = 0.48 & \\ M_{13,2} = 0.48 & \end{array}$$

The above result will be used for verification of mathematical model in the test set. Subsequently, when results in the previous steps prove that there is no misclassification, we start to estimate the gap that may exist between classes. For that reason, the following formula was utilized to find out the utmost gap between classes by maximizing S as objective function respect to the predefined constraints.

The result shows that the value of S is 0.25 (the gap between two classes) and important criteria are 6, 7, 13, 14, 15, and 17 in this analysis.

The numerical solution of model (11)-(17) gave a separation without error but the optimal solution was degenerated, i.e $S = U_1 = U_2 = U_3 = L_2 = L_3 = L_4 = 0$. Note that the degenerated solution always exist in this model e.g. S=0, $W_{13,1} = M_{19,1} = 1$, and upper and lower bounds are one. Important criteria are criteria 19, and 13. In the next level, we deiced to find the maximal possible gap which may exist between upper and lower bounds of the model. Therefore, the mathematical model was developed as follows:

$$\max S$$

subject to

$$L(a) = \sum_{i=1}^{m} \sum_{j=1}^{r_{ai}-1} W_{ij} - \sum_{i=1}^{m} \sum_{j=r_{ai}}^{p_i-1} M_{ij} \quad \forall a \in C_k$$

$$\left.\begin{array}{l} \forall a \in C_1 : U_1 \geq L(a), \\ \forall a \in C_2 : L_2 \leq L(a), U_2 \geq L(a), \\ \forall a \in C_3 : L_3 \leq L(a), U_3 \geq L(a), \\ \forall a \in C_4 : L_4 \leq L(a) \end{array}\right\}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{p_i-1} M_i j = 1, \quad \sum_{i=1}^{m} \sum_{j=1}^{p_i-1} W_{ij} = 1$$

$$U_1 + S \leq L_2, \;\; U_2 + S \leq L_3, \;\; U_3 + S \leq L_4$$

$$\forall a, S, e_{U(a)}, e_{L(a)} \geq 0$$

$$U_1, U_2, U_3, L_2, L_3 \text{ are unrestricted}$$

Then the results are changed as follow:

$$\begin{array}{ll} S = 0.23 & L2 = 0.05 \\ L3 = -0.26 & L4 = 0.23 \\ U1 = -0.73 & U2 = -0.5 \\ M32 = 0.19 & W62 = 0.23 \\ M43 = 0.35 & W141 = 0.04 \\ M65 = 0.23 & W142 = 0.23 \\ M133 = 0.23 & W152 = 0.23 \\ & W153 = 0.04 \\ & W193 = 0.04 \\ & W196 = 0.19 \end{array}$$

5 Verification of the mathematical model and validation for test set

For validation and verification, we have decided to choose new data set called test set and apply the same classification method developed in the previous section to the set.

The list of 39 new countries (alternatives) which selected for verification of our model is shown in Table 3 with their levels and classification. Classifying criteria of each country is done by considering the same process as at th old countries classification. As result of new classification, between 39 countries only four of them were misclassified (have negative value and shown by * in the table 3) which are Chile, Jamaica, Malaysia, and Philippine. This result is an excellent evidence for acceptance of our model.

No.	Country Name	Criterion						S Value
		g_4	g_7	g_{11}	g_{13}	g_{15}	g_{19}	
45	Bahrain	2	1	2	6	6	1	0.48
46	Cyprus	1	1	2	6	6	1	0.48
47	Greece	1	1	2	6	6	1	0.48
48	Ireland	1	2	2	6	6	1	0.5
49	Israel	2	2	2	6	6	1	0.5
50	Kuwait	2	2	1	6	6	1	0.5
51	Saudi Arabia	3	1	2	6	2	1	0.5
52	Singapore	1	2	4	6	6	1	0.98
53	United Arab Emirates	2	2	3	6	6	1	0.98
54	Albania	2	1	2	6	2	1	0.01
55	Armenia	1	1	3	1	1	1	0.52
56	Azerbaijan	1	1	4	1	1	1	0.52
57	Bangladesh	2	1	2	6	1	1	0.52
58	Belarus	1	1	4	1	1	1	0.52
59	Bosnia and Herzegovina	1	1	3	6	1	1	0.04
60	Bulgaria	1	1	2	3	3	1	0.49
61	Cameroon	3	1	1	6	1	1	0.5
62	Chile*	1	1	3	6	3	1	-0.47
63	Costa Rica	2	1	2	6	1	1	0.52
64	Cte d'Ivoire	3	1	1	6	1	1	0.5
65	Cuba	1	1	1	6	1	1	0.52
66	Egypt, Arab Rep.	3	1	2	6	1	1	0.5
67	Georgia	1	1	3	1	1	1	0.52
68	Ghana	3	1	2	6	1	1	0.5
69	Iran, Islamic Rep.	2	1	5	6	1	1	0.04
70	Jamaica*	2	1	4	6	3	1	-0.47
71	Kazakhstan	1	1	3	1	1	1	0.52
72	Lebanon	2	1	3	6	1	1	0.04
73	Libya	3	1	1	6	1	1	0.5
74	Malaysia*	2	1	4	6	3	1	-0.47
75	Pakistan	3	1	1	6	1	1	0.5
76	Panama	2	1	2	6	2	1	0.01
77	Philippines*	3	1	2	6	2	1	-0.01
78	Romania	1	1	3	4	2	1	0.01
79	South Africa	2	1	1	6	3	1	0.01
80	Sudan	3	1	2	5	1	1	0.5
81	Tajikistan	3	1	2	1	1	1	0.98
82	Turkmenistan	2	1	5	1	1	1	0.52
83	Vietnam	2	1	4	6	1	1	0.04

Table 3. Test set countries with classification and number of levels

6 Logical Analysis of Data

Logical Analysis of Data (LAD) is a classification method [4]. It can be used if there are only two classes and the objects are described by the same set of attributes. It was applied to separate high income countries from the other ones. The roles of the two classes are not symmetric. LAD tries

to cover one of the two classes called positive class in a way that the elements of the opposite class called negative are not covered. It creates sets of constraints such that each set is satisfied mainly but not exclusively by the positive objects. Set of constraints is called pattern. Each constraint concerns to one attribute and claims that it must be either greater or equal or less or equal than a certain threshold called cut value. For example a set of constraints is:

$$g_{15} \geq 2, g_{13} \geq 3, g_{17} \geq 2. \tag{18}$$

It is satisfied by all high income countries of the training set but it is also satisfied by some upper middle, medium and low income countries. Thus in the case of (18) the positive class is the set of high income countries and the other countries form the negative class. A set of constraints like (18) is called pattern. The quality of a pattern is expressed by two parameters. Prevalence means that what is the percentage of the objects in the positive class which satisfy the pattern. As all countries of the positive class satisfy (18) therefore its prevalence is 100%. Homogeneity is the percentage of the elements of the positive class among all objects satisfying the pattern. In the case of (18) all elements of the positive class satisfy it, i.e. 26 in the training set, and further 8 out of 18 from the negative class, thus its homogeneity is the percentage 26 out of 34, i.e. 76.47. Generally one pattern is not enough for a perfect separation of the two classes. Then a subset of patterns must be selected such that each positive object satisfies at least one pattern, i.e. it satisfies all conditions of the pattern. However here are patterns which perfectly separate the upper middle, medium and low income countries from the high income countries. Perfect separation requires that both the prevalence and the homogeneity of the pattern are 100%. The perfectly separating patterns are described in Table 4.

No.	Constraints	Test Set	
		High-income countries	Non-high-income countries
1	$g_{15} \leq 3$, $g_4 \leq 2$, $g_1 \leq 2$	None	61, 64, 66, 68, 73, 75, 77, 80 81
2	$g_{15} \leq 3$, $g_4 \leq 2$, $g_2 = 1$	None	61, 64, 66, 68, 73, 75, 77, 80 81
3	$g_{15} \leq 3$, $g_7 = 1$, $g_4 \leq 2$	None	61, 64, 66, 68, 73, 75, 77, 80 81
4	$g_{15} \leq 3$, $g_8 = 1$, $g_4 \leq 2$	None	61, 64, 66, 68, 73, 75, 77, 80 81
5	$g_{15} \leq 3$, $g_{14} \leq 2$, $g_1 \leq 2$	None	All
6	$g_{15} \leq 3$, $g_{14} \leq 2$, $g_2 = 1$	None	All
7	$g_{15} \leq 3$, $g_{14} \leq 2$, $g_7 = 1$	None	All
8	$g_{15} \leq 3$, $g_{14} \leq 2$, $g_8 = 1$	None	All

Table 4. Patterns which separate the upper middle, medium and low income countries from the high income countries on the Test set and have only three constraints. Constraints are given in the order as they are found by LAD. The order reflects the importance of the constraints.

Notice that the 8 patterns of Table 4 are the combination of only 7 constraints. The constraints are given in the order as they are found by LAD. The order reflects the importance of the constraints. Thus the most important constraint is $g_{15} \leq 3$, i.e. The value of criteria 15(mobile cellular subscription) is below average. The constraints $g_{14} \leq 2$ and $g_4 \leq 2$ still have high importance and the other constraints are supplementary ones only. Patterns $\mathcal{P}_1 = (g_{15} \leq 3, g_4 \leq 2)$ and $\mathcal{P}_2 = (g_{15} \leq 3, g_{14} \leq 2)$ are analyzed in Table 5. Pattern $\mathcal{P}_2$ and generally patterns containing its two constraints are more robust than pattern $\mathcal{P}_1$ and the children of $\mathcal{P}_1$.

No.	Constraints	Training Set		Test Set	
		High-income countries	Non-high-income countries	High-income countries	Non-high-income countries
1	$g_{15} \leq 3$, $g_4 \leq 2$	8 (Canada)	all countries	None	54-60, 62, 63, 65, 67, 69-72, 74, 76, 78, 79, 82, 83
2	$g_{15} \leq 3$, $g_{14} \leq 2$	30 (Oman)	all countries	None	all countries

Table 5. The behavior of the patterns $\mathcal{P}_1 = (g_{15} \leq 3, g_4 \leq 2)$ and $\mathcal{P}_2 = (g_{15} \leq 3, g_{14} \leq 2)$.

If the use of four constraints is allowed then further 31 perfect patterns can be obtained. They contain the above mentioned 7 constraints and only three more: $g_9 = 1$, $g_12 \geq 2$, and $g_5 = 1$. All of the 39 perfect patterns contain the constraint g15?3 and exactly one of $g_{14} \leq 2$ and $g_4 \leq 2$. Thus these three constraints are the most important ones in separating upper middle, medium and low income countries from high income countries. If the roles of the two sets of countries are interchanged then the results are different. It is the above mentioned property of LAD that the roles of the two sets are not symmetric. Perfect pattern, i.e. a pattern having 100 percent prevalence and homogeneity, does not exist. Tables 6 and 7 describe the patterns with high prevalence and homogeneity, respectively.

No.	Constraints	Homogeneity	Training set Non-high income countries
1	$g_{15} \geq 2$, $g_{13} \geq 3$	74.86%	7, 10, 26, 31, 32, 38, 40, 41, 42
2	$g_{15} \geq 2$, $g_{17} \geq 2$	72.22%	1, 6, 10, 26, 31, 32, 38, 40, 41
3	$g_{15} \geq 2$, $g_{13} \geq 3$, $g_{17} \geq 2$	76.47%	10, 26, 31, 32, 38, 40, 41, 42
4	$g_{15} \geq 2$, $g_{13} \geq 3$, $g_{18} \geq 2$	74.29%	7, 10, 26, 31, 32, 38, 40, 41, 42

Table 6. Patterns separating the high-income countries from the other ones with prevalence 1.

No.	Constraints	Prevalence	Homogeneity	Training set	
				High-income countries	Non-high-income countries
1	$g_{15} \geq 4$	92.31%	100%	8, 30	None
2	$g_{15} \geq 3$, $g_{17} \geq 2$ $g_4 = 1$	96.15%	96.15%	30	32
3	$g_{15} \geq 3$, $4 \geq g_{17} \geq 2$	92.30%	96.00%	25, 30	32
4	$g_{15} \geq 3$, $5 \geq g_{17} \geq 2$	96.15%	96.15%	30	32

Table 7. Patterns separating the high-income countries from the other ones with high homogeneity. Further patterns with homogeneity 100% exist but all of them contains the condition g15?4 with some supplementary constraints.

The countries 10, 26, 31, 32, 38, 40, 41 are non-high income countries such that they satisfy the patterns with prevalence 100%, i.e. they are close to be high income countries in a certain sense. All of them with the exception of country 31(Paraguay) belong to the upper-middle income class. Paraguay is a lower-middle income class country. Thus, the majority of these countries have a chance to enter the high-income category. It is not possible to achieve homogeneity 100% without the condition $g_{15} \geq 4$. This fact emphasizes again the importance of the criterion g_{15}.

As it mentioned before the high important criteria (more robust) are criteria 15, 14 and 4, which are mobile cellular subscription (per 100 people), military expenditure (% of GDP), and fertility rate, total (births per woman).

7 Summary and conclusion

In this paper, we try to find the most important criteria that affect World Bank's classification of countries. In order to reach this goal, mathematical models have developed using the MHDIS method. The training set was gathered and filtered, then the mathematical models have run in LINGO 12.0 program to find significant criteria, and misclassified countries have identified. In the last step to verify and validate the models, test set gathered from the same database to verify the results and complete the study.

According to the results of MHDIS method, which has discussed in Section 3, the most important criteria are GDP growth (annual %), GNI per capita (PPP, current international dollar), Gross capital formation (% of GDP), Inflation (GDP deflator, annual %), Mobile cellular subscriptions (per 100 people), and Population growth (annual %). Subsequently, we have determined the maximal gap (S) between classes, and the model has generalized to four classes instead of two. Finally, in the last step of the application of the MHDIS method, the numerical methods which were determined on the training set were applied to a new set of countries called test set. It is the verification and validation of our model. As result of verification and validation show that the method is powerful as only four countries out of 39 were misclassified.

In the Section 6 the method Logical Analysis of Data has been applied. LAD determines set of constraints called pattern for separation of the two classes. A pattern is perfect if all objects of one side satisfy it and at the same time no object of the other side satisfies the constraints of the pattern. Several perfect patterns have been obtained to separate upper middle, medium and low-income countries from the high-income countries on the training set.Perfect pattern does not exist in the opposite direction. However there are some patterns of good quality. The significant constraints that have crucial effect on separation are based on the criteria 15 (mobile cellular subscription), 14 (military expenditure), and 4 (fertility rate). The use of criteria 4 and 14 has a mutually exclusive character. On the other hand all patterns use criterion 15. The patterns based on the pair (15, 14) are the most robust ones, i.e. mobile cellular subscription and military expenditure give a good basis to decide if a country belongs to the high income category. Another valuable result that find out in LAD analysis was the number of non-high-income countries (10, 26, 31, 32, 38, 40, 41), which have potential to become member of high-income countries. It means that those countries may possibly belong to the high-income countries, because they satiety the patterns with prevalence 100% that satisfy by high-income countries as well. Interestingly, althoug it is claimed that World Bank classsification is based on GNI per capita, the GNI per capita data of the WB database have only a marginal role in the classification. Furthermore, more important factors (criteria) are: GDP growth (annual %), GNI per capita (PPP, current international dollar), Gross capital formation (% of GDP), Inflation (GDP deflator, annual %), Mobile cellular subscriptions (per 100 people), Population growth (annual %), military expenditure (annual %), and fertility rate (total birth par woman).

References

[1] Boros E., Hammer P. L., Ibaraki T., Kogan A. Logical Analysis of Numerical Data. Mathematical Programming 79 (1997), 163-190.

[2] Dahl H., Meeraus A. Zenios SA. Some financial optimization models: Risk management. In: Zenios SA, editor. Financial optimization. Cambridge: Cambridge University press 1993.

[3] Doumpos M., Zopounidis C., 2001. Assessing financial risks using a multicriteria sorting procedure: the case of country risk assessment. International Journal of Management Science 29, 97-109.

[4] Hammer P.L., Kogan A., Lejeune M.A. Reverse-engineering country risk rating: A combinatorial Non-recursive model. Rutgers, The State University of New Jersey , Rutgers, The State University of New Jersey and George Washington University, March 2007

[5] Haque N.U., Kumar M.S., Mark N., Mathieson D. 1996. The Economic Content of Indicators of Developing Country Creditworthiness. International Monetary Fund Working Paper 43 (4), 688-724.

[6] John M. Mulvey, Danish P. Rosenbaum, Bala Shetty. Strategic financial risk management and operation research. European Journal of Operation Research 97(1997), 1-16.

[7] Zopounnidis C, Doumpos M. Multicriteria sorting method. In: Floudas CA Pardalos PM, Editors. Encyclopedia of optimization. Dordrecht: Kluwer Academic Publishers, 2000.

[8] Zopounnidis C, Doumpos M.,Multicriteria classification and sorting method: A literature review. European Journal of Operational Research 138, 229-246.