December 2006

A Guide to PCA_DEA Version 1:

(Principal Component Analysis & Data Envelopment Analysis)

Computer Program

by

Nicole Adler and Ekaterina Yazhemsky

School of Business Administration

Hebrew University of Jerusalem

Mount Scopus 91905 Israel

e-mail: msnic@huji.ac.il

1. Introduction

PCA_DEA is a computer program for Windows (based on MATLAB code) which computes Data Envelopment Analysis (DEA) relative efficiency measures. It may be useful if there are a relatively large number of variables with respect to observations and discrimination between efficient and inefficient decision making units (DMU's) is problematic. Information about Data Envelopment Analysis can be found on the website http://deazone.com. The explanation about combining DEA and PCA methodologies can be find in Adler and Golany (2001, 2002, 2007) and Adler and Yazhemsky (2007). If you have questions which are not answered in the following paragraphs or if you have suggestions for further developments, send an email to msnic@huji.ac.il

2. Preparing the data

PCA_DEA accepts data in MS Excel format. The size of your analysis is limited by the memory of your PC; there is theoretically no limitation to the number of observations (DMU's) or inputs and outputs in PCA_DEA. The dataset should be collected in one worksheet. The name of the worksheet must be "data". The first column contains the DMU names (the first character is an alpha and all subsequent characters are alphanumeric). The first line contains the input/output names. The variables should be ordered as follows:

1. inputs to transform into principal components,
2. inputs to incorporate in original units,
3. outputs to transform into principal components,
4. outputs to incorporate in original units.

It is necessary to close the Excel file before running PCA_DEA. You should save the changes, close the file and then load the data in PCA_DEA. Note that PCA_DEA automatically removes any rows containing missing values from the calculations.

3. Starting PCA_DEA

For MATLAB users:

Make sure that files dialog1.m, gui_pca_dea.m, gui_pca_dea.fig, PCA.m, primal.m, dual.m are saved in the same folder as your data file. When you have prepared the data in Excel as described above, you can start PCA_DEA by running dialog1.m.

For standalone application users:

Make sure that PCADEA_pkg.zip package (which includes MCRInstaller.exe PCADEA.exe, PCADEA.ctf[1], PCADEA_mcr) and data files are saved in the same target directory. In order to reduce startup time of the application, add the target directory (i.e. the address where your datafile sits) as the first path in the user variables list.

> Right-click on **"**My Computer" and choose "Properties". In the box "System Properties" that opens, click the "Advanced" tab to obtain the dialog box. Next, click the button "Environment Variables". "Environment Variables" lists two kinds of variable - those that apply only to the current user and those that apply to the whole system. To create a new path, use the "New" button. Be sure to remember to separate directory names with a semicolon.

Run MCRInstaller.exe[2] **once** in the folder where you have collected the data and saved the program.

> The MCRInstaller opens a command window and begins preparation for installation. When the MCR Installer wizard appears, click 'Next' to begin the installation. Click 'Next' to continue. In the Select Installation Folder dialog box, specify where you want to install the MCR and whether you want to install the MCR for just yourself or others. Confirm your selection by clicking 'Next'. The MCRInstaller automatically:
>
> - Copies the necessary files to the target directory you specified.
> - Registers the components as needed.
> - Updates the system path to point to the MCR binary directory, which is <target_directory>/<version>/runtime/bin/win32. [3]

When the installation completes, click 'Close' on the Installation Completed dialog box to exit. This procedure is only required once, afterwards, simply run PCA_DEA normally. When you have prepared the data in Excel as described above, you can start PCA_DEA by running PCADEA.exe.

---

[1] CTF archives contain M-files that need to be extracted from the archive before they can be executed. The CTF archive and PCADEA_mcr directory automatically expand the first time you run the MATLAB Compiler-based component.

[2] The MCR is a version-specific file (Version 4.5 2006b MATLAB compiler).

[3] On Windows XP, this directory is automatically added to your path.

4. Running a DEA model

The first dialog box enables the user to choose the data file. The second dialog window displays all options available in the current version of PCA_DEA. After selection of the type of efficiency estimator, model orientation, returns-to-scale assumption, the total number of inputs (in categories 1 and 2) and outputs (3 and 4), the number of inputs to transform into principal components (category 1) and the number of outputs to transform into principal components (category 3) and the percent of information to retain in the model[4], the program is activated by clicking on button "RUN". [5]

Choosing an efficiency estimator

Additive (Charnes et al. (1985))

Pareto-Koopmans (mixed) efficiency: A DMU is fully efficient if and only if it is not possible to improve any input (or output) without worsening one or more of its other inputs (or outputs). An observation is rated as relatively efficient if, and only if, there are no output shortfalls or resource wastage at the optimal solution. Due to its units invariance properties, a normalized, weighted, additive DEA (Lovell and Pastor (1995)) was used in place of the simple additive.

Radial (CCR (Charnes et al. (1978)), BCC (Banker et al. (1984))

Debreu-Farrell (weak) efficiency ignores the presence of non-zero slacks. Radial (technical) inefficiency means that all inputs can be simultaneously reduced (by θ) without altering the proportions in which they are utilized.

Returns to Scale

Returns to scale refers to increasing or decreasing efficiency based on size.

- constant (CRS)
- variable (VRS)

Constant Returns to Scale means that the producers are able to linearly scale the inputs and outputs without increasing or decreasing efficiency.

---

[4] The simulation study suggests that the optimal index for the CRS (VRS) case ought to be equal to 80 (76) %. In other words, the data may be reduced to a few uncorrelated principal components that describe at least 80 (76) % of the variance of the original data. The higher the %, the greater the number of PCs retained.

[5] Each variable is divided by the corresponding standard deviation, then the correlation matrix of standardized variables and PCs are calculated. According to the level of information to retain, the program automatically chooses the number of PCs. Finally PCA_DEA linear program is used to derive efficiency scores.

Orientation

An input oriented (IN) measure quantifies the input reduction which is necessary to become efficient holding the outputs constant. Symmetrically, an output oriented (OUT) measure quantifies the necessary output expansion holding the inputs constant. A non-oriented (NON) measure quantifies necessary improvements when both inputs and outputs can be improved simultaneously.

Summary of the basic DEA models

Model	CCR_IN	CCR_OUT	BCC_IN	BCC_OUT	Weighted additive_ CRS	Weighted additive_ VRS
Sign X Y	Semi-p[6] Semi-p	Semi-p Semi-p	Semi-p Free	Free Semi-p	Semi-p Semi-p	Free Free
Translation invariance[7] X Y	No No	No No	No Yes	Yes No	No No	Yes Yes
Units invariance[8]	No	No	Yes	Yes	Yes	Yes
Efficiency	Technical	Technical	Technical	Technical	Mixed	Mixed
Score	(0,1]	[1,∞)	(0,1]	[1,∞)	[0,∞)	[0,∞)

5. Results

Once the computations are completed, the message 'Optimization terminated successfully' will be received and PCA_DEA will display the results in two new worksheets in the original data file. The name of the first worksheet specifies which model was computed, e.g. radial_OUT_VRS_100_2_3 contains the results of a DEA based on the data stored in the worksheet 'data' with radial estimator, variable returns to scale (BCC), output orientation and the PCs (2 input-PCs, 3 output-PCs) explaining 100% of the correlation in the original input and output matrices ('full information'). This worksheet contains:

[6] A semi-positive matrix has each of its columns semi-positive. A semi-positive column vector has all its elements nonnegative and at least one of its elements strictly positive.
[7] An efficiency measure is independent of the linear translation of the input and output variables.
[8] An efficiency measure is independent of the units in which the input and output variables are measured.

- DMU names;
- benchmarks arranged by corresponding intensities in descending order;
- efficiency score as defined above;
- weights (shadow prices) on inputs/outputs (W_varname);
- slacks (S_varname);
- corresponding intensities of reference DMU's (l_DMUname).

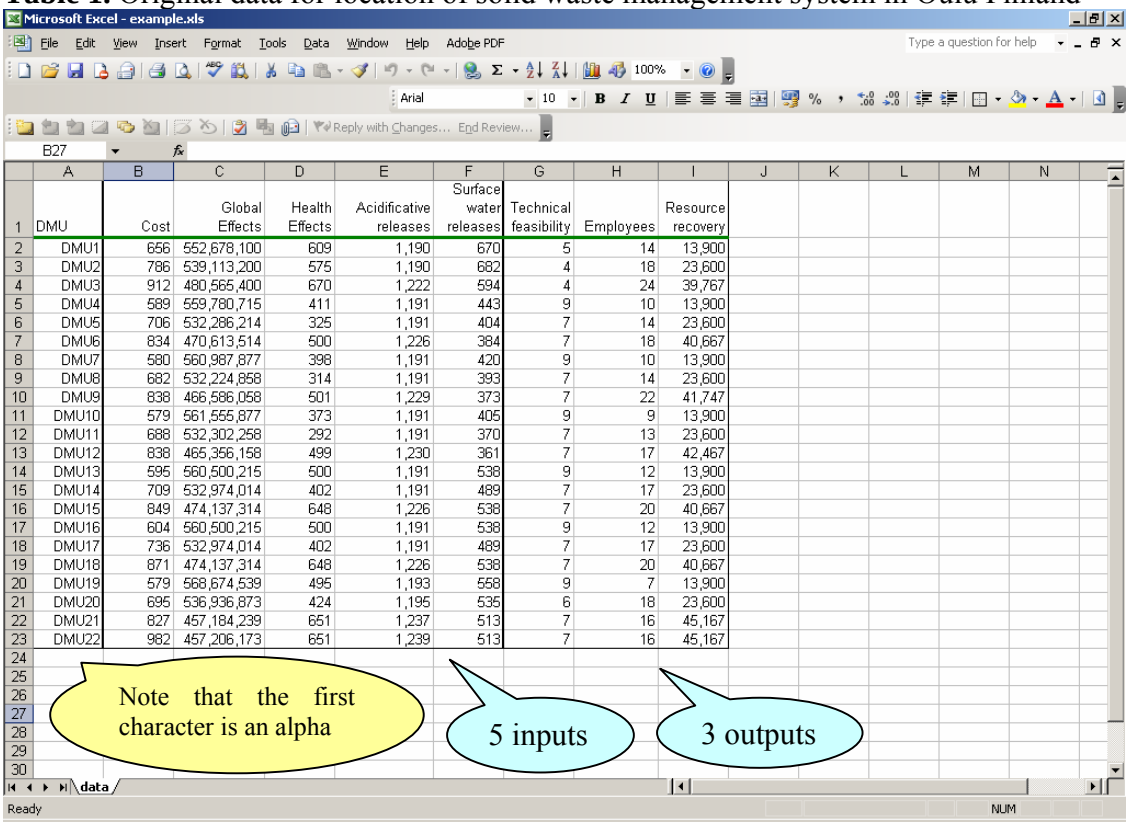
The worksheet 'PCA' describes the percentage of variance of the data explained by the input/output PCs in descending order (the first row corresponds to the first principal component). If PCA_DEA is run again on the same data file for a modified model, a new worksheet will be added to the file while the worksheet 'PCA' will be unchanged. Numbers are presented with 6 digits after the decimal point.

6. Example

In order to illustrate the PCA_DEA application, we will demonstrate the results of a dataset existing in the literature. The numerical illustration of PCA_DEA methodology was presented in Adler and Golany (2006) in which 22 solid waste management treatment systems in the Oulu region of Finland were compared over 5 inputs and 3 outputs. The original data is presented in worksheet 'data' in Excel file example.xls (Table 1). Table 2 shows the first dialog box requiring the user to retrieve a data file. The second dialog window (Table 3) displays all options available to the PCA_DEA model. The first calculations were executed for a radial input oriented constant returns-to-scale standard DEA as shown in Table 3. Worksheet 'PCA' includes the results of the principal component analysis of inputs and outputs (Table 4). As shown in Table 5 it was decided to transform all 5 inputs and 3 outputs into principal components and retain in the analysis at least 95 % of information. As a result 2 PCs on each side were included in the analysis. Table 6 presents 'radial_IN_CRS_95_2_2' worksheet. The first column includes original DMU names. In the second column benchmarks were chosen from the subset considered radially efficient (corresponding intensities are nonzero). Only 2 DMU's remain relatively efficient (the score in column three is equal to 1), namely DMU's 9 and 12. Both of them are Pareto-Koopmans efficient too (without positive slacks in L-S columns). Weights (shadow prices) on inputs/outputs are presented in

columns D-K. Note that weights on two variables (Global Effects and Employees) are very small or zero, therefore additional restrictions in LP's may be useful (assurance region). The second version of PCA_DEA will include this option. Moreover we will split the option '% of information retained' into input and output side.

Table 1. Original data for location of solid waste management system in Oulu Finland



	DMU	Cost	Global Effects	Health Effects	Acidificative releases	Surface water releases	Technical feasibility	Employees	Resource recovery
2	DMU1	656	552,678,100	609	1,190	670	5	14	13,900
3	DMU2	786	539,113,200	575	1,190	682	4	18	23,600
4	DMU3	912	480,565,400	670	1,222	594	4	24	39,767
5	DMU4	589	559,780,715	411	1,191	443	9	10	13,900
6	DMU5	706	532,286,214	325	1,191	404	7	14	23,600
7	DMU6	834	470,613,514	500	1,226	384	7	18	40,667
8	DMU7	580	560,987,877	398	1,191	420	9	10	13,900
9	DMU8	682	532,224,858	314	1,191	393	7	14	23,600
10	DMU9	838	466,586,058	501	1,229	373	7	22	41,747
11	DMU10	579	561,555,877	373	1,191	405	9	9	13,900
12	DMU11	688	532,302,258	292	1,191	370	7	13	23,600
13	DMU12	838	465,356,158	499	1,230	361	7	17	42,467
14	DMU13	595	560,500,215	500	1,191	538	9	12	13,900
15	DMU14	709	532,974,014	402	1,191	489	7	17	23,600
16	DMU15	849	474,137,314	648	1,226	538	7	20	40,667
17	DMU16	604	560,500,215	500	1,191	538	9	12	13,900
18	DMU17	736	532,974,014	402	1,191	489	7	17	23,600
19	DMU18	871	474,137,314	648	1,226	538	7	20	40,667
20	DMU19	579	568,674,539	495	1,193	558	9	7	13,900
21	DMU20	695	536,936,873	424	1,195	535	6	18	23,600
22	DMU21	827	457,184,239	651	1,237	513	7	16	45,167
23	DMU22	982	457,206,173	651	1,239	513	7	16	45,167

Table 2. Loading a data set

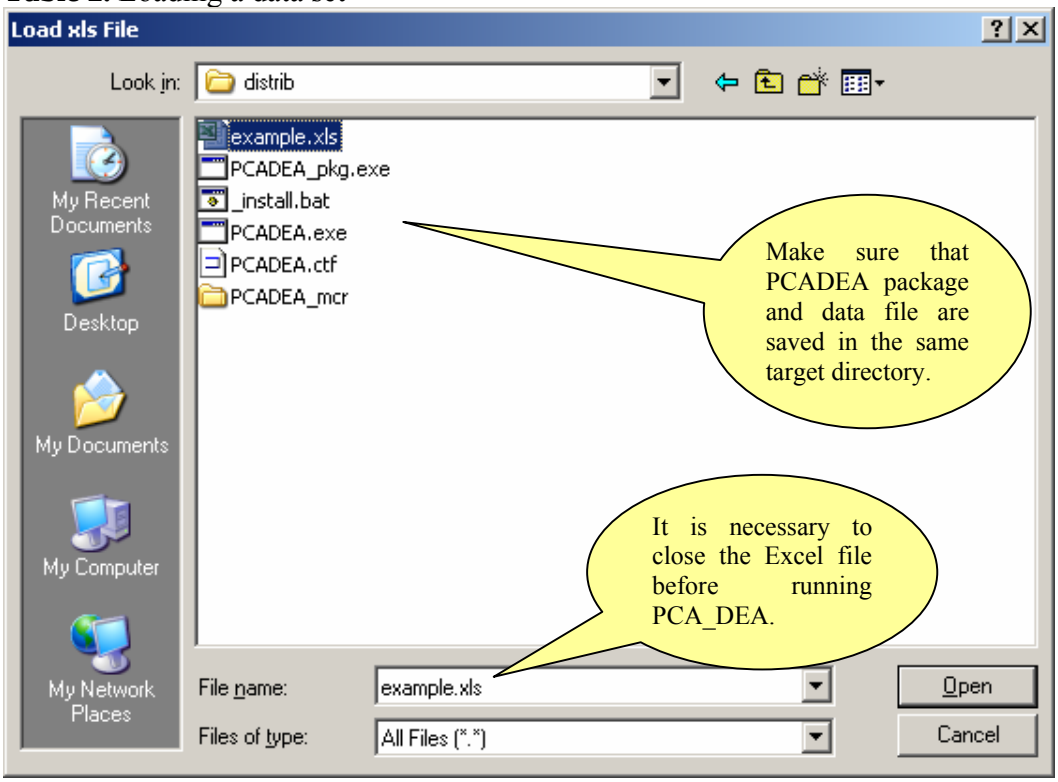


Table 3. Choosing a standard DEA

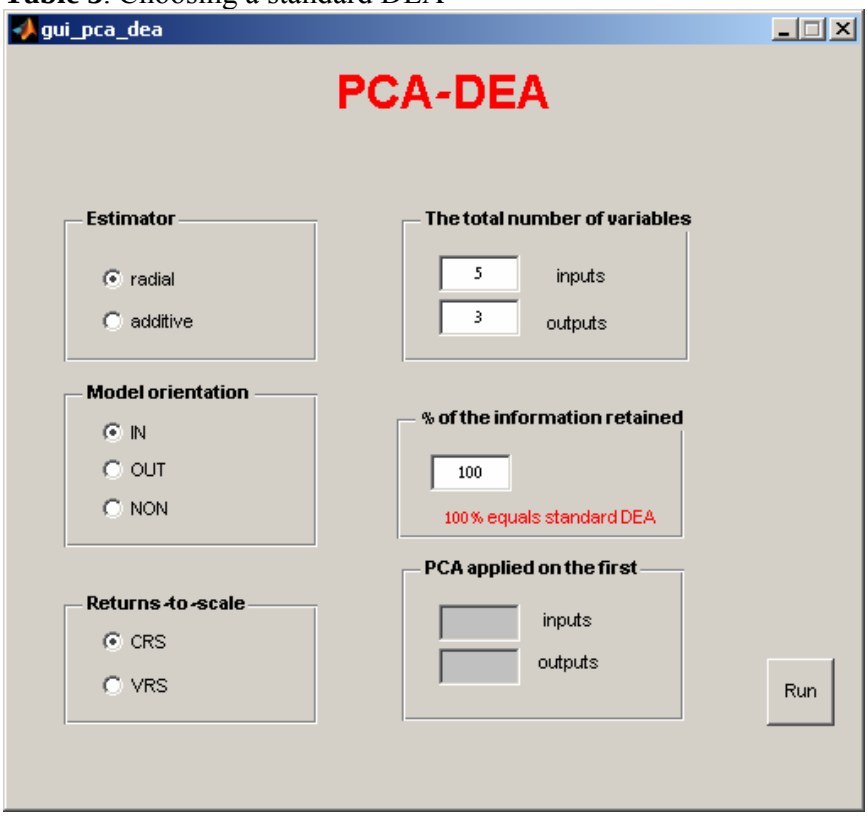


Table 4. Percentage of variance of the data explained by input/output PCs.

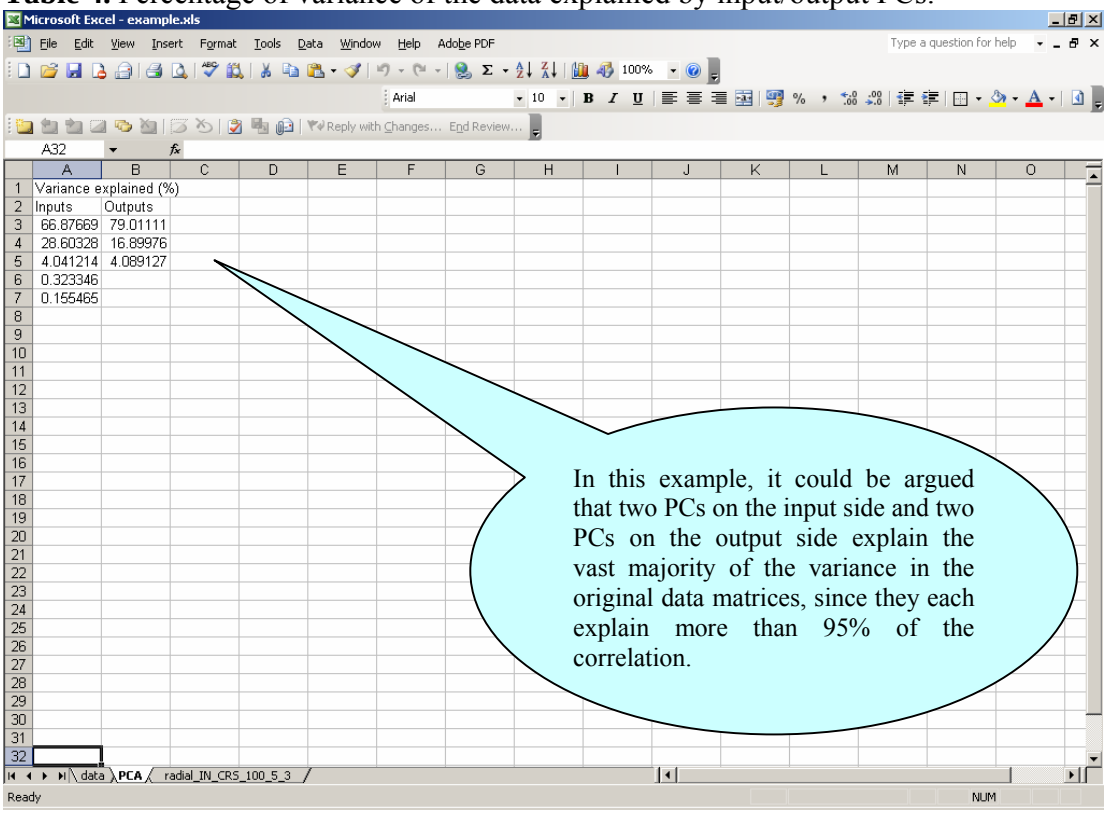


Table 5. Choosing a PCA_DEA model

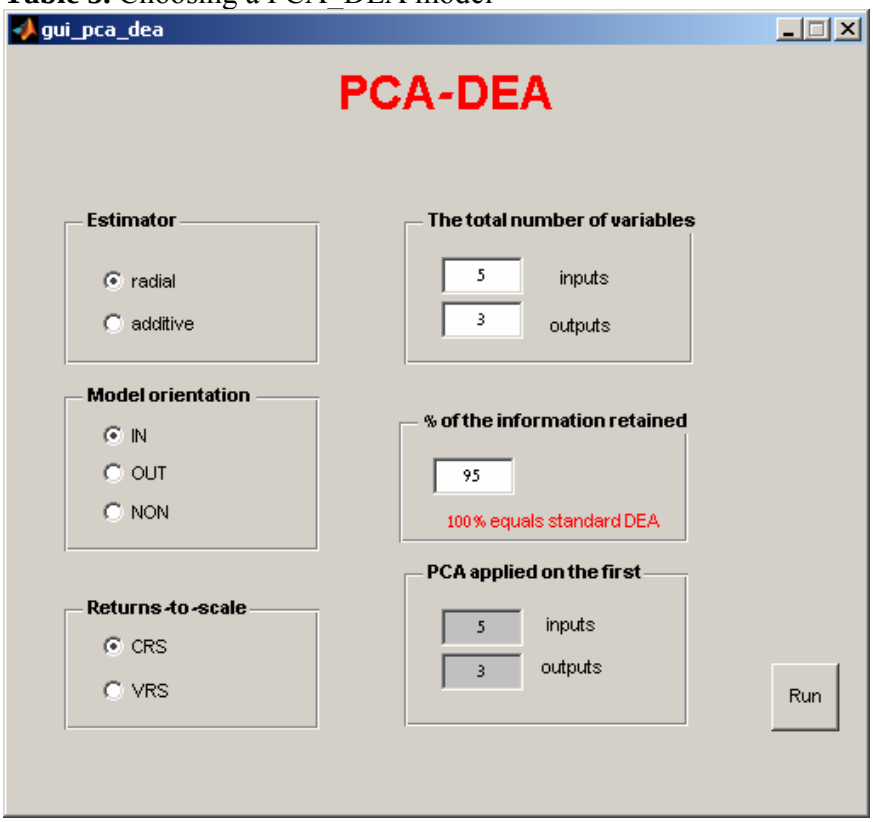


Table 6. PCA_DEA results

Microsoft Excel - example.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	DMU	Benchmarks	Score	W_Cost	W_Global	W_Health	W_Acidific	W_Surface	W_Technic	W_Employ	W_Resour	S_Cost	S_Global	S_Health
2	DMU1	DMU12	0.445421	0.011392	0	0.045961	0.004695	0.057606	0.097796	0	0.1077	0	9.40795	0
3	DMU2	DMU9	0.540273	0.011319	0	0.045664	0.004665	0.057233	0	0.069653	0.129399	0	10.59389	0
4	DMU3	DMU9	0.8069	0.011297	0	0.045576	0.004656	0.057123	0	0.069519	0.129149	0	12.17278	0
5	DMU4	DMU12	0.903834	0.014636	0	0.059048	0.006032	0.074008	0.125641	0	0.138364	0	5.04338	0
6	DMU5	DMU12	0.901421	0.015556	0	0.062758	0.006411	0.078659	0.133536	0	0.147059	0	0.799029	0
7	DMU6	DMU12, DMU9	0.966581	0.014104	0	0.056901	0.005813	0.071317	0.101052	0.01564	0.14034	0	1.249212	0
8	DMU7	DMU12	0.927645	0.015022	0	0.060603	0.006191	0.075958	0.128951	0	0.142009	0	3.888412	0
9	DMU8	DMU12	0.918009	0.015842	0	0.063913	0.006529	0.080106	0.135993	0	0.149765	0	0.09492	0
10	DMU9	DMU9	1	0.013497	0.008572	0.081649	0.000453	0.110109	0.016102	0.076026	0.15897	0	0	0
11	DMU10	DMU12	0.951365	0.015406	0	0.062153	0.006349	0.0779	0.132248	0	0.145641	0	2.740586	0
12	DMU11	DMU12	0.962957	0.015832	0.010946	0.098601	0	0.133508	0.142652	0	0.157098	0	0	0
13	DMU12	DMU12	1	0.013709	0.008756	0.083088	0.000431	0.11208	0.099686	0.018275	0.143732	0	0	0
14	DMU13	DMU12	0.807238	0.013072	0	0.052737	0.005388	0.066099	0.112213	0	0.123577	0	9.946299	0
15	DMU14	DMU12	0.810548	0.013988	0	0.056431	0.005765	0.070729	0.120074	0	0.132234	0	5.558237	0
16	DMU15	DMU9, DMU12	0.818806	0.011859	0	0.047844	0.004888	0.059965	0.084967	0.01315	0.118002	0	9.826343	0
17	DMU16	DMU12	0.806458	0.013059	0	0.052686	0.005382	0.066035	0.112105	0	0.123457	0	9.932412	0
18	DMU17	DMU12	0.80804	0.013944	0	0.056257	0.005747	0.07051	0.119702	0	0.131825	0	5.528173	0
19	DMU18	DMU9, DMU12	0.817054	0.011834	0	0.047741	0.004877	0.059837	0.084786	0.013122	0.117749	0	9.793901	0
20	DMU19	DMU12	0.798639	0.012933	0	0.052175	0.00533	0.065395	0.111018	0	0.12226	0	10.6305	0
21	DMU20	DMU12	0.700026	0.013389	0	0.054016	0.005518	0.067702	0.114935	0	0.126575	0	6.707362	0
22	DMU21	DMU12	0.898225	0.012029	0	0.04853	0.004958	0.060826	0.103261	0	0.113718	0	9.495607	0
23	DMU22	DMU12	0.884218	0.011841	0	0.047773	0.00488	0.059877	0.101651	0	0.111945	0	9.209194	0

Note that all the original data was normalized by dividing through by the standard deviation, otherwise the LPs may be infeasible due to the substantial differences in

7. Disclaimer

The author of the program described here accepts no responsibility for damages resulting from the use of this software and makes no warranty, either express or implied, including, but not limited to, any implied warranty of fitness for a particular purpose. The software is provided as it is, and you, its user, assume all risks when using it.

8. References

Adler, N., B. Golany. 2001. Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to Western Europe. European Journal of Operational Research 132 18-31.

Adler, N., B. Golany. 2002. Including Principal Component Weights to improve discrimination in Data Envelopment Analysis. Journal of the Operational Research Society 53 985-991.

Adler N., and Golany B. 2007. Data reduction through principal component analysis (DEA-PCA). Cook W. and Zhu J. (eds): Modeling Problem Structure and data varieties using data envelopment analysis: A Problem-Solving Handbook, Springer: New York. Unedited Version.

Adler N., E. Yazhemsky. 2007. Improving discrimination in Data Envelopment Analysis: PCA-DEA versus Variable Reduction. Which method at what cost? Working Paper (Submitted).

Charnes, A., W.W. Cooper and E. Rhodes. 1978. Measuring the efficiency of decision making units. European Journal of Operational Research 2 429-444.

Charnes, A., W.W. Cooper, B. Golany, L. Seiford and J. Stutz. 1985. Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. Journal of Econometrics 30 91-107.

Lovell, C.A.K. and J.T. Pastor. 1995. Units invariant and translation invariant DEA models. Operational Research Letters 18 147-151.