Replication Programs for “DEA Forest for Feature Importance: an Efficiency Analysis of Power Grid Firms”

By Tianhao Yi, Lisha Li, Jiaxuan Zhang and Zhiyong Li

1. Monte Carlo simulation

1.1. Main running files

- R\Simulation_DEA.R
- R\Simulation_SEDEA.R
- R\Simulation_NRDEA.R
- R\Simulation_BDEA.R
- R\Simulation_DEAF.R

Each file corresponds to a Monte Carlo simulation of a model. Before running, it is necessary to manually set the variables “Num_X” (number of inputs: 4, 6, 8, 10, 30), “rho” (data correlation in DGP: 0.2, 0.5, 0.85) and the Production Function (Search for the corresponding production function name in the code and replace it throughout the entire text, for example, search for “ReducedForm” and replace it with “CobbDouglas”). Other parameters do not need to be changed.

The experimental results generated by R language have all been packaged and placed in “Matlab/Result”, and divided into three folders “CobbDouglas”, “ReducedForm”, and “Translog” according to different production functions.

- Matlab\Table_result. m

To obtain **Tables 1-5** and **Appendix A** in the paper, use “Table_result. m” in Matlab and adjust the parameters “Fun” (Production Function Name), “Model” (Model Name), and “rho” (data correlation in DGP). The relevant table values are in the parameter “SP_mean”.

1.2. Production Function for DGP in Monte Carlo Simulation

- R\Fun_CobbDouglas.R
- R\Fun_ReducedForm.R
- R\Fun_Translog.R

Each function corresponds to a production function, as detailed in **Equation 9** in the paper.

2. Empirical

2.1. Transmission and distribution data

Name	Sources
Capital increment	China Electricity Industry Statistical Compilation
Increased network length	China Electricity Industry Statistical Compilation
Increased transformer capacity	China Electricity Industry Statistical Compilation
Network length	China Electricity Industry Statistical Compilation
Transformer capacity	China Electricity Industry Statistical Compilation
Number of transformers	China Electricity Industry Statistical Compilation
GDP	CSMAR Database (China Stock Market & Accounting Research Database)
Line loss	China Electricity Industry Statistical Compilation
Outage duration per customer	China Electricity Industry Statistical Compilation
Thermal generation	China Electricity Industry Statistical Compilation
Hydro power generation	China Electricity Industry Statistical Compilation
Wind power generation	China Electricity Industry Statistical Compilation
Photovoltaic power generation	China Electricity Industry Statistical Compilation
Precipitation	China Meteorological Yearbook
Hail damage area	China Meteorological Yearbook
Floods, landslides and mudslides damage area	China Meteorological Yearbook
Direct economic losses from natural disasters	China Meteorological Yearbook
Number of customers	China Electricity Industry Statistical Compilation
Number of household customers	China Electricity Industry Statistical Compilation
Number of non-household customers	China Electricity Industry Statistical Compilation
Electricity delivered	China Electricity Industry Statistical Compilation
Primary sector consumption	China Electricity Industry Statistical Compilation
Secondary sector consumption	China Electricity Industry Statistical Compilation
Tertiary sector consumption	China Electricity Industry Statistical Compilation
Household sector consumption	China Electricity Industry Statistical Compilation
Inter-provincial electricity transfer	China Electricity Industry Statistical Compilation

The data of China's power transmission and distribution scenarios is obtained by organizing the China Electricity Industry Statistical Compilation and China Metrological Yearbook, while GDP data is obtained from the CSMAR Database. The average value from 2016 to 2021 is shown in the file "Data\Data_all.xlsx". Meanwhile, the file "Data\Data.csv" is a simplified version used for subsequent efficiency evaluations. See data description details in paper **Section 4.1**.

2.2. Calculate the T&D efficiency of each province

- R\One_DEA.R
- R\One_SEDEA.R
- R\One_NRDEA.R
- R\One_BDEA.R
- R\One_DEAF.R

Use the above program to obtain **Tables 8 &9**. Each file represents an empirical program related to a model. **It is worth noting that** because DEAF and BDEA require multiple iterations to take the average result internally, even if the random seed is not fixed, the relevant results will not change much. Specifically, the efficiency rankings between provinces remain largely consistent.

2.3. Calculation file of feature importance in empirical analysis

- R\Importance_DEAF_new.R
- R\Importance_DEAF_old.R

The results shown in **Table 10** of the paper are generated by the two programs mentioned above, with the “Importance_DEAF_new.R” used to generate the results of “SP_new”.

2.4. Calculating the correlation coefficient between data

- Matlab\Data_Corr.m

Use this program in Matlab to obtain **Table 11** in the paper.

3. Explanation of other documents

- R\DEAF.R & R\DEAF_CCR.R

VRS version of DEAF (“DEAF.R”) and CRS version of DEAF (“DEAF_CCR.R”). Both are input-oriented.

- R\BDEA.R & R\BDEA_CCR.R

VRS version of BDEA (“BDEA.R”) and CRS version of BDEA (“BDEA_CCR.R”). Both are input-oriented.

- R\BCC.R & R\BCC_XX.R p

Used to calculate BCC models in BDEA and DEAF. Both are based on the VRS assumption.

- R\CCR.R & R\CCR_XX.R

Used to calculate CCR models in BDEA and DEAF. Both are based on the CRS assumption.

- R\GenerateData.R

DGP file for Monte Carlo simulation.

- R\Disturb.R

Used to artificially add perturbations to generate observation data based on raw data.

- R\LCV.R

The Likehood cross validation function is used to automatically calculate the most suitable bandwidth h in Bootstrap DEA & DEA Forest.