Saint Petersburg State University
Department of mathematical game theory and statistical decisions

Master's dissertation

Vadim Glukhov

Exploratory data analysis and university performance assessment using DEA and SFA modeling

Specialization 01.04.02
Applied Mathematics and Computer Science
Master's Program Game Theory and Operations Research

Scientific supervisor:
professor Parilina Elena

Reviewer:
associate professor, Samara University Kuznetsova Olga

Saint-Petersburg
2020

Contents

Introduction

Efficiency measurement of company's activity plays a vital role in its further development. Managers can assess information how certain department works and compare it with other departments and branches of the company. This information is crucial, when it is necessary to allocate money and resources inside the company, open new branch or close the existing one.

It is remarkable that not only commercial organizations try to improve their efficiency. Many nonprofit organizations such as public universities, charity funds, society institutions and their activity can be also estimated by numerical approaches.

World university rankings are the most common way to compare universities in different countries across all continents. They are used for about 30 years since they were first developed and they become more and more important guideline for numerous universities. Although these rankings are comprehensive and trustworthy, there are many issues associated with them: small amount of attributes, controversial attributes, different weights of attributes. Also, there rankings do not use solid mathematical model, but just a few formulas to aggregate attributes. The purpose of the research is to provide a mathematically consistent analysis to compare universities using well-known intelligible benchmarking approaches such as Data Envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA).

Problem Statement

The purpose of the research is to develop university rankings for Russian universities presented in all world university rankings using modern benchmarking approaches. To reach the goal, the following tasks were formulated.

1. Overview relevant scientific publications.
2. Provide an analysis of existing approaches of efficiency measurement.
3. Gather data of world university rankings and compare them using statistical tests.
4. Gather data of Russian universities, provide exploratory data analysis.
5. Choose inputs and outputs of universities and reduce their dimension with principal component analysis.
6. Using data envelopment analysis and stochastic frontier analysis, develop university rankings.
7. Compare developed rankings between each other and with other world university rankings.

Literature review

Decision-Making Unit (DMU) is an entity which works and has some efficiency that can be estimated related to other entities. This definition is widely used in publications and, in general, aggregates the definitions of firms, branches, departments and other entities.

Efficiency measurement theory is actively used in its modern way since 1957, when M.J. Farrell published his famous article ”The Measurement of Productive Efficiency” [11], where he introduced a new approach to efficiency measurement. The Farrell output efficiency measures how output can be proportionally increased using the same amount of inputs. It is noteworthy that the theory can be used for any type of DMUs: from government and charity funds to commercial enterprises. Universities' performance can also be assessed using efficiency methods in many difference ways. The most common and appropriate way is university rankings.

University rankings first were created in the second half last century, and now they are becoming more and more popular and widely used, because every university tries to attract attention of high quality professors and gifted students round the world. Some researchers [26] even compare university rankings with the Olympic Games, drawing an analogy between various aspects of these completely different competitions. Although rankings are widely used by experts, students, and mass media, there are several problems with them.

The first, and the main, problem is discrepancy between ranking systems [25]. Study [5] illustrated that the intersection of indicators in several rankings is small what means that the ranking systems evaluate different areas of education and non of them gives a comprehensive analysis of universities. Piro and Sivertsen [18] confirmed it with their research that some rankings concern more about research excellence (ARWU), whereas others (THE) concern about several different topics such as teaching, industrial collaboration and internationalization. Moreover, Pietrucha found [20] how other factors such a GDP per capita and political stability in a country influence university's rank. Therefore it becomes unclear how to decide which university to choose for further studying. Knowing that authors of [8] tried to generate new indicators which would cover all rankings and approximately combine all possible indices.

There are numerous researches about the comparison of the most known world university rankings. [6, 7, 13, 17, 22] Most of the researchers coincide in their opinion that rankings are inconsistent and, therefore, there is no single ranking that can be trusted entirely.

Bougnol and Dula considered [2] several technical pitfalls of the world university rankings such as co-linearity of the attributes, controversial in the attributes and not transparent techniques. Earlier these authors compared [3] the actual ranking report from University of Florida and the calculated one using mathematical programming approach - DEA. Their results show that although the ranks of first 15 universities are equal, there is still a weak correlation between other universities with Spearman's correlation coefficient equal to 0.55.

Often authors consider a particular attribute in the rankings to show inefficiency of its usage. French researchers pointed [19] that, in general, rankings take into account papers impact on the whole world, not a certain country. For instance, if a French language paper did a great influence within France, but not across the world, then the affiliated university would get much lower points than the similar English language paper. Knowing such problems Spanish authors proposed [21] to use national rankings as complements to world rankings "as the latter usually offer a poor representation of national university systems".

While some authors describe inefficiency of modern world rankings, other authors use methods from different areas of mathematics such a operations research and econometrics to design their own ranking for a small number of universities and compare it with the existed ones. Zhang, Qian and Zhao [27] used multi-agent approach based on Malmquiest index and SE-DEA to compare university efficiencies from different point of views: perspective of society, the productivity of scientific research and some others. They considered such factors as management, perspective of society, scientific research, technical facilities. Wang Hongli and Jia Yue went deeper into efficiency of scientific research and compare 58 subjects for 15 universities using similar apparatus as the latter authors. Chiang Kao and Hwei-Lan

Pao [14] did an analogous research on research performance in management of Taiwan universities.

It is very common for researchers to analyze the topic from the perspective of the country where they live. Iranian researchers [15] used efficiency analysis to assess Islamic Azad University's 28 branches in East Ajerbaijan. Norwegian authors considered [10] Nordic universities from the perspective of the impact of world rankings on their strategic plan. They realised that pursuit of rankings does not matter to them, although some countries, like Russia, have the governmental programs to help universities to advance in rankings. Another country with governmental program for boosting its universities higher in rankings is Ghana. Andy Brock [4] analyzed the question of the dependency between budgeting from government and university efficiency in this country.

Daniel Schwekendiek compared [24] approaches of two countries towards better performance of their universities. The author showed different approaches of Korea and Germany. While Korean universities achieved big successes in a short run due to "big-push" from the government, German universities are playing in the long run not giving the best results in the first several years after education reforms.

Manuel Salas-Velasco [23] provided Data Envelopment Analysis for Spanish universities and found the particular parameters in which universities were more or less efficient than others. Other authors [12] analyzed Spanish universities using clustering and factor analysis for a range of inputs and outputs of universities. Another research [9] described a comprehensive picture of the higher education in Spain, its drawbacks and prospects in terms of efficiency.

Funding is one of the most important features that can significantly boost university performance. This idea was elaborated [1] by Spanish researchers, who compared funding effects on the position of universities in rankings during the period of time.

Chapter 1. Benchmarking

1.1 Theory of Production

As efficiency primarily refers to firm and its resources, it is necessary to introduce some ideas from mathematical economics, such as *production function*, *production possibility curve* and isocurves: *isoquant* and *isocost*.

First of all, we need to explain two terms: *input* and *output*. Input is something we use as a resource to produce goods. For example, it could be labour, capital, raw materials, etc. Output is a quantity of something that we produce as a firm. For example, output of a grocery store is a number of sales per some period of time. Undoubtedly, there could be more than one output but for simplification and better graphical representation in the examples of this section only one output is used.

Inputs and outputs are connected by several ways. For instance, *production function* shows how many quantities of output will be produced, if we spend a certain amount of resources (inputs). There are many different kinds of production functions, but the most common and widely used are:

— Linear production function: $Q = a_0 + a_1K + a_2L$.

— Cobb-Douglas production function: $Q = AK^{\alpha}L^{\beta}$.

— Leontief production function: $Q = min(\frac{K}{a}; \frac{L}{b})$.

where a_0, a_1, a_2 - coefficients of linear model, A - total factor productivity, K - capital, L - labor, a, b - technologically determined constants, α - elasticity of capital, β - elasticity of labour.

An example of Cobb-Douglas production curve when capital K is fixed is shown in Figure 1.1. In this example $A = 1.1, K = 4, \alpha = 0.8, \beta = 0.2$.

Cobb-Douglas function will be used for further analysis as it has a reasonable form: earlier (smaller) quantities of inputs stronger affect output than later (larger).

Production possibility curve (PPC), or *technology set*, illustrates various combinations of outputs that can be produced by a constant level of inputs.

If firm A has two outputs: quantity of sales and quantity of clients then its PPC could be presented as depicted in Figure 1.2. Each combination of Q_1 and Q_2 on the curve shows the possible output while inputs are constants. Increasing the amount of Q_1 causes decreasing the amount of Q_2.

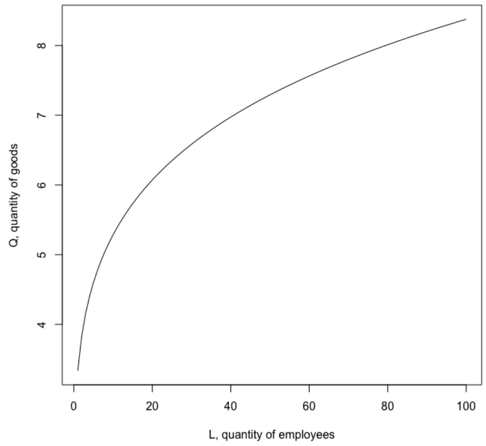

Figure 1.1: Cobb-Douglas production curve.

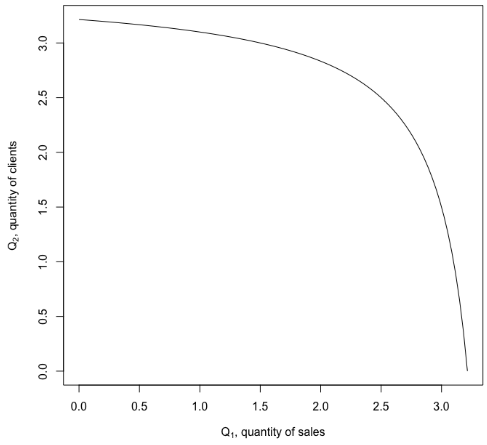

Figure 1.2: Production possibility curve.

Isoquant curve shows different combinations of inputs to produce a constant level of output. Output quantity is the same during the whole curve. As an example we can consider a firm with two inputs and one output. Isoquant of the firm is shown in Figure 1.3.

Isocost curve shows different combinations of inputs while the total amount of costs remains unchangeable. It could be easily seen that the slope of isocost is just a ratio of minus quantity of price of the first input divided by the price of the second input.

1.2 Efficiency measurement approaches

Some simple approaches of efficiency measurement were known many years. They were intuitive and appropriate to assess productivity of resources separately from each other. Many of them are widely used now when it is not necessary to provide time consuming calculations. As an example, we can recall *workforce productivity* or *capital intensity*, which are just a result of division of two measures. Unfortunately, these measures are very simple and cannot show the whole picture of the firm behaviour.

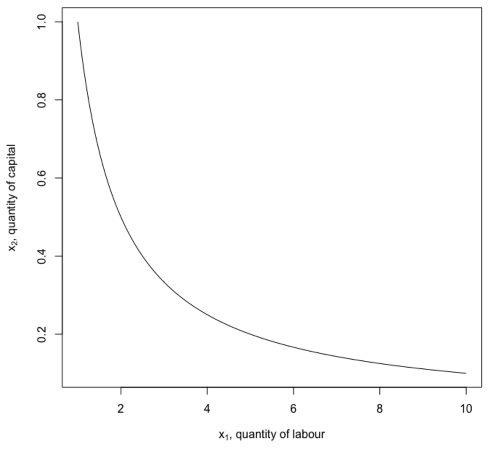

Figure 1.3: Isoquant curve.

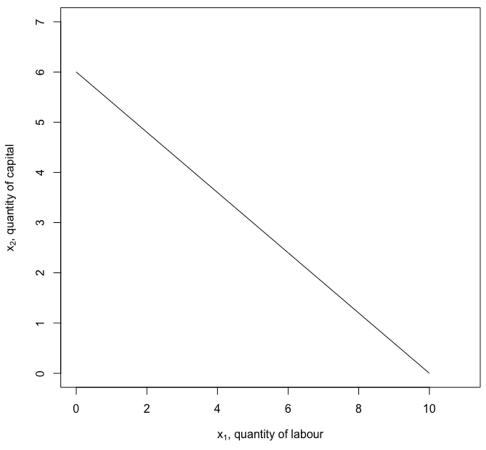

Figure 1.4: Isocost curve.

There are numerous different techniques for efficiency measurement. Most common and widely used of them are: *input-oriented* and *output-oriented* approaches. Both include such an important term as technology set defined above as a production possibility curve. *Technology set (T)* could be defined as the set of combinations of input and output such that the input can actually produce the output:

$$T = \{(x, y)|x \text{ can produce } y\}.$$

One of the most widely used approaches in efficiency measurement is **Farrell efficiency**, which was suggested by Farrell and Debreu. The main idea is whether it possible to reduce the input without change the output. In terms of multiple inputs, the idea is to proportionally reduce all inputs.

The input-based Farrell efficiency of plan (x, y) relative to a technology T is defined as

$$E = min\{E > 0|(Ex, y) \in T\}$$

and means the maximal proportional contraction of all inputs x that allows to produce y. For instance, if E = 0.92, then it is possible to save 8% of all inputs producing the same amount of outputs.

Graphically it can be presented using actual combination of inputs and the optimal combination of inputs. (Figure 1.5)

As can be seen firm's input combination is far from the optimal one

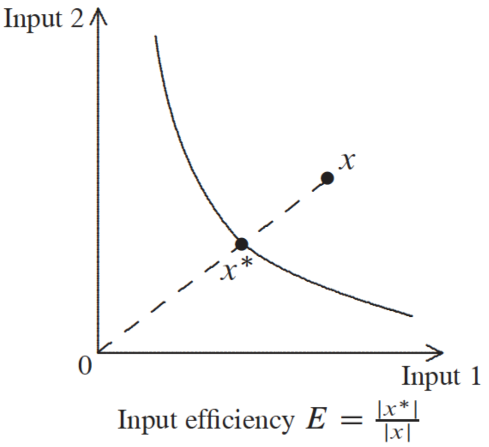

Figure 1.5: Input-oriented approach.

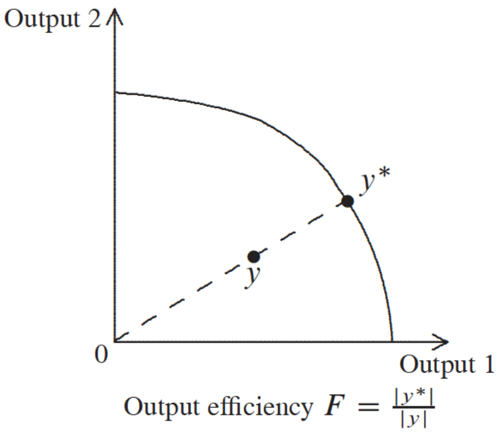

Figure 1.6: Output-oriented approach.

located on the isoquant because it is still possible to reduce inputs remaining the level of output the same. In this case, the efficiency score is calculated as $E = \frac{|x^*|}{|x|}$.

The output-based Farrell efficiency of plan (x, y) relative to a technology T is defined as

$$F = max\{F > 0|(x, Fy) \in T\}$$

and means the maximal proportional expansion of all outputs y having the same inputs x. For instance, if F = 1.13, then it is possible to increase output by 13% without increasing of inputs.

Graphically it can be presented using actual combination of outputs and the optimal combination of outputs. (Figure 1.6)

As can be seen firm's outputs combination is far from the optimal one located on the output isoquant, or transformation curve, because it is still possible to increase outputs with the same level of inputs. In this case, the efficiency score is calculated as $F = \frac{|y^*|}{|y|}$.

Another representation of Farrell efficiency is Shaphard distance functions, which are just the inverse of the Farrell ones:

$$D_i(x,y) = max\{D > 0|(\frac{x}{D}, y) \in T\} = \frac{1}{E(x,y)},$$

$$D_o(x,y) = min\{D > 0|(x, \frac{y}{D}) \in T\} = \frac{1}{F(x,y)}.$$

1.3 Data Envelopment Analysis

Data Envelopment Analysis is a non-parametric approach of efficiency measurement. It based on operations research, mathematical programming and management science. The background of the DEA approach is a production theory, and the crucial assumption is that all decision-making unites are efficient, i.e. there is no inefficiency.

Assumptions of the basic DEA models are usually about the technology T. The combinations of the following properties give different models according to return to scale:

— **Free disposability.** University can produce less with more: $(x, y) \in T, x^{'} \geqslant x, y^{'} \leqslant y \implies (x^{'}, y^{'}) \in T$.

— **Convexity.** Any weighted average of feasible production plans is feasible as well: $(x, y) \in T, (x^{'}, y^{'}) \in T, \alpha \in [0, 1] \implies \alpha(x, y) + (1 - \alpha)(x^{'}, y^{'}) \in T$.

— **γ-returns to scale.** Production can be scaled with any of a given set of factors: $(x, y) \in T, k \in \Gamma(\gamma) \implies k * (x, y) \in T$, where γ - type of return to scale.

— **Additivity, replicability.** The sum of any two feasible production plans is feasible as well: $(x, y) \in, (x^{'}, y^{'}) \in T \implies (x + x^{'}, y + y^{'}) \in T$.

The most common types of return to scale are:

1. FDH - Free disposability hull.
2. VRS - Varying return to scale.
3. DRS - Decreasing return to scale.
4. IRS - Increasing return to scale.
5. CRS - Constant return to scale.

6. FRH - Free replicability hull.

Another prerequisite for DEA model is an idea of *minimal extrapolation*. In most cases, actual technology T is not known, so it is necessary to estimate technology T^* by existing data. When we consider candidate technologies $T^{'}$ that are subsets of $\mathbb{R}^m_+ \times \mathbb{R}^n_+$ and that

1. contains data: $(x^k, y^k) \in T^{'}, k = 1, ..., K$;
2. satisfies the regulatory assumptions,

the set of such candidate technologies can be denoted as

$$\tau = \{T^{'} \subset \mathbb{R}^m_+ \times \mathbb{R}^n_+ | T^{'} \text{ satisfy (1) and (2)}\},$$

The minimal extrapolation principle states that we estimate the unknown technology T by the set

$$T^* = \bigcap_{T^{'} \in \tau} T^{'}.$$

For analysis of universities we assume that return to scale is variable as it is an aggregation of the other types (increasing, constant and decreasing). Therefore minimal extrapolation technology for variable return to scale is

$$T^*(drs) = \{(x, y) \in \mathbb{R}^m_+ \times \mathbb{R}^n_+ | \exists \lambda \in \Lambda^K(vrs) : x \geqslant \sum_{k=1}^{K} \lambda^k x^k, y \leqslant \sum_{k=1}^{K} \lambda^k x^k\},$$

where $\Lambda^K(vrs)$ is

$$\Lambda^K(vrs) = \{\lambda \in \mathbb{R}^K_+ | \sum_{k=1}^{K} \lambda^k = 1\}.$$

Combining Farrell idea of proportional improvements (decreasing inputs, increasing outputs) and minimal extrapolation principle, we can define DEA formulation. The first formulation is for input case. The goal is to measure the Farrell efficiency of firm o as the input efficiency

$$E^o = E((x^o, y^o); T^*) = min\{E \in \mathbb{R}_+ | (Ex^o, y^o) \in T^*\}.$$

Replace $T^*(\lambda)$ with the technology defined above, the general formulation is

$$\min_{E,\lambda^1,\ldots,\lambda^K} E,$$

$$s.t. Ex_i^o \geqslant \sum_{k=1}^{K} \lambda^k x_j^k, i = 1, \ldots, m,$$

$$y^o \leqslant \sum_{k=1}^{K} \lambda^k y_j^k, j = 1, \ldots, n,$$

$$\lambda \in \Lambda^K(vrs).$$

The second formulation is for output case. The goal is to measure the Farrell efficiency of firm o as the output efficiency

$$F^o = F((x^o, y^o); T^*) = max\{F \in \mathbb{R}_+ | (x^o, Fy^o) \in T^*\}.$$

Replace $T^*(\lambda)$ with the technology defined above, the general formulation is

$$\max_{F,\lambda^1,\ldots,\lambda^K} F$$

$$s.t. x_i^o \geqslant \sum_{k=1}^{K} \lambda^k x_j^k, i = 1, \ldots, m$$

$$Fy^o \leqslant \sum_{k=1}^{K} \lambda^k y_j^k, j = 1, \ldots, n$$

$$\lambda \in \Lambda^K(vrs)$$

As can be seen, it is a typical optimization problem that can be solved by linear programming for m linear inputs.

1.4 Stochastic Frontier Analysis

Stochastic Frontier Analysis is a parametric approach of efficiency measurement based on economics theory and econometrics. Unlike DEA, it proposes an idea that a firm could be inefficient. Also, there is a stochastic component which obviously exists in real life. Nevertheless it requires satisfaction of some additional assumptions, which make this approach less attractive to use.

The main notion of Stochastic Frontier Analysis is to find the approximation of production function having inputs and outputs for several universities. The task of data approximation has been solved successfully many years ago by a approach called regression analysis. Regression model can be represented as

$$y^k = f(x^k; \beta) + v^k, v^k \sim N(0, \sigma^2), k = 1, ..., K,$$

where K - number of DMUs, x^k - input matrix for k-th DMU, y^k - output vector for k-th DMU, v^k - noise for k-th DMU.

In terms of universities, this means that all universities are efficient, but the deviation for production curve is only due to some random factor. The model above is approximated by Ordinary Least Squares (OLS) method.

The *deterministic version* of regression model is

$$y^k = f(x^k; \beta) - u^k, u^k \sim H, k = 1, ..., K,$$

where v^k - inefficiency for k-th DMU, H - some probability distribution with support only on $\mathbb{R}_+$.

The approach above is closer to desired one by which we want to approximate universities data, but by nature there are always some random factors affect data.

The third approach is *Corrected Ordinary Least Squares (COLS).* It is just OLS shifted upward with the maximum error term.

$$\min_{\beta} \sum_{k=1}^{K} (y^k - f(x^k; \hat{\beta})^2,$$

$$\beta_{00} = max\{y^k - f(x^k; \hat{\beta}) | k = 1, \ldots, K\}.$$

COLS is *corrected* because it eliminates OLS problem that some DMUs are above the production curve what is not possible.

Finally, Stochastic Frontier Analysis is an approach that combine OLS and deterministic approaches. General formulation of SFA is following:

$$y^k = f(x^k; \beta) + v^k - u^k,$$

$$v^k \sim N(0, \sigma_v^2), u^k \sim N_+(0, \sigma_u^2), k = 1, ..., K.$$

As can be seen, distribution for inefficiency terms are half-normal, because inefficiencies cannot be negative. Also, there is an assumption that v^k and u^k are independent. Comparison SFA with ordinary regression model and corrected OLS is shown in Figure 1.7.

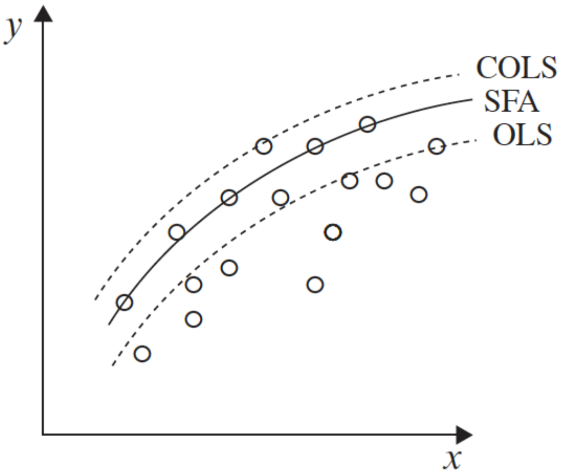

Figure 1.7: Comparison of ordinary regression model and SFA.

The task is to determine β, σ_v^2 and σ_u^2. The main problem is to separate inefficiency and noise, because we can only calculate the total error terms:

$$\epsilon = v^k - u^k = y^k - f(x^k, \hat{\beta}).$$

This problem may be solved analysing distributions of noise and inefficiency and their intercepts. Let us denote $\sigma^2 = \sigma_v^2 + \sigma_u^2$ and $\lambda = \sqrt{\frac{\sigma_u^2}{\sigma_v^2}}$, then we can estimate σ^2 and λ instead of variance of errors. Examples of some cases of new definitions are the following:

— If $\lambda = 0, \sigma^2 = 1, \sigma_v^2 = 1, \sigma_u^2 = 0$, then distributions overlap each other with mean zero and total distribution is as the overlap.

— If $\lambda = 0, \sigma^2 = 1.5, \sigma_v^2 = 1, \sigma_u^2 = 0.7$, then inefficiency distribution is skewed slightly right. To total distribution is very tall.

— If $\lambda = 0, \sigma^2 = 2, \sigma_v^2 = 1, \sigma_u^2 = 1$, then inefficiency distribution is skewed slightly right. To total distribution is tall.

— If $\lambda = 0, \sigma^2 = 6, \sigma_v^2 = 1, \sigma_u^2 = 2.2$, then inefficiency distribution is largely skewed right. To total distribution is little taller that the efficiency one and has the same shape.

All of the above cases are shown in Figure 1.8.

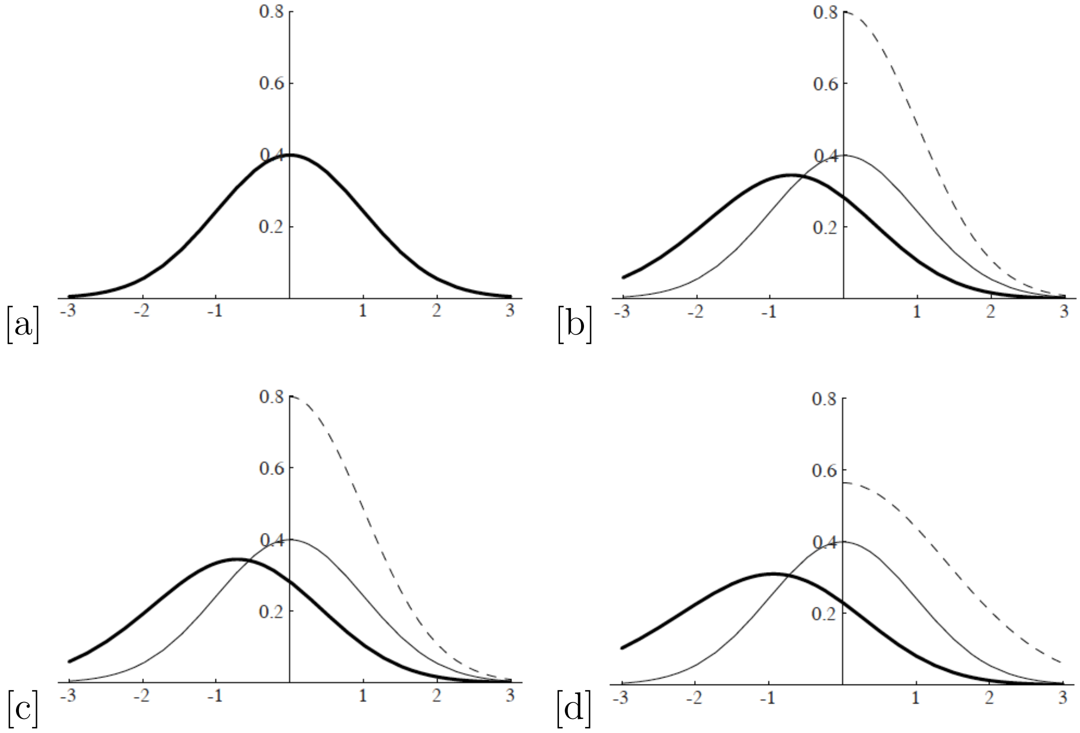

Figure 1.8: (a) $\lambda = 0, \sigma^2 = 1, \sigma_v^2 = 1, \sigma_u^2 = 0$; (b) $\lambda = 0, \sigma^2 = 1.5, \sigma_v^2 = 1, \sigma_u^2 = 0.7$; (c) $\lambda = 0, \sigma^2 = 2, \sigma_v^2 = 1, \sigma_u^2 = 1$; (d) $\lambda = 0, \sigma^2 = 6, \sigma_v^2 = 1, \sigma_u^2 = 2.2$.

σ^2 and λ can be estimated using *maximum likelihood estimation.* The log-likelihood function for the estimation is shown below.

$$l(\beta, \sigma^2, \lambda) = -\frac{1}{2}Klog(\frac{\pi}{2}) - \frac{1}{2}Klog(\sigma^2) + \sum_{k=1}^{K} log\Phi(-\frac{\varepsilon_k \lambda}{\sqrt{\sigma^2}}) - \frac{1}{2\sigma^2}\sum_{k=1}^{K}\varepsilon_k^2,$$

where ε_k - total error term of DMU k.

Log-likelihood function above cannot be computed analytically, so the only way is to use numerical optimization using software program. When σ^2 and λ are calculated, it is easy to return to variance of noise and inefficiency:

$$\lambda = \sqrt{\frac{\sigma_u^2}{\sigma_v^2}} \implies \sigma_u^2 = \lambda^2 \sigma_v^2$$

$$\sigma^2 = \sigma_v^2 + \sigma_u^2 = \sigma_v^2 + \lambda^2\sigma_v^2 = \sigma_v^2(1+\lambda^2) \implies \sigma_v^2 = \frac{1}{1+\lambda^2}\sigma^2$$

$$\sigma_u^2 = \lambda^2\sigma_v^2 = \frac{\lambda}{1+\lambda^2}\sigma^2$$

As we got a variance for half-normal distribution of inefficiency term, we still do not have firm-specific efficiency. We know the total inefficiency error term ϵ, but not the efficiency term. But it can be estimated as expected value of conditional distribution:

$$TE = E(e^{-u}|\epsilon) = \frac{\Phi(\mu_*/\sigma_* - \sigma_*)}{\Phi(\mu_*/\sigma_*)} e^{\frac{1}{2}\sigma_*^2 - \mu_*},$$

where

$$\mu_* = -\epsilon\frac{\sigma_u^2}{\sigma^2} = -\epsilon\frac{\lambda^2}{1+\lambda^2},$$

$$\sigma_* = \sqrt{\frac{\sigma_u^2\sigma_v^2}{\sigma^2}} = \frac{\lambda}{1+\lambda^2}\sigma.$$

All the mentioned above is suitable for the case of one output. Multi-output model called *estimable stochastic distance function* and has the following form:

$$log(\frac{1}{x_m}) = log(D_i(\frac{x}{x_m}, y)) + v - u,$$

$$log(y_n) = -log(D_o(x, \frac{y_n}{y})) + v - u,$$

where D_i - input distance function, D_o - output distance function, $v \sim N(0, \sigma_v^2)$, $u \sim N_+(0, \sigma_u^2)$, m - number of inputs, n - number of outputs.

Interpretation of $\frac{x}{x_m}$ is $(\frac{x_1}{x_m}, ..., \frac{x_{m-1}}{x_m})$; interpretation of $\frac{y_n}{y}$ is $(\frac{y_n}{y_1}, ..., \frac{y_n}{y_{n-1}})$.

1.5 Principal Component Analysis

Principal component analysis provides linear dimensionality reduction. The main idea of the method is to approximate data with less dimensions.

The intuition is to approximate maximum variance of the data. Although the interpretation of principle components are not clear, this approach is better in most cases than elimination of inputs.

Each principal component is associated with variance of the data: the more principal component (PC) includes variance, the better it explains the initial data. The goal is to choose less PCs with the most explained variance. In order to explain more variance, highly correlated variables should be analysed; hence, from inputs and outputs highly correlated variables were selected.

The steps to derive principal components are:

1. Data normalization.
2. Computation of covariance matrix.
3. Computation of eigenvalues and eigenvectors of the covariance matrix.
4. Dimensionality reduction (computation of projections).

Data normalization is used in order to eliminate the difference in scales of variables.

Covariance matrix shows the dependence between variables. It is computed by the following formula:

$$Cov(x_i, x_j) = E[(x_i - E(x_i))(x_j - E(x_j))] = E(x_i, x_j) - E(x_i)E(x_j).$$

As data normalized, following equations are true:

$$E(x_i) = E(x_j) = 0$$

and

$$Cov(x_i, x_j) = E(x_i, x_j).$$

Covariance for the same variable is just its variance:

$$Cov(x_i, x_i) = Var(x_i).$$

The goal of PCA is to find a vector, which maximizes the variance of data. This vector is called *projection* of the data X. Projection can be presented as $v^T X$ with variance $Var(v^T X)$. Therefore, the variance of data is

$$Var(X) = E(XX^T)$$

and

$$Var(v^T X) = E((v^T X)(v^T X)^T) = v^T E(XX^T)v.$$

Variance is maximized while $v^T E(XX^T)v$ is maximized.

Rayleigh quotient for covariance matrix is

$$R(M, \bar{X}) = \frac{\bar{x}^T M \bar{x}}{\bar{x}^T \bar{x}} = \lambda \frac{\bar{x}^T \bar{x}}{\bar{x}^T \bar{x}} = \lambda,$$

where λ - eigenvalue of the covariance matrix. Therefore, maximum variance is in the direction of the eigenvalue with the maximum value.

The first principal component then can be derived as

$$PC_1 = (v^T X)^T v^T + m,$$

where m - vector of means to counteract the normalization, v^T - eigenvalue with the maximum value.

Other principal components can be derived using other eigenvalues in decreasing order.

Chapter 2. Data Processing

2.1 Data collection

The research operates with 4 rankings from different issuers:

- Academic Ranking of World Universities (ARWU) (China, since 2003)
- QS World University Rankings (QS) (UK, since 2004)
- Times Higher Education World University Rankings (THE) (UK, since 2004)
- The Center of World University Rankings (CWUR) (UAE, 2012)

As ARWU does not provide portable version of its ranking list, it is necessary to gather data from its official website. It can be done using *web scraping*. The notion of web scraping is crawling through website using certain software to gather necessary data. In this case, web scraper was written using Python programming language. As the output program returns excel file with columns: rank, university and 6 columns for university's features. After data is taken, universities are corresponded with the ranks of rankings manually.

The data are taken for all rankings for the last available year. The sample of the data for this research includes 313 universities. All rankings have their own features so it is important to elaborate each of them in order to work with data correctly.

CWUR is the most complete and not ambiguous ranking system, where every university has its own unique rank. There are 2000 universities in this ranking system.

ARWU has 1000 universities, where exact ranks are known for the first 100 universities. Other universities divided by groups with particular ranges of ranks. For instance, there are 10 groups for ARWU: 101-150, 151-200, 201-300, 301-400, 401-500, 501-600, 601-700, 701-800, 801-900, 901-1000 ranks with the corresponding number of universities in each group.

It becomes a problem to work with such not unique defined data. For this ranking system and three other, the possible way to tackle the problem is to replace range rank for university by single rank derived from the uniform distribution with boundaries as boundaries of range rank. In this case, some ranks can be the same, so if they will be used in rank tests, it is necessary to use the correction according to ties in data.

THE has 10 groups of ranks: 201-250, 251-300, 301-350, 351-400, 401-500, 501-600, 601-800, 801-1000, 1001+. The total number of universities in this ranking is 1397.

QS has 15 groups of ranks: 501-510, 511-520, 521-530, 531-540, 541-550, 551-560, 561-570, 571-580, 581-590, 591-600, 601-650, 651-700, 701-750, 751-800, 801-1000. Thus, QS ranking is the second most elaborated, so rank replacement here will be better that in other two ones. The total number of universities in this ranking is 1002.

An example of the universities ranking table is illustrated in Table 2.1.

University	QS	THE	ARWU	CWUR
The University of Melbourne	32	38	64	41
New York University	29	39	26	30
Fudan University	109	40	132	101-150
KAIST	110	41	199	201-300
The University of Sydney	60	42	100	80
The University of New South Wales	71	43	113	94
...	...	...	...	...
Ural Federal University	1000+	364	-	701-800

Table 2.1: Subset of universities ranking table.

As can be seen, there are two types of ranges: with both boundaries and with the lower boundary. The second interesting thing is the scatter of ranks. For example, KAIST has a difference for at least 160 ranks between two most "distant" rankings. Another notice is the lack of rank for particular university in one or more rankings. These universities will be excluded from the analysis of the universities concordance.

Although the universities data are highly used in rankings, it cannot be used due to the fact that it is already processed and modified into different attributes. Moreover, there are no full databases and datasets ready to download in the Internet, so it becomes impossible to provide the analysis for all world universities. Nevertheless, some countries provide online tables and summaries about their universities, so there was a decision to provide the analysis for Russian universities which are presented in all four rankings.

2.2 Rankings consistency

The first point of the research is to prove inconsistency between rankings. The simplest metric to measure data proximity is to use *Spearman's pairwise rank correlation coefficient*:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

where $d_i = Rank(X_i) - Rank(Y_i)$ - difference between the two ranks of each observation, n - number of observations:

In case of four rankings, there will be 4x4 matrix with pairwise-correlation coefficients.

Ranking	QS	THE	ARWU	CWUR
QS	1	0.74	0.75	0.76
THE	0.74	1	0.57	0.58
ARWU	0.75	0.57	1	0.91
CWUR	0.76	0.58	0.91	1

Table 2.2: Correlation matrix between university rankings.

Heat map can be used for better interpretation of the correlation matrix as it highlights cells with common behavior: the lighter cell illustrates the higher correlation. Here and after heat maps will be used instead of correlation matrices because of better interpretation.

The interesting point is that correlation is significantly different for different part of ranks. In general, it decreases when lower rank universities are considered. Thus, universities with ranks 1-100 for QS ranking have higher correlation than the next hundred of universities. 2.2 It can be seen, when correlation matrix is computed for each of the hundred of universities.

In order to aggregate all the pairwise correlation coefficients and get only one metric, *Kendall's coefficient of concordance* can be used:

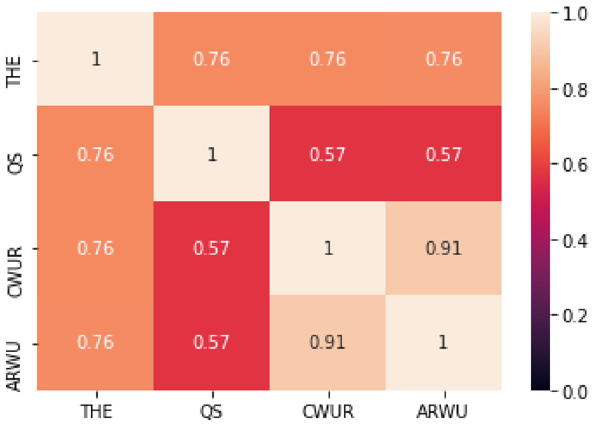

Figure 2.1: Heat map of the correlation matrix.

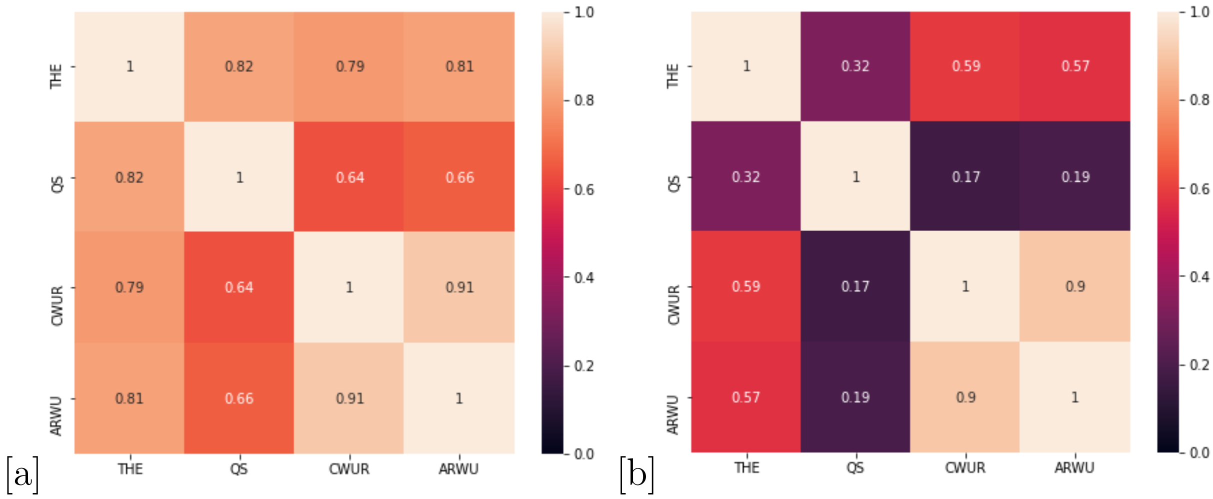

Figure 2.2: (a) Heat map of correlation matrix for the universities with ranks 1-100 in QS ranking; (b) Heat map of correlation matrix for the universities with ranks 101-200 in QS ranking.

$$W = \frac{12S}{m^2(n^3 - n)},$$

where

$$S = \sum_{i=1}^{n} (R_i - \bar{R})^2,$$

$$\bar{R} = \frac{1}{n} \sum_{i=1}^{n} R_i,$$

$$R_i = \sum_{j=1}^{m} r_{i,j},$$

where $r_{i,j}$ - rank for university i given by ranking j, m - number of rankings, n - number of universities.

with the following *Spearman's rank correlation coefficient* between all $\binom{2}{m}$ possible pairs of rankings between rankings:

$$\bar{r_s} = \frac{mW - 1}{m - 1}.$$

The results for the given rankings data are:

$$W = 0.79,$$

$$\bar{r_s} = 0.72.$$

Correlation coefficient means that in 72% of observations are correlated between each other. This number is even less than three fourth, so the conclusion is that the data is inconsistent, specially when not the first top-100 universities are considered.

2.3 Inputs and outputs

First, it is necessary to gather data of Russian universities. Fortunately, information is available on [34] website but not in a portable format. For the second time web scraping is used here. Web scraper goes through every region of Russia and universities of regions getting all the available information presented in tables. Even though data for only 11 universities is needed, web scraper gathers is for all Russian universities. It is very important and useful in the next studies when all Russian universities will be considered.

Inputs and outputs for Russian universities are takes from [34]. The desired universities are the ones which are presented in all the mentioned above rankings on the last year. There are 11 Russian universities taken for the further analysis:

- MSU - Moscow State University
- SPBU - Saint-Petersburg State University
- NSU - Novosibirsk State University

- MIPT - Moscow Institute of Physics and Technology
- HSE - Higher School of Economics
- MEPHI - National Research Nuclear University MEPhI
- ITMO - ITMO University
- TSU - Tomsk State University
- TSPU - Tomsk Polytechnic University
- KFU - Kazan Federal University
- URFU - Ural Federal University

All universities above are located in 6 regions. Although capital region in most countries is the most wealthy for most of top universities, in case of Russia it does not work.

Region	Number of universities
Moscow	2
Saint-Petersburg	2
Tomsk	2
Ekaterinburg	1
Kazan	1
Novosibirsk	1

Table 2.3: Number of universities per region.

There are several outputs for every university where 5 outputs are common for all universities and other outputs are unique for every university. Common outputs are:

- E.1 Educational activities.
- E.2 Research activities.
- E.3 International activities.
- E.4 Financial and economic activities.
- E.5 Faculty salaries.

There are numerous inputs divided by several groups similar to outputs:

- Educational activities.
- Research activities.
- International activities.
- Financial and economic activities.
- Infrastructure.
- Staff composition.

Some of the inputs are aggregated or exclude each other. *Total square of campus* includes *Total square of labs* and *Total square of dormitories*, so they cannot be used together. *The share of university income from the federal budget* and *The share of university income from non-budgetary sources* are mutually excluded because the sum of them is 100%.

The following inputs are taken for further analysis, where some of them are aggregated in order to decrease number of inputs:

- Education activities
  - 1.1. Average score of Unified State Exam (USE) of newcomer bachelor students study for money of the Russian Federation.
  - 1. Number of students on bachelor, specialist and master programs.
  - 9. The number of enterprises that are the basis of practice with which contractual relations are drawn up.
- Research activities
  - 2.1 + 2.2 + 2.3. Number of citations for the last 5 year in Web of Science, Scopus and RSCI.
  - 2.4 + 2.5 + 2.6. Number of scientific papers in Web of Science, Scopus and RSCI per 100 faculty members.
  - 2.7. Total amount of Research & Development.

  - 2.16. Number of grants for the last year per 100 faculty members.
  - 13 + 14. Number of business incubators and technology parks.

- International activities
  - 3.1 + 3.2. Share of foreign students.
  - 3.8. Share of foreign faculty members.
  - 3.9. Number of faculty members working in the university for more than 1 semester.
  - 35. Number of scientific papers published with foreign co-authors.

- Financial and economic activities
  - 4.1. Income of the university per one faculty member.
  - 4.3. Proportion of the average faculty wage to the average wage in the region's economy.
  - 48. University income.

- Infrastructure
  - 5.1. The total area of educational and laboratory facilities per student.
  - 5.6. The number of personal computers per student.
  - 40 + 41 + 42. The area of educational, research and laboratory buildings, and the area of dormitories.
  - 46. Share of personal computers with Internet access.

- Staff composition
  - 6.1 + 6.2. Share of faculty with doctoral and PhD degrees.
  - 6.4. Number of faculty members with doctoral and PhD degrees per 100 students.
  - 28. Faculty average salary.

According to the data of the universities and the SFA approach, there is one problem: the number of observations (11) is much less than the number of variables (24), what is case of regression means that the variance is infinite, so it cannot be used. As the number of universities cannot be increased, the only possible way it to exclude several of inputs. Simple elimination is not productive because it is a loss of information. The best way to handle this problem is to use *principal component analysis*.

Heat map for outputs will be presented in the next section (Figure 2.7), while heat maps for positively and negatively highly correlated inputs are presented in Figure 2.3.

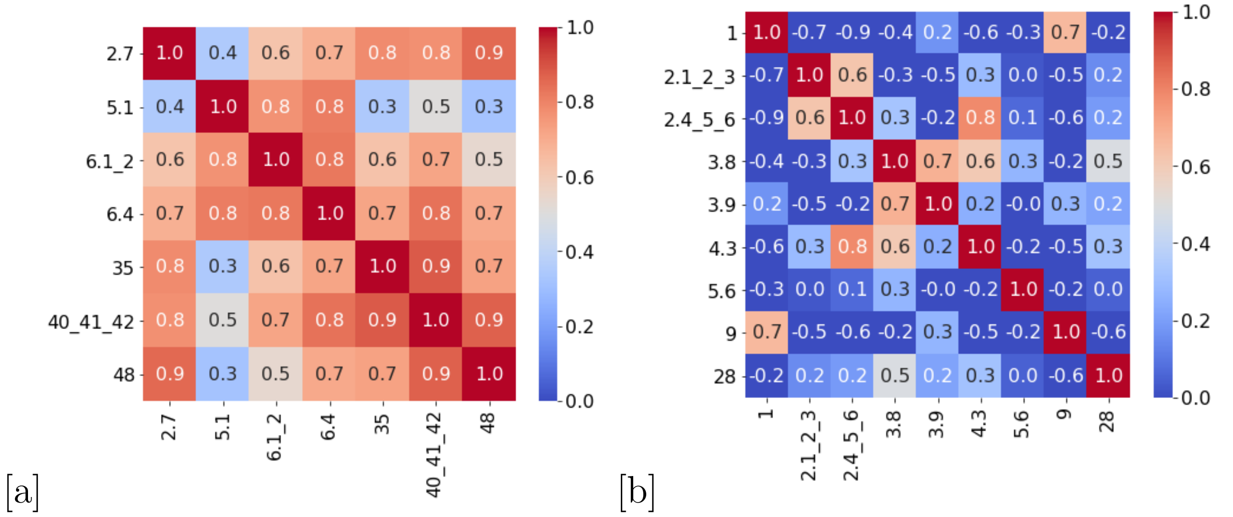

Figure 2.3: (a) Heat map of correlation matrix for positively correlated inputs (b) Heat map of correlation matrix for negatively correlated inputs

First two principal components as axes for each set of variables are plotted in Figure 2.4, where red lines show the initial variables. Plots in Figure 2.4 are called *biplots*. Slopes and directions of lines according to each other can tell something about the relationship between initial variables.

Slopes show the relationships between variables and principal components. Horizontal lines are highly correlated with the parallel axis. Directions illustrate type of relationship: positively correlated variables are co-directed with acute angle, while negatively correlated variables have an obtuse angle between them. It is explicitly seen on biplots: positively correlated inputs and outputs are co-directed with acute angle, while negatively correlated inputs have an obtuse angle between them.

Most principal components for each set of variables are shown in the

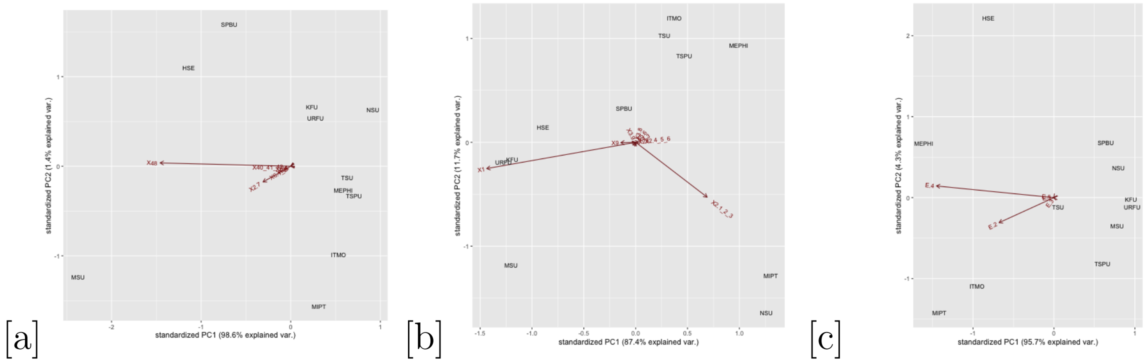

Figure 2.4: (a) Principal components of positively correlated inputs; (b) Principal components of negatively correlated inputs; (c) Principal components of positively correlated outputs.

decreasing order of variance on *screeplots* in Figure 2.5. As can be seen, for positively correlated inputs it is enough to use the first principal component as it explains 98.6%. For negatively correlated inputs the first two components explain 99.7%, what enough to explain variance of these variables. For outputs the first principal component explains 95.6% of data variance. Therefore data were reduced by 16 variables from 24 to 10.

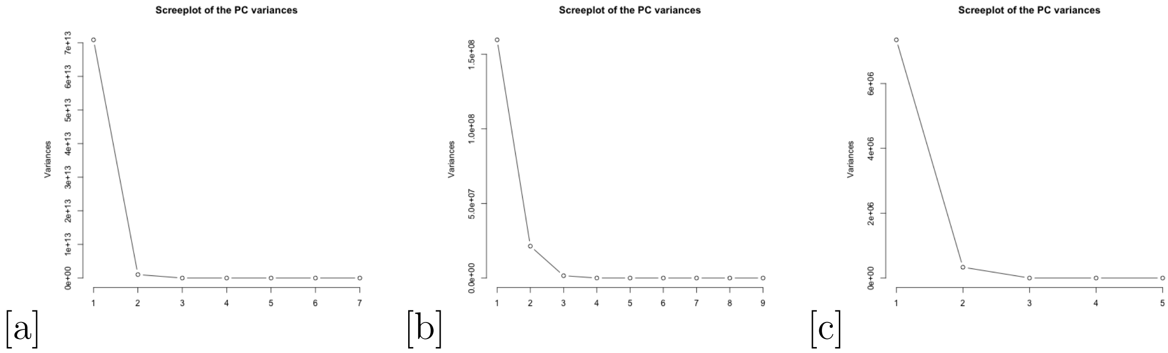

Figure 2.5: (a) Screeplot of PC variances of positive correlated inputs; (b) Screeplot of PC variances of negative correlated inputs; (c) Screeplot of PC variances of highly correlated outputs.

2.4 Exploratory Data Analysis

The important part of the research is exploratory data analysis that allows to understand data well using wide range of visualisations. The data are presented for one year across all the observed Russian universities if not stated explicitly.

The first step is to inspect correlation matrices for all inputs and all outputs. Heat maps for correlation matrices are shown below.

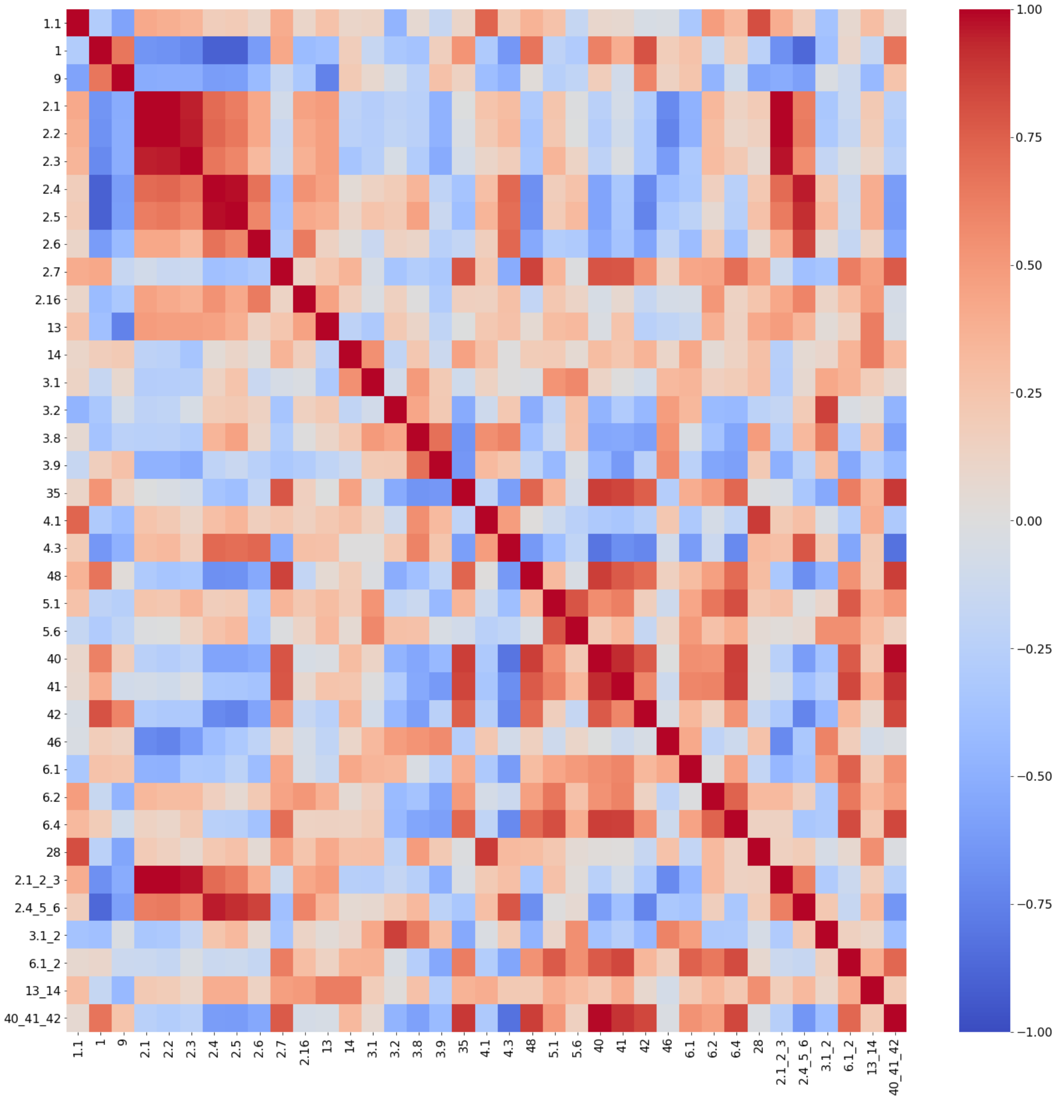

Figure 2.6: Inputs heat map.

As can be seen, there are interesting thing about correlation plots. For instance, indicators 2.1, 2.2 and 2.3 are linearly connected, what means that the proportions of citations in Web of Science, Scopus and RSCI are the same for every university. 2.4 and 2.5 has the correlation coefficient equal 1, so number of papers in Scopus in RSCI are also linearly dependent. Area of educational, research and living area for students are also highly correlated.

There are several inverse linearly dependent input indicators, most of which are in different groups. For instance, the more students are in university the less scientific papers are written there.

The most different indicator in outputs is international activities. All its correlations with other indicators are about 0, although indicators E.1,

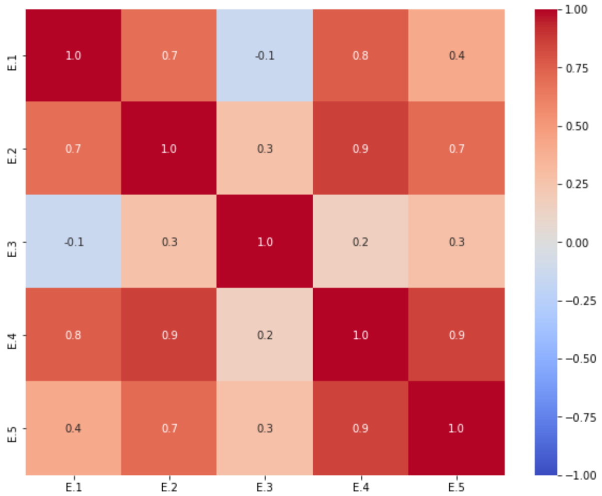

Figure 2.7: Output heat map.

E.2, E.4 and E.5 are correlated with each other for at least 0.4 correlation coefficient.

Several plots in Figure 2.8 show the change of university ranks across several years for QS ranking. There is a strong trend of improving positions in QS Ranking, although some universities had "bad" years, when they lost their positions. The highest rank across all universities for all years is 84 - for Moscow State University in 2020. Also, there is a competition between universities for positions: Saint-Petersburg University had been the second university in Russia until 2019, but in 2020 it lost position and gave it to Novosibirsk State University.

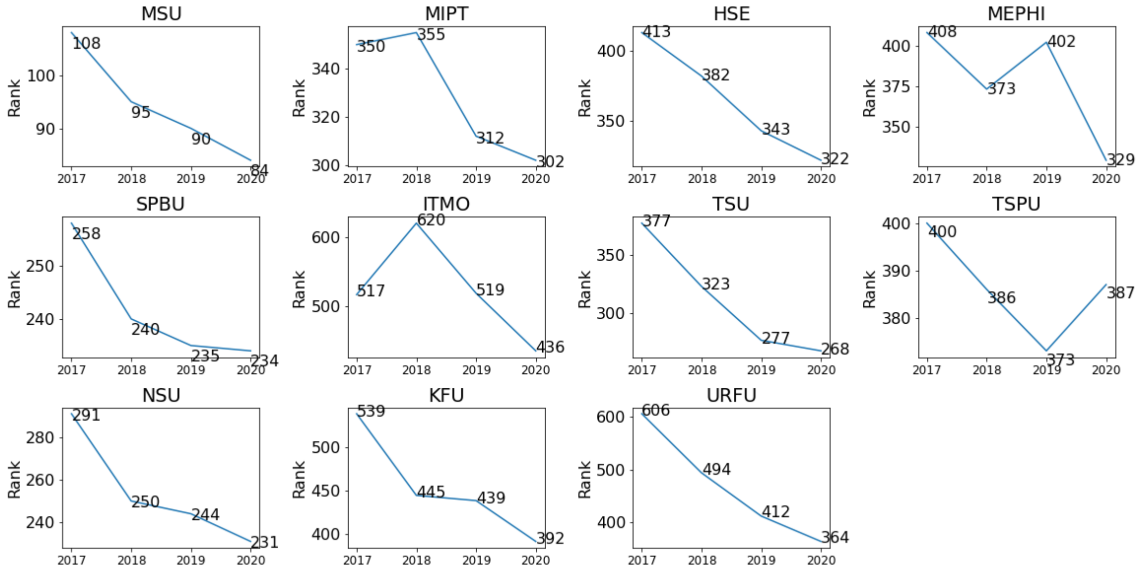

Figure 2.8: Change in ranks across the time for Russian universities in QS ranking.

To understand the difference between universities, plots of inputs across the universities can be derived. It is not necessary to describe all plots, but the most significant ones.

It is interesting to notice the number, or more precisely the proportion, of foreign students in each university (Figure 2.9). As can be seen, two of three most advanced universities in this criterion are not from the large federal cities - Moscow and Saint-Petersburg. The noticeable fact is foreign students for Tomsk universities are mostly from neighboring countries, while MEPhI has 5% more students from far abroad.

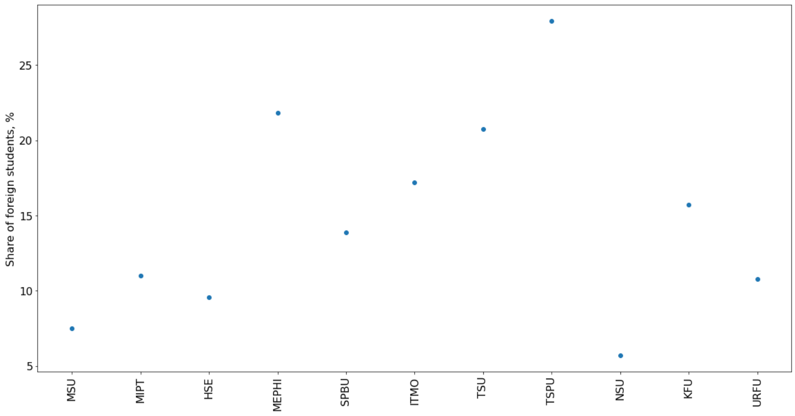

Figure 2.9: Share of foreign students in Russian universities.

Number of citations per 100 faculty members for each university and different databases is depicted in Figure 2.10. For most universities data are condensed, so there is no difference across databases, although for top three universities citations of Russian papers are less than the English ones. Moreover, the most cited database is Scopus, while the least cited is RSCI. It is easily explained by the number of potential readers.

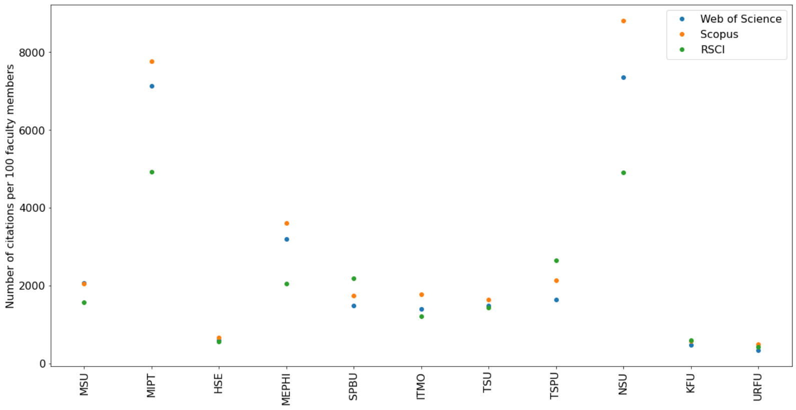

Figure 2.10: Number of citations per 100 faculty members by database.

Despite the above, number of publications in each database does not correlates with the number of citations. (Figure 2.11) The most popular database is Russian RSCI, while there are two universities, which are most presented in the international databases.

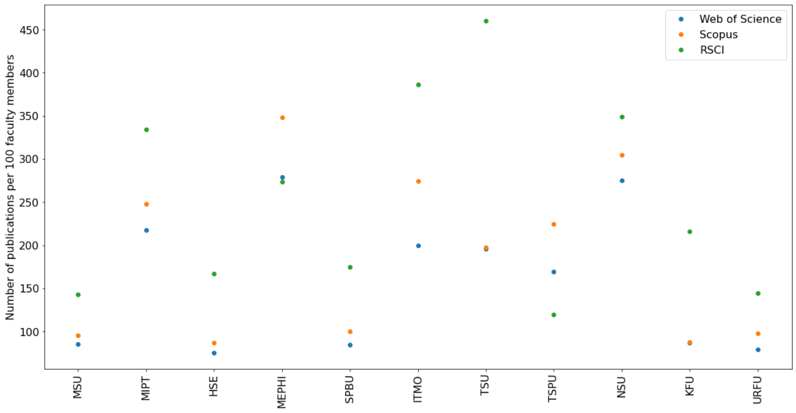

Figure 2.11: Number of publications per 100 faculty members by database.

Possession of some sort of resources may be crucial for students to choose a certain university. For instance, some cannot afford paying rent, so they rely on university to provide a living area during the period of study. It is interesting to notice that some universities have dormitories area even more than educational and laboratory square. (Figure 2.12)

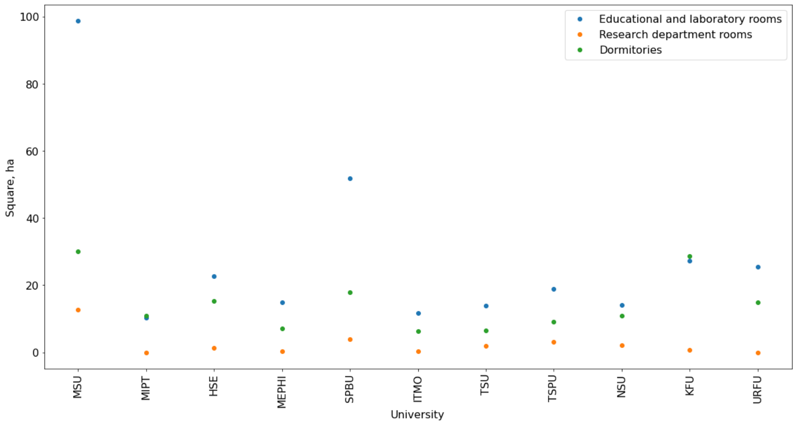

Figure 2.12: Square of university facilities.

To sum up, there is a large difference between universities according to different aspects of their activity. Although some universities prevail in the amount of resources, it does not mean that they manage them more effectively.

Chapter 3. Modeling

3.1 DEA ranking

The DEA model shows efficiencies equal 1 for all universities, what means that all universities are ideally efficient. It was obtained for the both models: input-oriented and output-oriented. It means that efficiencies are very close to each other so they are indistinguishable. Therefore, all universities must have the same rank in ranking. Of course, this situation is almost impossible; hence, the DEA ranking model should not be relied at when making decisions. Nevertheless, it is possible to inspect the aggregated frontier, which is presented in Figure 3.1.

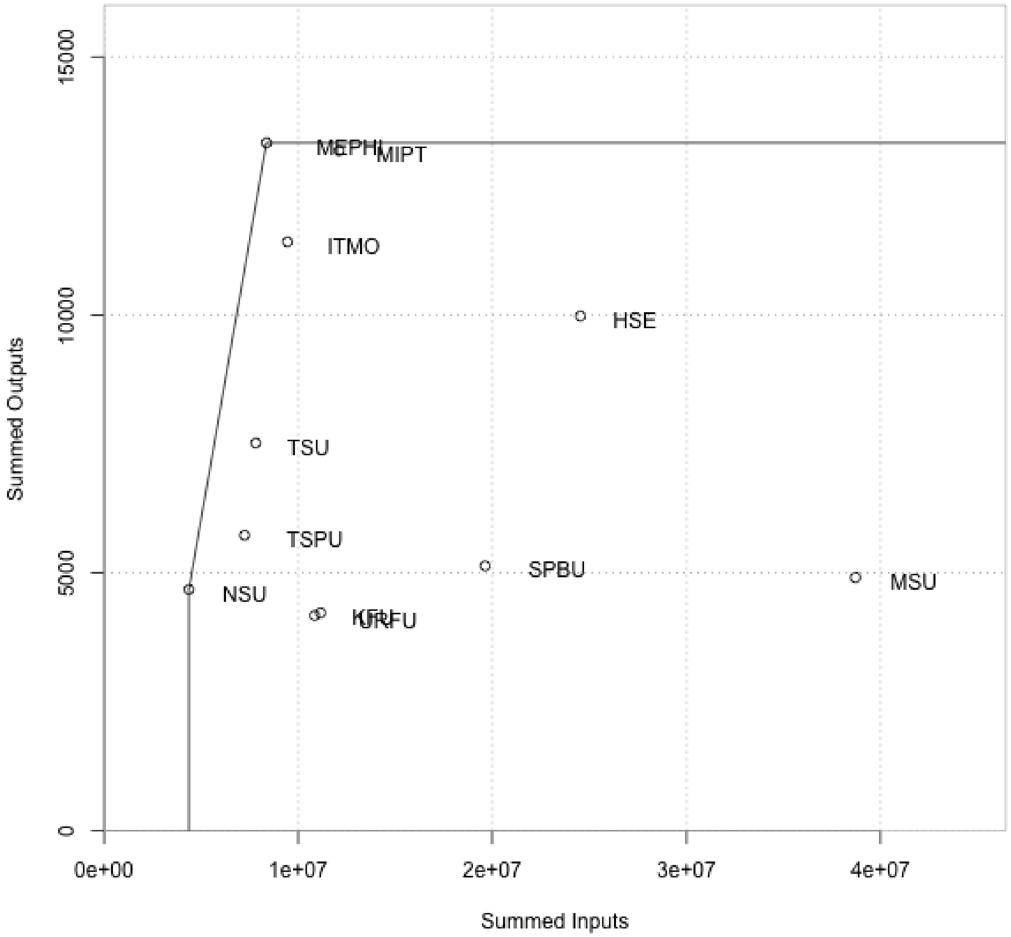

Figure 3.1: Frontier of universities aggregated for inputs and outputs.

Although there are two the most efficient universities on the plot - MEPHI and NSU, they cannot be considered as the universities which works better than others, because the plot is just an example of summarized inputs and outputs.

3.2 SFA ranking

Implementation of the SFA model to the data shows the following results. List of universities with the corresponding efficiency scores are presented below in ascending order.

- ITMO: 0.9858
- KFU: 0.9933
- MIPT: 0.9934
- MSU: 0.9934
- TSPU: 0.9941
- TSU: 0.9946
- NSU: 0.9953
- SPBU: 0.9958
- HSE: 0.9961
- URFU: 0.9965
- MEPHI: 0.9975

As can be seen, it is very close to the DEA efficiency scores, even though they are not the same. Therefore, we can conclude that considered 11 Russian universities are equally effieint comparing to each other. A very important note is that **the efficiencies will be differ, when considering another set of universities or another set of inputs/outputs**.

Graphically SFA scores are shown in figure 3.2.

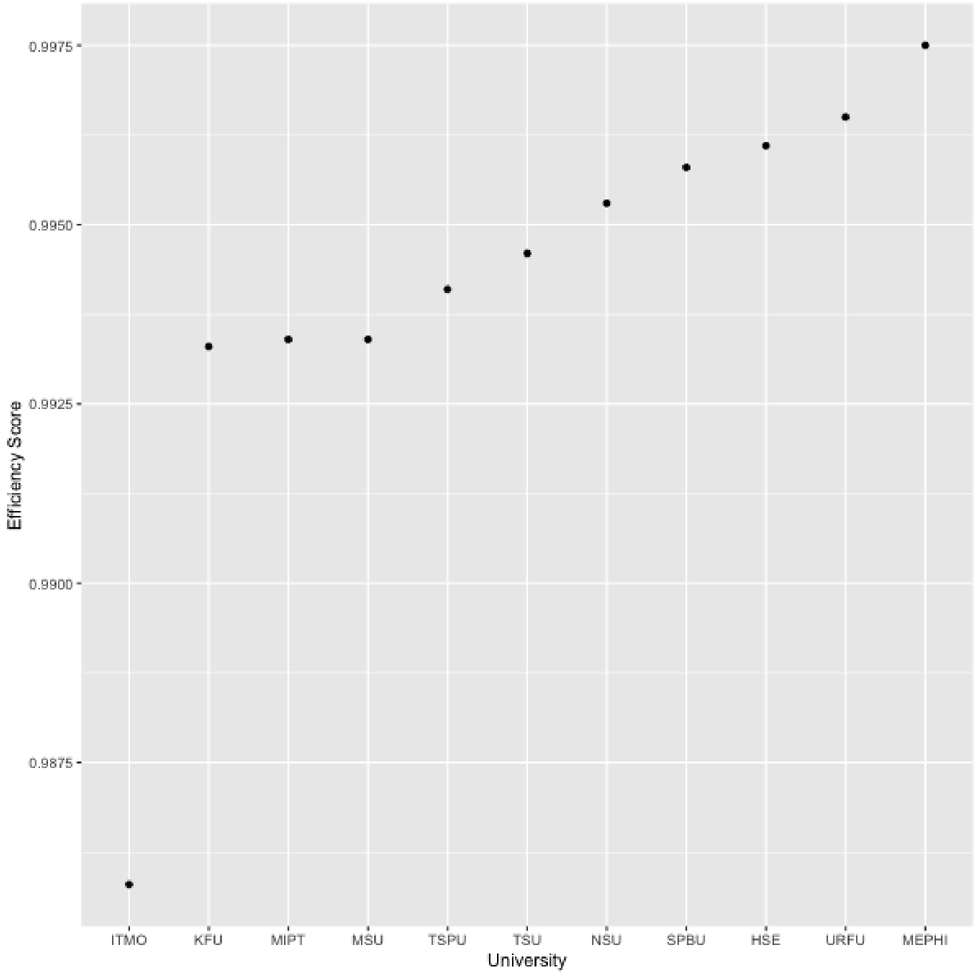

Figure 3.2: Universities with efficiency scores by SFA model.

As can be seen, all universities are highly effective, even though ITMO university is significantly less than others. As all universities are rather efficient, the error term of the model is very low:

$$\lambda = 1,$$

$$\sigma^2 = 1.67 * 10^{-4},$$

$$\sigma_v^2 = 8.354 * 10^{-5},$$

$$\sigma_u^2 = 8.354 * 10^{-5},$$

$$l(\beta, \sigma^2, \lambda) = 28.954.$$

The interesting fact is that more resources does not mean more efficiency for university, because it can use it inefficiently. It is clearly illustrated by MSU, which has the most resources in most categories of inputs, but it only takes the fourth place in ranking.

3.3 Results

Results show that more or less all analyzed Russian universities are equally efficient. While DEA assign equally efficiency scores to all universities, SFA distinguished them by the fourth decimal place.

As it is completely incorrect to compare the SFA model and world university rankings directly, because the set of universities are different, it is still possible to range rankings by the similarity to the SFA model. First, it is necessary to transform ranks to efficiency scores, as the opposite is not possible. All universities in the following analysis are Russian. The next formula is used for each ranking system:

$$ER_i = \frac{R_{max} - R_i}{R_{max}},$$

where R_i - rank of i-th university, R_{max} - maximum rank in ranking, ER_i - efficiency score of i-th university.

Maximum number of rank for each ranking was described previous. Efficiency scores for Russian universities are presented in Table 3.1.

University	SFA	QS	THE	ARWU	CWUR
MSU	0.9934	0.92	0.86	0.91	0.89
MIPT	0.9934	0.70	0.83	0.53	0.73
HSE	0.9961	0.68	0.79	0.08	0.56
MEPHI	0.9975	0.67	0.65	0.36	0.64
SPBU	0.9958	0.77	0.45	0.66	0.72
ITMO	0.9958	0.56	0.67	0.17	0.50
TSU	0.9946	0.73	0.62	0.18	0.51
TSPU	0.9941	0.61	0.49	0.11	0.30
NSU	0.9953	0.77	0.58	0.53	0.70
KFU	0.9933	0.61	0.55	0.15	0.46
URFU	0.9965	0.64	0.28	0.26	0.40

Table 3.1: Russian universities ranks converted into efficiency scores with scores from SFA model.

As we use not all, but only Russian universities, it is not possible to use metrics of ranking quality: *precision*, *average precision*, *mean average precision*, *discounted cumulative gain*, etc. The evaluation metric in our case is *MAE* - median absolute error. Median is used because the distributions of errors are not symmetric. Median absolute errors for each ranking are presented in Table 3.2.

Ranking	QS	THE	ARWU	CWUR
MAE	0.3	0.38	0.64	0.41

Table 3.2: Median absolute errors of the developed SFA model.

The results show that the most closed ranking to the developed one is QS, which has the least error. As QS and THE rankings are similar to each other, the error between them is not large. The most distinct ranking is ARWU with error twice more the least one. Therefore, using such simple metric as median absolute error and developed SFA mode, we can conclude that the most appropriate ranking is QS.

3.4 Further development

The developed models can be improved in many different ways.

First, all Russian universities can be considered, as the program for collecting data of universities is already developed. In this case, there will be the most comprehensive analysis of Russian universities ever developed. This will also solve the problem of lack of universities, because size of sample will be much large than number of variables.

Second, if all Russian universities are taken, number of inputs and outputs can be increased to take all the existing data. Therefore, principal component analysis is not needed in this case and universities can be compared without data manipulation and dimensionality reduction.

Third, costs for inputs can be introduced, what makes rankings sensible since the financial point of view. As prices are associated with inputs for every university, there will not be a problem that different country regions have different prices.

The approach and the steps presented in the research are universal; hence, the procedure of the analysis is reproducible and can be used as a baseline for benchmarking studies in different areas of life.

Conclusion

The purpose of the research was the development university rankings for Russian universities presented in all world university ranking using modern benchmarking approaches - data envelopment analysis and stochastic frontier analysis.

Required preliminary steps were done: overview of literature, data collection, exploratory data analysis, dimensionality reduction of inputs and outputs in order to use SFA modeling.

Classical and modern scientific publications were analysed, several efficiency measurement approaches were introduced with the most important one - Farrell efficiency.

An important step is data collection which was done using two web scrapers written in Python. They gathered data and organized it for further analysis.

Exploratory data analysis helps to understand data and choose the most important variables for further analysis. Also, it shows some interesting facts about universities, such that there are some universities which have dormitory square more than educational and research square.

Principal component analysis was used because there was a problem when the number of observations was less than the number of variables. As some inputs and outputs are highly correlated, it was easy to aggregate them into principal components.

The results of DEA and SFA modeling show that all Russian universities are almost equally efficiency. While DEA ranking assigned equally rank 1 to all universities, SFA showed the distinction in the fourth decimal place.

Bibliography

[1] Benito, M., Gil, P. & Romera, R. Funding, is it key for standing out in the university rankings?. Scientometrics 121, 771–792 (2019).

[2] Bougnol, M., Dulá, J.H. Technical pitfalls in university rankings. High Educ 69, 859–866 (2015) doi:10.1007/s10734-014-9809-y

[3] Bougnol, M., Dulá, J.H. Validating DEA as a ranking tool: An application of DEA to assess performance in higher education. Ann Oper Res 145, 339–365 (2006) doi:10.1007/s10479-006-0039-2

[4] Brock, A. Budgeting models and university efficiency: A Ghanaian case study. High Educ 32, 113–127 (1996).

[5] Buela-Casal, G., Gutiérrez-Martínez, O., Bermúdez-Sánchez, M.P. et al. *Comparative study of international academic rankings of universities.* Scientometrics 71, 349–365 (2007) doi:10.1007s11192-007-1653-8

[6] Çakır, M.P., Acartürk, C., Alaşehir, O. et al. A comparative analysis of global and national university ranking systems. Scientometrics 103, 813–848 (2015).

[7] Chen, K., Liao, P. A comparative study on world university rankings: a bibliometric survey. Scientometrics 92, 89–103 (2012).

[8] Daraio, C. and Bonaccorsi, A. (2017), Beyond university rankings? Generating new indicators on universities by linking data in open platforms. Journal of the Association for Information Science and Technology, 68: 508-529. doi:10.1002/asi.23679

[9] De Filippo, D., Casani, F., García-Zorita, C. et al. Visibility in international rankings. Strategies for enhancing the competitiveness of Spanish universities. Scientometrics 93, 949–966 (2012).

[10] Elken, M., Hovdhaugen, E. & Stensaker, B. Global rankings in the Nordic region: challenging the identity of research-intensive universities?. High Educ 72, 781–795 (2016).

[11] M. J. Farrell *Journal of the Royal Statistical Society. Series A (General)* Vol. 120, No. 3 (1957), pp. 253-290.

[12] Gómez, I., Bordons, M., Fernández, M.T. et al. Structure and research performance of Spanish universities. Scientometrics 79, 131–146 (2009).

[13] Johnes, J. University rankings: What do they really show?. Scientometrics 115, 585–606 (2018).

[14] Kao, C., Pao, H. An evaluation of research performance in management of 168 Taiwan universities. Scientometrics 78, 261 (2009) doi:10.1007/s11192-007-1906-6

[15] Khalil Jahangiri email ; Ali Rezazadeh1; Mohsen Poorebadollahan2

[16] Lovell CAK (1993). *"Production Frontiers and Productive Efficiency."* in Fried HO and SS Schmidt (eds.) the Measurement of Productive Efficiency: Techniques and Applications, Oxford U.K.: 3-67.

[17] Moed, H.F. A critical comparative analysis of five world university rankings. Scientometrics 110, 967–990 (2017).

[18] Piro, F.N., Sivertsen, G. How can differences in international university rankings be explained?. Scientometrics 109, 2263–2278 (2016) doi:10.1007/s11192-016-2056-5.

[19] van Raan, A.F.J., van Leeuwen, T.N. & Visser, M.S. Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings. Scientometrics 88, 495–498 (2011) doi:10.1007/s11192-011-0382-1.

[20] Pietrucha, J. Country-specific determinants of world university rankings. Scientometrics 114, 1129–1139 (2018) doi:10.1007/s11192-017-2634-1

[21] Robinson-García, N., Torres-Salinas, D., Delgado López-Cózar, E. et al. An insight into the importance of national university rankings in an international context: the case of the I-UGR rankings of Spanish universities. Scientometrics 101, 1309–1324 (2014) doi:10.1007/s11192-014-1263-1.

[22] Safón, V. What do global university rankings really measure? The search for the X factor and the X entity. Scientometrics 97, 223–244 (2013).

[23] Salas-Velasco, M. The technical efficiency performance of the higher education systems based on data envelopment analysis with an illustration for the Spanish case. Educ Res Policy.

[24] Schwekendiek, D. Recent changes in World University Rankings: an explorative study of Korea and Germany. Asia Eur J 13, 361–377 (2015).

[25] Shehatta, I., Mahmood, K. *Correlation among top 100 universities in the major six global rankings: policy implications.* Scientometrics 109, 1231–1254 (2016) doi:10.1007/s11192-016-2065-4

[26] Yudkevich, M., Altbach, P.G. & Rumbley, L.E. Global university rankings: The "Olympic Games" of higher education?. Prospects 45, 411–419 (2015) doi:10.1007/s11125-015-9365-y

[27] Xueqian Zhang, Yalin Qian & Qiangqiang Zhao, Performance Evaluation of Scientific Research in China's First-class Universities from the Perspective of Multi-agent Based on Malmquist index, SE-DEA and SFA Respectively. 2nd International Conference on Economy, Management and Entrepreneurship (ICOEME 2019), doi:https://doi.org/10.2991/icoeme-19.2019.64.

[28] Timothy J. Coelli, D.S. Prasada Rao, Christopher J. O'Donnell, and George E. Battese. *An Introduction to Efficiency and Productivity Analysis.* Springer Science+Business Media, Inc., 2005.

[29] Peter Bogetoft, Lars Otto. *Benchmarking with DEA, SFA and R.* Springer Science+Business Media, LLC., 2011.

[30] QS World University Rankings - Top Universities, `https://www.topuniversities.com/university-rankings`

[31] World University Rankings — Times Higher Education (THE), `https://www.timeshighereducation.com/world-university-rankings`

[32] ARWU World University Rankings 2019 — Academic Ranking of World Universities
http://www.shanghairanking.com/

[33] CWUR — Center for World University Rankings https://cwur.org/

[34] Monitoring — Information and analytical materials based on the results monitoring the quality of training
http://indicators.miccedu.ru/monitoring/?m=spo

Appendix

Appendix 1. Programming code

Program is written in Python and R programming languages.

```
# coding=utf−8
import requests
import pandas as pd
import numpy as np
import seaborn as sns
import geopandas, geoplot, random, re, scipy
from bs4 import BeautifulSoup as BS
from mpl_toolkits.axes_grid1 import make_axes_locatable
import matplotlib.pyplot as plt
import matplotlib
from matplotlib.pyplot import figure, savefig
from sklearn.decomposition import PCA
from sklearn.metrics import mean_absolute_error as mae

font = {’size’: 16}
matplotlib.rc(’font’, **font)

rankings = [’THE’, ’QS’, ’CWUR’, ’ARWU’]
countries = {
    ’AG’: ’Argentina’, ’AU’: ’Austria’, ’AUS’: ’Australia’, ’BE’: ’Belgium’, ’
        BEL’: ’Belarus’, ’BZ’: ’Brazil’,
    ’CA’: ’Canada’, ’CH’: ’China’, ’CHI’: ’Chile’, ’COL’: ’Colombia’, ’CZ’: ’
        Czech’, ’DN’: ’Denmark’,
    ’EG’: ’Egypt’, ’ES’: ’Estonia’, ’FI’: ’Finland’, ’FR’: ’France’, ’GE’: ’
        Germany’, ’GR’: ’Greece’,
    ’IN’: ’India’, ’ID’: ’Indonesia’, ’IR’: ’Ireland’, ’IRN’: ’Iran’, ’IT’: ’Italy’,
        ’IS’: ’Izrael’,
    'JP': 'Japan', 'KO': 'Korea', 'KZ': 'Kazakhstan', 'LE': 'Lebanon', 'MA':
        'Malaysia', 'ME': 'Mexico',
    'NE': 'Netherlands', 'NO': 'Norway', 'NZ': 'New␣Zealand', 'PH': '
        Philippines', 'PAK': 'Pakistan',
    'POL': 'Poland', 'POR': 'Portugal', 'RU': 'Russia', 'SA': 'South␣Africa',
        'SAU': 'Saudi␣Arabia',
    'SG': 'Singapore', 'SP': 'Spain', 'SW': 'Sweden', 'TH': 'Thailand', 'UA':
        'Ukraine',
    'UAE': 'United␣Arab␣Emirates', 'UK': 'United␣Kingdom', 'US': 'United
        ␣States␣of␣America'
}

media_dir = 'Media/'
EDA_dir = 'Media/EDA/'
rankings_dir = 'Media/Rankings'

inputs_all = list(map(str, list(range(1, 57))))
inputs_all.remove('47')

edu = list(map(str, list(range(1, 10))))
science = list(map(str, list(range(10, 21))))
staff = list(map(str, list(range(21, 30))))
inter = list(map(str, list(range(30, 39))))
infra = list(map(str, list(range(39, 47))))
fin = list(map(str, list(range(48, 57))))



inputs = ['1.1', '1', '9', '2.1', '2.2', '2.3', '2.4', '2.5', '2.6', '2.7', '2.16', '13', '
    14', '3.1', '3.2',
          '3.8', '3.9', '35', '4.1', '4.3', '48', '5.1', '5.6', '40', '41', '42', '46', '
              6.1', '6.2', '6.4', '28',
          '2.1_2_3', '2.4_5_6', '3.1_2', '6.1_2', '13_14', '40_41_42']
inputs_grouped = ['1.1', '1', '9', '2.7', '2.16', '3.8', '3.9', '35', '4.1', '4.3', '48',
            '5.1', '5.6', '46', '6.4', '28', '2.1_2_3', '2.4_5_6', '3.1_2', '6.1_2', '13
                _14', '40_41_42']
outputs = ['E.1', 'E.2', 'E.3', 'E.4', 'E.5']

def set_rank(rank_range):
    """
    Set university rank if it is defined as range, but not a single value

    Parameters:
    rank_range (str): Consists university rank. Two formats: "a−b" or "c
        +", where a, b, c are integers and a < b

    Return
    int: University rank as an integer
    """

    range_min_max = re.search('(\d+)?−?(\d+)?\+?', rank_range)

    if range_min_max[2] is None:
        return int(range_min_max[1])
    else:
        rank = random.randint( int(range_min_max[1]), int(
            range_min_max[2]) )
        return int(rank)

def kendall_w(data):
    """
    Calculate and return Kendall's coefficient of concordance

    Parameters
    data (Pandas dataframe): Dataframe of data

    Return
    int: W coefficient
    """

    m = 4
    n = len(data)
    Ri = data.sum(axis = 1)
    T = 0
    for ranking in data:
        vc = data[ranking].value_counts()
        d = vc[vc > 1]
        t = np.sum(list(map(lambda t: t ** 3 - t, d)))
        T = T + t

    W = (12 * np.sum(Ri) - 3 * m ** 2 * n * (n + 1) ** 2) / (m ** 2 *
        n * (n ** 2 - 1) - m * T)
    rS = (m * W - 1) / (m - 1)

    return (T, W, rS)

url = 'http://indicators.miccedu.ru/monitoring/index.php?m=vpo'
r = requests.get(url)
r.encoding = r.apparent_encoding
page = BS(r.text, 'html.parser')

columns = []
uni_id_list = []
uni = []
data = pd.DataFrame(data = [[0, 1, 2, 3, 4]],
                    columns = ['uni_id', 'index', 'desc', 'dim', 'measure'])

for state in page.select('p.MsoListParagraph_a[href]'): # through regions
    region_url = host + str(year) + '/' + state['href']
    r = requests.get(region_url)
    r.encoding = r.apparent_encoding
    state_page = BS(r.content)
    for uni in state_page.select('.blockcontent_tr_td.inst_a[href]'): #
       through unis
        uni_id = int(uni['href'][12:])
        uni.append(uni.get_text())
        uni_id_list.append(uni_id)
        uni_url = host + '_vpo/' + uni['href']
        r = requests.get(uni_url)
        r.encoding = r.apparent_encoding
        uni_page = BS(r.content)

        for indicator in uni_page.select('table#analis_dop_tr'): # through
           the last table
            fields = []
            td_num = len(indicator.find_all('td'))

            if td_num == 4: # skip headings and not full rows
                for td in indicator.select('td'):
                    fields.append(td.get_text())
                row = pd.DataFrame(data = [[uni_id, fields[0], fields[1],
                   fields[2], fields[3]]],
                                        columns = ['uni_id', 'index', 'desc', '
                                           dim', 'measure'])
                data = data.append(row)

data = data.iloc[3:, :]
data.set_index('uni_id', inplace = True)
data.to_excel('russian_uni.xlsx', index = False)

url = 'http://indicators.miccedu.ru/monitoring/_vpo/inst.php?id='
unis = pd.read_excel('Data/Russian_Universities_Initial.xlsx')

for index, uni in unis.iterrows():
    r = requests.get(url + str(uni.Uni_Id))
```

```
r.encoding = r.apparent_encoding
page = BS(r.text, 'html.parser')


for indicator in page.select('table#analis_dop_tr'): # through the last table
    fields = []
    td_num = len(indicator.find_all('td'))

    if td_num == 4: # skip headings and not full rows
        not_valid = False
        for ix, td in enumerate(indicator.select('td')):
            if re.search('', td.get_text()) or (ix == 2 and re.search('', td.get_text())):
                not_valid = True
                break
            else:
                fields.append(td.get_text())

        if not_valid == False:
            unis.loc[index, str(fields[0])] = float(fields[3].replace('␣', '').replace(',', '.'))

for table in page.select('table.napde'):
    row = []
    for ix, indicator in enumerate(table.select('td')): # through the last table
        if ix in [0, 1, 2, 3]:
            continue
        else:
            res = divmod(ix, 4)
            if res[1] in [0, 3]:
                row.append(indicator.get_text())
                if res[1] == 3:
                    unis.loc[index, str(row[0])] = float(row[1].replace('␣', ''
                        ).replace(',', '.'))
                    row = []

    # working with outpus
    # there is an additional empty tbody before thead of the result table, so
        we cannot use BS
    # use regex instead
    Es = re.findall('(E\.\d)</td>', str(page))
    output_values = re.findall('right␣center␣no−repeat;\">(\d+(?:,\d+)?)
        <', str(page))
    income = re.findall('<span␣style=\"\">(\d+(?:,\d+)?)<', str(page))
    output_values.insert(4, income[0])

    for ix, val in enumerate(output_values):
        unis.loc[index, Es[ix]] = float(val.replace(',', '.'))



unis['2.1_2_3'] = unis['2.1'] + unis['2.2'] + unis['2.3']
unis['2.4_5_6'] = unis['2.4'] + unis['2.5'] + unis['2.6']
unis['3.1_2'] = unis['3.1'] + unis['3.2']
unis['6.1_2'] = unis['6.1'] + unis['6.2']
unis['13_14'] = unis['13'] + unis['14']
unis['40_41_42'] = (unis['40'] + unis['41'] + unis['42']) / 10000

for index, university in unis.iterrows():
    for ranking in ['THE', 'QS', 'CWUR', 'ARWU']:
        value = unis.loc[unis.University == university[0], ranking]
        if not (isinstance(value[index], int) or value[index] is None or
            isinstance( value[index], float)):
            unis.loc[unis.University == university[0], ranking] = set_rank(
                value[index])
```

```
unis.to_excel('Data/Russian_Universities.xlsx', index = False)
unis.to_csv('Data/Russian_Universities.csv', index = False)

unis = pd.read_excel('Data/Russian_Universities.xlsx')
data = pd.read_excel('Data/Russian_Universities_TS_Initial.xlsx')
data.loc[(data.Year == 2020) & (data.Ranking == 'THE')].Rank = unis.
    THE
data.loc[(data.Year == 2020) & (data.Ranking == 'QS')].Rank = unis.QS
data.loc[(data.Year == 2020) & (data.Ranking == 'ARWU')].Rank = unis
    .ARWU
data.loc[(data.Year == 2020) & (data.Ranking == 'CWUR')].Rank = unis
    .CWUR

for index, row in data.iterrows():
    rank = row[2] # rank
    if not isinstance(rank, int) and not (rank is None) and not
        isinstance( rank, float):
         data.loc[index, 'Rank'] = set_rank(rank)

data.to_excel('Data/Russian_Universities_TS.xlsx', index = False)

url = 'http://www.shanghairanking.com/ARWU2019.html'
r = requests.get(url)
page = BS(r.text, 'html.parser')

data = pd.DataFrame(data = [[0, 1, 2, 3, 4, 5, 6, 7]],
                         columns = ['Rank', 'Uni', 'Alumni', 'Award', 'HiCi', 'N
                            &S', 'PUB', 'PCP'])

for i, uni in enumerate(page.select('table#UniversityRanking_tr')):
    if i == 0:
        continue
    else:
        fields = []
        for j, indicator in enumerate(uni.select(’td’)):
            if j in [0, 1, 5, 6, 7, 8, 9, 10]:
                if j == 1:
                    ind = (indicator.select(’a’))[0].text
                else:
                    ind = indicator.text
                fields.append(ind)
        row = pd.DataFrame(data = [[fields[0], fields[1], fields[2], fields[3],
            fields[4], fields[5], fields[6], fields[7]]],
                                            columns = [’Rank’, ’Uni’, ’Alumni’, ’
                                                Award’, ’HiCi’, ’N&S’, ’PUB’, ’
                                                PCP’])
        data = data.append(row)

data.set_index(’Rank’, inplace = True)
data.to_excel(’Data/ARWU_Ranking_Full.xlsx’)

world = (pd.read_excel(’Data/All_Universities_2020.xlsx’)).iloc[:500, :29]
uni = world.University
world = world.set_index(’University’)
world = world.dropna()

world[’Country’] = [countries[code] for code in world.loc[:, ’Country_Code’
    ]]

# Fill ranged ranks (101–150, 800–1000)
for ranking in rankings:
    for rank_i in range(len(world[ranking])):
        rank = world[ranking][rank_i]
        if isinstance(rank, str):
            world.loc[:, ranking][rank_i] = set_rank(rank)

# Convert to ranks
rdata = (world.reset_index())[rankings].rank(axis = 0, ascending = True)
```

```
rdata[’University’] = uni
rdata = rdata.set_index(’University’)

world.THE = rdata.THE
world.QS = rdata.QS
world.CWUR = rdata.CWUR
world.ARWU = rdata.ARWU
rdata.head()

# All data
corr_matrix = scipy.stats.spearmanr(world.loc[:, rankings])[0]
sns.heatmap(corr_matrix, vmin = 0, vmax = 1, annot = True, xticklabels
    = rankings, yticklabels = rankings)
plt.savefig(media_dir + ’spearman_corr.png’, bbox_inches=’tight’)

# 1−100 and 101−200
f, axes = plt.subplots(1, 2, figsize = (16, 6))
corr_matrix_100 = scipy.stats.spearmanr(world.loc[:, rankings][1:100])[0]
corr_matrix_200 = scipy.stats.spearmanr(world.loc[:, rankings][101:200])[0]

sns.heatmap(corr_matrix_100, vmin = 0, vmax = 1, annot = True, ax =
    axes[0], xticklabels = rankings, yticklabels = rankings)
sns.heatmap(corr_matrix_200, vmin = 0, vmax = 1, annot = True, ax =
    axes[1], xticklabels = rankings, yticklabels = rankings)
plt.savefig(media_dir + ’spearman_corr_1−100−200.png’, bbox_inches=’
    tight’)

rusuni = pd.read_excel(’Data/Russian_Universities.xlsx’)
rusuni.dropna(inplace = True)
rusuni.reset_index(drop = True, inplace = True)
rusuni = rusuni.loc[:, inputs + outputs + rankings + [’University’, ’Region’
    ]]
rusuni.head()
```

```
corr = rusuni.loc[:, inputs].corr()
figure(figsize = (25, 25))
sns.heatmap(corr, vmin = -1, vmax = 1, cmap = 'coolwarm', fmt=".1f")
plt.savefig(media_dir + 'heatmap_inputs.png', bbox_inches='tight')

corr = rusuni.loc[:, inputs_grouped].corr()
figure(figsize = (25, 25))
sns.heatmap(corr, vmin = -1, vmax = 1, cmap = 'coolwarm', fmt=".1f")
plt.savefig(media_dir + 'heatmap_inputs_grouped.png', bbox_inches='tight')

# Data exploration before PCA
f, axes = plt.subplots(1, 2, figsize = (16, 6))
positive_cor = ['2.7', '5.1', '6.1_2', '6.4', '35', '40_41_42', '48']
negative_cor = ['1', '2.1_2_3', '2.4_5_6', '3.8', '3.9', '4.3', '5.6', '9', '28']
output_cor_shrink = ['E.1', 'E.2', 'E.4', 'E.5']
output_cor = ['E.1', 'E.2', 'E.3', 'E.4', 'E.5']

corr_matrix_positive = rusuni.loc[:, positive_cor].corr()
corr_matrix_negative = rusuni.loc[:, negative_cor].corr()

sns.heatmap(corr_matrix_positive, vmin = 0, vmax = 1, annot = True, ax
    = axes[0], cmap = 'coolwarm', fmt=".1f")
sns.heatmap(corr_matrix_negative, vmin = 0, vmax = 1, annot = True, ax
    = axes[1], cmap = 'coolwarm', fmt=".1f")
plt.savefig(media_dir + 'spearman_corr_strong_weak.png', bbox_inches='
    tight')

print('Number_of_variables:', len(corr) - len(positive_cor) - len(
    negative_cor) + 3)

corr = ['1.1', '13_14', '2.16', '3.1_2', '4.1', '46']
corr_matrix = rusuni.loc[:, corr].corr()
sns.heatmap(corr_matrix, vmin = 0, vmax = 1, annot = True, cmap = '
    coolwarm', fmt=".1f")
```

```
from sklearn.decomposition import PCA

pca_pos = PCA(n_components = 2)
pca_pos.fit(rusuni.loc[:, positive_cor].T)
PC1 = np.round(pca_pos.components_[0], 4)
# print('Explained variance ration positive: ', pca_pos.
    explained_variance_ratio_)
print('PC1␣explained:', pca_pos.explained_variance_ratio_[0])


pca_neg = PCA(n_components = 2)
pca_neg.fit(rusuni.loc[:, negative_cor].T)
PC2 = np.round(pca_neg.components_[0], 4)
PC3 = np.round(pca_neg.components_[1], 4)
# print('Explained variance ration negative: ', pca_neg.
    explained_variance_ratio_)
print('PC2,␣PC3␣explained:', pca_neg.explained_variance_ratio_[0:2])


pca_out = PCA(n_components = 2)
pca_out.fit(rusuni.loc[:, output_cor].T)
PC4 = np.round(pca_out.components_[0], 4)
# print('Explained variance ration negative: ', pca_neg.
    explained_variance_ratio_)
print('PC4␣explained:', pca_out.explained_variance_ratio_[0])

pca_out = PCA(n_components = 2)
pca_out.fit(rusuni.loc[:, output_cor_shrink].T)
PC5 = np.round(pca_out.components_[0], 4)
# print('Explained variance ration negative: ', pca_neg.
    explained_variance_ratio_)
print('PC5␣explained:', pca_out.explained_variance_ratio_[0])
```

```
rusuni.loc[:, 'INPUT1'] = PC1
rusuni.loc[:, 'INPUT2'] = PC2
rusuni.loc[:, 'INPUT3'] = PC3
rusuni.loc[:, 'OUTPUT_FULL'] = PC4
rusuni.loc[:, 'OUTPUT_SHRINK'] = PC5

rusuni.to_csv('Data/Russian_Universities.csv', index = False)

corr = rusuni.loc[:, outputs].corr()
figure(figsize = (10, 8))
sns.heatmap(corr, annot = True, vmin = -1, vmax = 1, cmap = '
    coolwarm', fmt=".1f")
plt.savefig(media_dir + 'heatmap_outputs.png', bbox_inches='tight')

for input in inputs:
    figure(figsize = (20, 10))
    plt.plot(rusuni.University, rusuni[input])
    plt.title(input)
    plt.xticks(rotation = 90)
    plt.savefig(EDA_dir + 'Distributions/' + input + '.png', bbox_inches=
        'tight')
    plt.close()

for output in outputs:
    figure(figsize = (20, 10))
    plt.plot(rusuni.University, rusuni[output])
    plt.title(output)
    plt.xticks(rotation=90)
    plt.savefig(EDA_dir + 'Distributions/' + output + '.png', bbox_inches
        ='tight')
    plt.close()

figure(figsize = (20, 10))
plt.plot(rusuni.University, rusuni['3.1_2'], 'o')
plt.xlabel('University')
plt.ylabel('Share␣of␣foreign␣students,␣%')
plt.xticks(rotation = 90)
plt.savefig(EDA_dir + '/3-1_2_explained.png', bbox_inches='tight')

figure(figsize = (20, 10))
plt.plot(rusuni.University, rusuni['2.1'], 'o')
plt.plot(rusuni.University, rusuni['2.2'], 'o')
plt.plot(rusuni.University, rusuni['2.3'], 'o')
plt.legend(['Web␣of␣Science', 'Scopus', 'RSCI'])
plt.xlabel('University')
plt.ylabel('Number␣of␣citations␣per␣100␣faculty␣members')
plt.xticks(rotation = 90)
plt.savefig(EDA_dir + '/2-1_2_3_explained.png', bbox_inches='tight')

figure(figsize = (20, 10))
plt.plot(rusuni.University, rusuni['2.4'], 'o')
plt.plot(rusuni.University, rusuni['2.5'], 'o')
plt.plot(rusuni.University, rusuni['2.6'], 'o')
plt.legend(['Web␣of␣Science', 'Scopus', 'RSCI'])
plt.xlabel('University')
plt.ylabel('Number␣of␣publications␣per␣100␣faculty␣members')
plt.xticks(rotation = 90)
plt.savefig(EDA_dir + '/2-4_5_6_explained.png', bbox_inches='tight')

figure(figsize = (20, 10))
plt.plot(rusuni.University, rusuni['40'] / 10000, 'o')
plt.plot(rusuni.University, rusuni['41'] / 10000, 'o')
plt.plot(rusuni.University, rusuni['42'] / 10000, 'o')
plt.legend(['Educational␣and␣laboratory␣rooms', 'Research␣department␣
    rooms', 'Dormitories'])
plt.xlabel('University')
plt.ylabel('Square,␣ha')
plt.xticks(rotation = 90)
plt.savefig(EDA_dir + '/40_41_42_explained.png', bbox_inches='tight')

for input in inputs:
    figure(figsize = (20, 10))
    sns.boxplot(rusuni[input])
    plt.title(input)
    plt.xticks(rotation = 90)
    plt.savefig(EDA_dir + 'BoxPlots/' + input + '.png', bbox_inches='
        tight')
    plt.close()

for output in outputs:
    figure(figsize = (20, 10))
    sns.boxplot(rusuni[output])
    plt.title(output)
    plt.xticks(rotation=90)
    plt.savefig(EDA_dir + '/BoxPlots/' + output + '.png', bbox_inches='
        tight')
    plt.close()

ts = pd.read_excel('Data/Russian_Universities_TS.xlsx')
QS = ts.loc[(ts.University.isin(rusuni.University)) & (ts.Ranking == 'QS')]
QS.dropna(inplace = True)

fig, axes = plt.subplots(4, 3,figsize = (16, 14), constrained_layout = True)

k = 0
for i, row in enumerate(axes):
    for j, col in enumerate(row):
        if k < len(rusuni.University):
            uni_name = rusuni.University[k]
            u = QS.loc[QS.University == uni_name]
            u.reset_index(drop = True, inplace = True)
            axes[i, j].plot(u.Year, u.Rank)
            axes[i, j].set_xticks(u.Year)
            axes[i, j].set_xticklabels(u.Year, fontsize = 12)
            axes[i, j].set_title(uni_name)
            axes[i, j].set_ylabel('Rank')


            for ix, _ in enumerate(u.Year):
                axes[i, j].text(u.Year[ix], int(u.Rank[ix]) - 3, int(u.Rank[ix]))
        k = k + 1

axes[-1, -1].axis('off')
plt.savefig(media_dir + 'Russian_TS.png')

rusuni = pd.read_excel('Data/Russian_Universities.xlsx')
rusuni.dropna(inplace = True)
rusuni.reset_index(drop = True, inplace = True)
rusuni = rusuni.loc[:, inputs + outputs + rankings + ['University', 'Region']]
rusuni.head(15)

# number of universities in each ranking
THE_len = 1397
QS_len = 1002
CWUR_len = 2000
ARWU_len = 1000

rusuni['THE_eff'] = (THE_len - rusuni.THE) / THE_len
rusuni['QS_eff'] = (QS_len - rusuni.QS) / QS_len
rusuni['CWUR_eff'] = (CWUR_len - rusuni.CWUR) / CWUR_len
rusuni['ARWU_eff'] = (ARWU_len - rusuni.ARWU) / ARWU_len
rusuni['SFA_eff'] = [0.9934, 0.9934, 0.9961, 0.9975, 0.9958, 0.9858, 0.9946, 0.9941, 0.9953, 0.9933, 0.9965]
```

```
rusuni.head(11)

print(’MAE␣THE:’, mae(rusuni.SFA_eff, rusuni.THE_eff))
print(’MAE␣QS:’, mae(rusuni.SFA_eff, rusuni.QS_eff))
print(’MAE␣ARWU:’, mae(rusuni.SFA_eff, rusuni.ARWU_eff))
print(’MAE␣CWUR:’, mae(rusuni.SFA_eff, rusuni.CWUR_eff))


Sys.setlocale(category = ”LC_ALL”, locale = ”UTF−8”)

library(knitr)
library(Benchmarking)
library(ggplot2)
library(dplyr)

thesis_url <− ’˜/Google␣Drive///Research/Programs/’
data <− read.csv(’˜/Google␣Drive///Research/Programs/Data/
    Russian_Universities.csv’,
                  header=TRUE, sep=”,”, dec=”.”) %>% na.omit()
uni <− data %>% select(University)

# INPUT / OUTPUT COLUMNS
input_cols <− c(’X1.1’, ’X1’, ’X9’, ’X2.7’, ’X2.16’, ’X3.8’, ’X3.9’, ’X35’, ’
    X4.1’, ’X4.3’,
                    ’X48’, ’X5.1’, ’X5.6’, ’X46’, ’X6.4’, ’X28’, ’X2.1_2_3’, ’
                        X2.4_5_6’, ’X3.1_2’,
                    ’X6.1_2’, ’X13_14’, ’X40_41_42’)
input_cols_PCA <− c(’X1.1’, ’X2.16’, ’X3.1_2’, ’X4.1’, ’X13_14’, ’X46’, ’
    INPUT1’, ’INPUT2’, ’INPUT3’)
output_cols <− c(’E.1’, ’E.2’, ’E.3’, ’E.4’, ’E.5’)
output_cols_PCA_shrink <− c(’E.3’, ’OUTPUT_SHRINK’)
output_cols_PCA_full <− c(’OUTPUT_FULL’)

# INPUTS / OUTPUTS
input <− as.matrix(data[, input_cols])
```

```
input_PCA <− as.matrix(data[, input_cols_PCA])
output <− as.matrix(data[, output_cols])
output_PCA_shrink <− as.matrix(data[, output_cols_PCA_shrink])
output_PCA_full <− as.matrix(data[, output_cols_PCA_full], ncol = 1)

# FRONTIER PLOT
plot.new()
png(paste0(thesis_url, ’Media/Benchmarking/DEA_Frontier.png’), width
   = 600, height = 600)
dea.plot.frontier(input, output, txt = as.matrix(uni, col = 1), GRID =
   TRUE, xlab = ”Summed_Inputs”, ylab = ”Summed_Outputs”)
dea.plot.frontier(input[2, ], input[4, ], txt = as.matrix(uni, col = 1),
   GRID = TRUE, xlab = ”X1”, ylab = ”X2”)
dev.off()

# DEA WITH SINGLE OUTPUT
dea_model_single <− dea(input, output_PCA_full, RTS = ”VRS”,
   ORIENTATION = ”IN”)
summary(dea_model_single)
dea_scores_single <− eff(dea_model_single)
dea_uni_single <− cbind(uni, dea_scores_single)
dea_uni_single
plot(sort(dea_scores_single))

# DEA WITH TWO OUTPUTS
dea_model_single <− dea(input, output_PCA_shrink, RTS = ”VRS”,
   ORIENTATION = ”IN”)
summary(dea_model_single)
dea_scores_single <− eff(dea_model_single)
dea_uni_single <− cbind(uni, dea_scores_single)
dea_uni_single
plot(sort(dea_scores_single))

# DEA WITH MULTIPLE OUTPUTS
```

```
dea_model_multiple <− dea(input, output, RTS = ”VRS”,
    ORIENTATION = ”IN”, digits = 9)
summary(dea_model_multiple)
dea_scores_multiple <− eff(dea_model_multiple)
dea_uni_multiple <− cbind(uni, dea_scores_multiple)
dea_uni_multiple
plot(sort(dea_scores_multiple))

# DEA WITH MULTIPLE OUTPUTS ADDITIVE
dea_mult_add <− dea.add(input, output)
summary(dea_mult_add)

# SFA WITH SINGLE OUTPUT | FINAL
sfa_model_multiple_full <− sfa(input_PCA, output_PCA_full)
summary(sfa_model_multiple_full)
sfa_scores_multiple_full <− eff(sfa_model_multiple_full)
sfa_uni_multiple_full <− data.frame(University = uni,
                                    Score = round(sfa_scores_multiple_
                                        full, 4) )
sfa_uni_multiple_full <− sfa_uni_multiple_full[order(sfa_uni_multiple_full$
    Score), ]
rownames(sfa_uni_multiple_full) <− NULL

png(paste0(thesis_url, ’Media/Benchmarking/SFA_scores.png’), width =
    600, height = 600)
ggplot(sfa_uni_multiple_full, aes(reorder(University, Score), Score)) +
  geom_point() +
  labs(x = ”University”,
       y = ”Efficiency␣Score”)
dev.off()

# SFA WITH MULTIPLE OUTPUTS SHRINK
input_sfa <− input_PCA / input_PCA[, ’X1.1’]
input_sfa <− input_sfa[, c(2:ncol(input_sfa))]
input_sfa <− cbind(input_sfa, output_PCA_shrink)
input_sfa[input_sfa <= 0] <− 0.01

output_sfa <− as.matrix( input_PCA[, ’X1.1’], ncol = 1 )

sfa_model_multiple_shrink <− sfa(log(input_sfa), −log(output_sfa))
summary(sfa_model_multiple_shrink)
sfa_scores_multiple_shrink <− eff(sfa_model_multiple_shrink)
sfa_uni_multiple_shrink <− cbind(uni, sfa_scores_multiple_shrink)
sfa_uni_multiple_shrink
plot(sort(sfa_scores_multiple_shrink))

# DIFF B/W SINGLE & MULTIPLE
cbind(sfa_scores_multiple_full, sfa_scores_multiple_shrink)

Sys.setlocale(category = ”LC_ALL”, locale = ”UTF−8”)

library(ggbiplot)
library(data.table)

thesis_url <− ’˜/Google_Drive///Research/Programs/’
data <− read.csv( paste0(thesis_url, ’Data/Russian_Universities.csv’),
    header=TRUE, sep=”,”, dec=”.”)
data <− na.omit(data)
head(data)

uni <− data$University
input <− c(’X1.1’, ’X1’, ’X9’, ’X2.7’, ’X2.16’, ’X3.8’, ’X3.9’, ’X35’, ’X4.1’,
    ’X4.3’, ’X48’, ’X5.1’, ’X5.6’, ’X46’, ’X6.4’, ’X28’, ’X2.1_2_3’, ’X2.4_5_6’, ’
    X3.1_2’, ’X6.1_2’, ’X13_14’, ’X40_41_42’)

input_positive_cor <− c(’X2.7’, ’X5.1’, ’X6.1_2’, ’X6.4’, ’X35’, ’X40_41_42’
    , ’X48’)
input_negative_cor <− c(’X1’, ’X2.1_2_3’, ’X2.4_5_6’, ’X3.8’, ’X3.9’, ’X4.3’
    , ’X5.6’, ’X9’, ’X28’)
output_cor <− c(’E.1’, ’E.2’, ’E.4’, ’E.5’)
output_cor_all <− c(’E.1’, ’E.2’, ’E.3’, ’E.4’, ’E.5’)

# PCA Input Positive
pca_x_pos <− prcomp(data[, input_positive_cor])
summary(pca_x_pos)

png(paste0(thesis_url, ’Media/PCA/Input_Positive_Variances.png’), width
    = 500, height = 500)
screeplot(pca_x_pos,
            type = ”l”,
            npcs = length(input_positive_cor),
            main = paste0(”Screeplot␣of␣the␣PC␣variances”) )
dev.off()

png(paste0(thesis_url, ’Media/PCA/Input_Positive_Biplot.png’), width =
    500, height = 500)
ggbiplot(pca_x_pos, labels = uni)
dev.off()

# PCA Input Negative
pca_x_neg <− prcomp(data[, input_negative_cor])
summary(pca_x_neg)

png(paste0(thesis_url, ’Media/PCA/Input_Negative_Variances.png’), width
    = 500, height = 500)
screeplot(pca_x_neg,
            type = ”l”,
            npcs = length(input_negative_cor),
            main = paste0(”Screeplot␣of␣the␣PC␣variances”) )
dev.off()

png(paste0(thesis_url, ’Media/PCA/Input_Negative_Biplot.png’), width =
    500, height = 500)
ggbiplot(pca_x_neg, labels = uni)
dev.off()

# PCA Output
pca_out <- prcomp(data[, output_cor])
summary(pca_out)

png(paste0(thesis_url, ’Media/PCA/Output_Variances.png’), width = 500,
     height = 500)
screeplot(pca_out,
            type = ”l”,
            npcs = length(output_cor),
            main = paste0(”Screeplot␣of␣the␣PC␣variances”) )
dev.off()

png(paste0(thesis_url, ’Media/PCA/Output_Biplot.png’), width = 500,
    height = 500)
ggbiplot(pca_out, labels = uni)
dev.off()

# PCA Output All
pca_out <- prcomp(data[, output_cor_all])
summary(pca_out)

png(paste0(thesis_url, ’Media/Benchmarking/Output_Variances.png’),
    width = 500, height = 500)
screeplot(pca_out,
            type = ”l”,
            npcs = length(output_cor_all),
            main = paste0(”Screeplot␣of␣the␣PC␣variances”) )
dev.off()

png(paste0(thesis_url, ’Media/Benchmarking/Output_Biplot.png’), width
    = 500, height = 500)
ggbiplot(pca_out, labels = uni)
dev.off()
```