Journal of Applied and Theoretical Social Sciences

ICCSOR

ISSN:2687-5861

Using MLT to Estimate Corruption Patterns from The BEEPS

Alexander Shemetev[1]

Abstract

The paper describes the opportunity for using machine learning technologies (MLT) for estimating corruption by clustering. We used the enhanced BEEPS data (The Business Environment and Enterprise Performance Survey (European Bank for Reconstruction and Development, 2014)). It contains 1672 variables and 59619 observations produced by well-respected agencies like Nielsen for the European Bank of Reconstruction and Development. The analysis of different indicators with the MLT allows us to cluster the countries by the types of potential corruption patterns. We suggested this method could overcome the shortcomings of the classical survey surveillance approach because we can estimate countries with some distortion or insufficiencies in the data (for example, when the business units may want to lie about the corruption due to some reasons). This gives us an additional measurement that can be used for analyzing the true corruption field. This can be useful for business units, scientific people, and policymakers for analyzing the patterns of corruption in different countries.

Keywords: Corruption, Cluster Analysis Of Corruption, Corruption Patterns, Eurasian Countries, Corruption Index

JEL Codes: D73, D79, R00

[1] Associate Professor at the Department of Public and Municipal Administration (and Management), the Russian Academy of National Economy and Public Administration (RANEPA), shemetev-aa@ranepa.ru

Journal of Applied and Theoretical Social Sciences

ICCSOR

ISSN:2687-5861

BEEPS’ten Yolsuzluk Modellerini Tahmin Etmek Üzere MLT Kullanımı

Özet

Makale, kümeleme yoluyla yolsuzluğu tahmin etmek için makine öğrenimi teknolojilerini (MLT) kullanma fırsatlarını anlatmaktadır. Makalede, Avrupa İmar ve Kalkınma Bankasının 2014 yılına ait Gelişmiş İş Ortamı ve İşletme Performansı Araştırması (BEEPS)verileri kullanıldı. Avrupa İmar ve Kalkınma Bankası için Nielsen gibi saygın kurumlar tarafından üretilen 1672 değişken ve 59619 gözlemden faydalanıldı. MLT ile farklı göstergelerin analizi, ülkeleri potansiyel yolsuzluk modellerine göre kümelememize olanak tanımaktadır. Bu yöntemin klasik alan araştırması yaklaşımının gözlem eksikliklerinin üstesinden gelebileceğini ifade ettik çünkü bu yöntemle verilerinde bazı çarpıklıklar veya yetersizlikler bulunan ülkeleri de tahmin edebilmemiz mümkün hale geldi. (örneğin, firmaların bazı nedenlerden dolayı yolsuzluk hakkında yalan söylemek isteyebileceği zamanlarda). Bu yöntem bize gerçek yolsuzluk alanını analiz etmek için kullanılabilecek ek bir ölçüm sağlamaktadır. Ulaştığımız sonuçlar, farklı ülkelerdeki yolsuzluk modellerini analiz etmek için firmalar, bilim adamları ve politika yapıcılar için yararlı olabilir.

Anahtar Kelimeler: Yolsuzluk, Yolsuzluğun Küme Analizi, Yolsuzluk Modelleri, Avrasya Ülkeleri, Yolsuzluk Endeksi

JEL Kodlar: D73, D79, R00

Introduction

Corruption is a very interesting topic for research. Some researchers say that corruption is something like a sort of business (Lindgreen & Lindgreen, 2004). Other researchers say that corruption is nothing like a business (for example, Катц, 2020; Sallaberry et al., 2020). They say that, unlike business, even if a corrupt official earns the same amount of money, - it may bring the negative effects up to 10 times higher than the profits that the corrupt official earned (Čábelková and Hanousek, 2004; Hanousek and Kochanova, 2016; Катц, 2020). Thus, even if corrupt officials and entrepreneurs may have the same houses built with comparable amounts of money, - the total impact on the economy may be dramatically different. Entrepreneur earns money by providing useful service commodities or some work. This creates a win-win situation for the economy, for the entrepreneur, and for society as a whole.

Our research question is if we can find any corruption patterns with the use of machine-learning technologies (MLT) for clustering by comparing different countries. Answering this question will allow us to answer the related research questions: what is corruption and what factors influence it the most; which countries surveyed by the BEEPS are the most corrupt?

Our main hypothesis is that there is a significant chance to see the corruption pattern within specific economies by using the MLT for clustering: even if the data from a specific country look biased or insufficient – we can still evaluate its corruption pattern by comparing the cluster this country is referred to. This hypothesis leads us for checking the related hypotheses: firm size should form one of the 2 corruption types – small firms "grease their wheels", while big firms use corruption to beat their competitors and take advantage of the market mostly. It is assumed the entire BEEPS firms and countries contain 3 to 4 clusters, – the optimal number of clusters is estimated by using the MLT. Our second related hypothesis is that clustering by corruption type is possible from usually limited input data we usually have and the limited trust factor of this input data. We argue the classical methods cannot deal with the limited and distorted input data that is a common case even in the qualitative databases about corruption like BEEPS. We propose a method of MLT for clustering the countries by corruption types and estimating the main factors of corruption.

Answering the research question and testing our main hypothesis is our main purpose.

Literature Review

There are many questions on how to estimate corruption. Corruption is something hidden, something that corrupt officials do not want to be obvious to the other people. Thus, we face the limitation of the input data and it can be quite difficult to estimate the true level of that corruption any country has (Mauro, 1995).Corruption can be a threat even on the macroeconomic level. There results from the reforms can be dramatically different if corruption persists(Babecký and Campos, 2011; Babecky and Havranek, 2014).Since many regulations of the modern economy are based on the DSGE (dynamic-stochastic general equilibrium) models, corruption may change the results of applications of this model. DSGE are models that allow us to answer a very narrow question due to the huge amount of input data. They are used by central banks and analytical agencies in different countries to model economic processes. DSGE models do not initially consider the possibility that a part of the economy may be hidden by a corruption component. The corruption component creates errors for these models at the input. As a result of these errors, we get an increase in the error of the analysis result of this model at the output. Therefore, the DSGE model may become inapplicable because of an incorrect analysis of the corruption component in the economy. Hence, the economy can become more complicated to predict and more difficult to manage (Slobodyan and Wouters, "Learning in an Estimated Medium-Scale DSGE Model", 2012; Slobodyan and Wouters, "Learning in a Medium-Scale Dsge Model with Expectations Based on Small Forecasting Models", 2012).

There is a concept of bureaucratic harassment (Kaufmann, 1997). There exists some very complicated bureaucratic procedure. It is so complicated that most entrepreneurs do not know how to overcome it without corruption. This type of corruption we could call as "grease the wheels"(Méon and Weill, 2010; Méon and Sekkat, 2005, p. 70). Undertakers pay to corrupt officials some bribes or some sums (or create some other benefits) in order for the business to go smoother. This looks like a business deal. Entrepreneurs pay money and receive some advantages from the government.

Another type of corruption is based on legal harassment(Kaufmann and Vicente, 2011; Jain, 2001). This is the most devastating form of corruption (Thompson, 2018). It maximizes the losses of business and society (Silver and Rand, 1978). We call it the corruption of the second type in our search. Undertakers try to beat their competitors with the corruption mechanisms. Entrepreneurs pay in order to receive some superior advantages over their competitors. This means that the entrepreneurs receive some superior competitive advantage. For example, this could be some right to build a supermarket at a certain place. Alternatively, this could be some specific license for selling some specific commodities right here and right now. Moreover, it can be some payment to control for participating in governmental purchases. This payment is usually made in order to make competitors off the market with the help of the government. This is not necessary to be actually a money payment. This can be, for example, sharing a specific part of stocks or of some obligations. This could be some presents to some corrupt officials like real estate or expensive jewelry or benefits for the business of the family members of the corrupt official. This type of corruption goes off the free market principles(Smith, 1776).

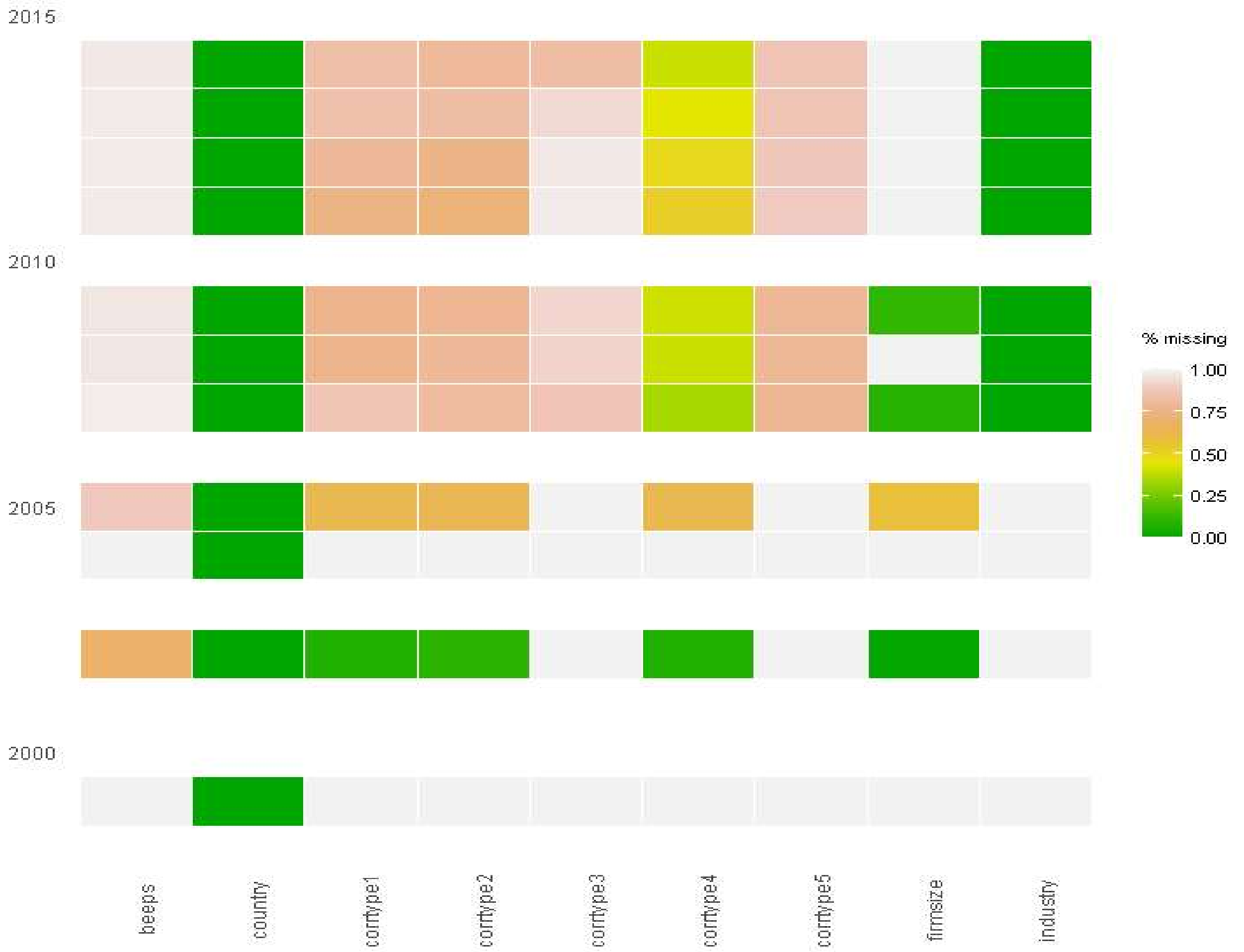
There is a multiple effect of corruption on the economy. It is proven that corruption makes the unofficial economy expanding. Corrupt officials cannot spend money as normal undertakers. They have to hide their wealth. Moreover, the gray market asks for the government for help in exchange for bribery. This creates this positive correlation between the core option and unofficial economy growth(World Bank, "Helping Countries Combat Corruption", 1997; World Bank, "Helping Countries Combat Corruption: The Role of the World Bank", 1997).In addition to this, corruption increases the tax burden on firms. Undertakers who use bribery, they escape partial losses in taxes, while the fair firms, at the same time, have to pay more taxes and, thus, their tax burden increases(World Bank, "Helping Countries Combat Corruption", 1997; World Bank, "Helping Countries Combat Corruption: The Role of the World Bank", 1997).

These creates unfair environment. It makes the fair agents to pay more and to suffer more losses, while it makes unfair agents to benefit and suffer fewer losses. It may create additional motivation for more firms to become corrupted. Moreover, this creates more incentives for the public officials to accept bribes and to participate in corruption. Fair public officials will suffer smaller incomes, while unfair public officials will enjoy additional incomes. This creates incentives for the public officials to deviate from a normal behavior and participate in different corruptions schemes. Thus, corruption creates more corruption. It creates penalty for fair agents and benefits for unfair ones.

Methodology and Data

We used the BEEPS (The Business Environment and Enterprise Performance Survey (European Bank for Reconstruction and Development, 2014)) data. It contains 1672 variables and 59619 observations. A high number of variables should help at selecting the proper parameters for better clustering. It is quite difficult to obtain any good data about corruption. We used the BEEPS survey. This is an extensive economic survey undertaken as a joint initiative of the World Bank and the European Bank for Reconstruction and Development. The survey provides data for 37 countries from 1999 to 2014. Almost all the countries belong to the Eurasian area. The survey consists of 17 sections none of which is directly devoted to corruption (topics like infrastructure, sales, competition, ...).

Figure 1. Descriptive Statistics Representing The Availability Of The BEEPS Data



- Note to chart 1: BEEPS – is a supportive variable related to the expected quality value of the original data devoted directly to corruption estimation (modelled estimation); the lacking data was modelled with the use of machine learning technologies and then compared with the original pattern for not having contradictions. For the firm size in the modern research waves we compared the data with the one received from the Russian statistical office (for modelling and making better expectations about the missing values in the research).

We have created a special code in the R programming language that would allow us to display the missing data in the databases like BEEPS. We named the R package "alexandershemetev" (Shemetev, 2020). The "alex_na_plot" function creates the plot like fig. 1 that shows green for years and indicators for which we have most of the data available. Indicators for which we have about half of the data are shown in yellow. Red color shows indicators for which we have approximately 25% of the data. The gray color is typical for the data for which we have the order of 0 to 10% of the data available. We collected the 1672 parameters into 6 large groups named "corrtype" and labeled them with numbers from 1 to 6 (it is related to the types of corruption some certain data could be applied for). We also entered data on the size of firms and industries. We created the generic BEEPS indicator that measures the overall quality of the data for assessing corruption using standard methods. We see that for almost all years we have very little sufficient data to use classical methods for estimating corruption in different countries, even within the framework of such qualitative research as the BEEPS. Thus, we assumed that the use of machine learning technology and the cluster approach should help us fill this gap in the original data. This approach allows us to conduct the analysis more qualitatively than if we would use the classical methods of interview analysis.

The paper suggests that if an analyst uses classical methods for analyzing corruption, he will face the problem that he will lack input data as seen in Figure 1. It is believed that the BEEPS study is one of the highest quality studies on corruption data. It is conducted by leading analytical agencies for the European Bank for Reconstruction and Development and the World Bank. Therefore, it is believed that databases like BEEPS are like a treasure trove of useful information about corruption in different sectors in different countries. This method of data analysis is suggested in this paper. It allows analyzing databases, such as BEEPS, which allows a better assessment of the pattern of corruption. Machine learning is used to match each country as accurately as possible to each individual cluster. The results of this cluster analysis are quite logical. The suggested method will make it possible to assess corruption even by the missing input data. This is the advantage of our proposed method.

Table 1. Descriptive Statistics Representing the BEEPS Data (As Taken From The BEEPS)

	Descriptive Statistics							
Indicator	**Variable Type**	**N obs.**	**Mean**	**Min.**	**25 %**	**Median**	**75 %**	**Max.**
BEEPS	Coded factor	4,090	0.980	0	1	1	1	1
year	Numeric	59,619	2007	1999	2005	2008	2009	2014
corrtype1	Coded factor	16,854	1.350	0*	1	2	2	2
corrtype2	Coded factor	16,577	1.036	0*	0	0	0	1
corrtype3	Coded factor	2,697	0.857	0*	2	2	2	2
corrtype4	Coded factor	30,176	1.286	0*	1	2	2	2
corrtype5	Coded factor	6,287	0.788	0*	1	2	2	2
corrtype6	Coded factor	59,619	0.067	0	0	0	0	1
Firm size	Coded factor	21,516	1.776	1	1	2	2	3
Industry	Coded factor	35,255	38.236	2	24	45	52	72

Notes to the table 1:

✓ “0*” is a numerical representation of the code “-9” from the BEEPS which means people denied to answer the question. This “0*” is not a numerical zero – this mostly is a factor variable that catches the people denied to reply for the certain questions related to a specific corruption type. Since the replies come as factors (except those directly related to a numerical scaling), we cared them properly over the analysis. The descriptive statistics for the coded factors is given to highlight the distribution of the coded variables (how often there can be met a specific factor in the dataset).

- ✓ Year row interpretation: the accurate mean value is 2007.129 – means year 2007 plus 12.9% of the time of year (late Winter- early Spring 2007); Standard deviation is 3.996; it means the standard deviation from the mean year 2007 is about 4 years.
- ✓ BEEPS – overall indicator performing the cases of corruption when firms receive additional benefits (0-1 estimate);
- ✓ year – specific years of the survey;
- ✓ **corrtype 1 to 6** – represent the different questions related to corruption in the basic BEEPS survey with encoded answers. The main questions that we paid attention to are represented below. We used most of the variables to control for the correctness of the chosen proxies for the MLT. We used all variables included in these categories separately as they are; we just used these macro-categories to make a simpler representation of the analysis and output.
- ✓ **Corruption type 1 (corrtype1)** is payments for implementing the basic needs for the functioning of any business (the main proxy is taken is attaching to electric supply without which no business (with rare exceptions) may run in the country. The main question here sounds like this: "informal gift/payment expected or requested for an electrical connection?", which can also be controlled by questions like: "how much of an obstacle is electricity to the current operations of this firm?". We consider electric supply to be a natural monopoly connected with government regulations – if bribery is high within the country, - asking to pay for electricity connection will be on the priority list. The paper considers that almost no business can operate without electricity. Electricity is something most firms need to receive from a natural monopoly (and relatively rarely from other sources). If government corruption exists, - the government allows accepting bribes by its controlled natural monopolies (like electric supplies). The government may also prevent any competitors of entering the electric markets (like alternative energy producers as competitors to the natural monopolies). Firm managers will more fair tell about difficulties in getting electricity than about direct bribery of officials. The research considers electric supply difficulties are something that everybody knows in the country and so managers will speak more freely. That is why this qualitative variable is a proxy for the real 1-st type of corruption in the country. This variable will correlate similar to the real index of the corruption of the 1-st type, but the input data will be fairer. "0*" means a person denied to reply to this question. "0" – means a person said no. "1" means a person confessed to some indirect clue of corruption (like saying he did not give any bribes as a top-manager or business owner, but telling, for example, he receives 50% of electricity from an owned generator and faces obstacles for operating in the business). This type 1 corruption means the government is unable to remove the obstacles from the natural monopolies by having an inefficient control system that should heathen the good background for the corruption of the 1st type. "2" means there are many shreds of evidence from the top-managers and business owners (respondents of the BEEPS) who directly confess the corruption of the 1st type exists in their country.
- ✓ **Corruption type 2 (corrtype2)** is related to licensing and obtaining the necessary legal documents to function in this country officially. Licensing is taken as a proxy for this corruption type estimation. We would be happy to see a fair reply to the main

question in this group: “unofficial payments to get licenses and permits”, “frequency of unofficial payments/gifts to obtain business licenses and permits” or “when you applied for an operating license was an informal gift requested?”. We understand that people, maybe, not willing to reply fairly to such questions. That is why we do not consider this question as a good proxy for the corruption of the 2nd type for all countries, but just as an important variable. In some countries, people speak more “yes” to these questions, but in some countries, people may reply “no” but hiding the true answer. We could try to establish it from the related questions like: “application to obtain an operating license submitted over last 2 years?” or “obstacle: business licensing and permits (yes/no)”. People will be more freely saying about the true term to get a license, than directly answering if they gave gifts to public officials. One can compare the variation. If it is high, say, for one firms and low for others, this may mean that some firms might use corruption for obtaining permits for operating. That is why a qualitative total output is used. “0*” means people (top-managers and business owners) denied to answer the set questions; “1” means highest sure there is a corruption of the 2-nd type judging by the answers to the questions. We also had a catch-question “change in total costs if business licensing and permits no longer an obstacle (yes/no/moderately/don’t know/deny to reply)”. If a business pays bribes for licenses but denies saying freely about this, it will, most probably, answer freely if an increase in the official cost of getting a license may become an obstacle. We suggested the sum of the bribe might be incomparable more to license cost; hence, business owners may lose their attention at the price of the license (our assumption). So, if a business owner pays bribes much higher than the license cost – it might not mention the official price of the license as substantial (we have special questions if a firm needs any license at all). We considered it as one of the good clustering questions. We did not put any regression coefficients or analyzing the scales like in the classical approach; we used MLT clustering to judge by answers if a country is within a specific cluster. This gives a huge advantage to the classical approach – we do not interpret or modify the data – we just cluster it with MLT. Thus, this can manage to get sibling-like countries. In addition, it will be sufficient to estimate properly just one from the entire scope (by concentrating all the data we have) to say about the entire cluster (about all the similar countries). In our opinion, this gives us a big advantage in our approach. The values used in the descriptive table mean next: “0*” means a person denied to reply to this question; “0” – means a person said no; “1” means there are many pieces of evidence from the top-managers and business owners (respondents of the BEEPS) who directly confessed the corruption of the second type exists in their country. Other questions used as control variables for clustering.

✓ **Corruption type 3 (corrtype3)** is created to estimate what percentage of sales a firm should pay to secure its normal functioning. Through the reason we have a limited number of “confessions”, we had to work only with the data when we are sure about the percentage of the contract. A set of questions is used:

1) “% of unofficial payments to get licenses and permits”;
2) “% of unofficial payments to get connected to public services”;
3) “% of unofficial payments to get licenses and permits”;
4) “% of unofficial payments to deal with taxes and tax collection”;

5) “% of unofficial payments to gain government contracts”;
6) “% of unofficial payments when dealing with customs/imports”;
7) “% of unofficial payments when dealing with courts”;
8) “% of unofficial payments when dealing with health/fire inspections”;
9) “% of unofficial payments to influence the content of new laws decrees”;
10) “% of unofficial payments for other things”.

For people who denied answering a question we used “0*”; “0” means people replied no; “1” means there are payments but they are relatively low (comparing to other countries); “2” – other (means firms pay grafts for functioning and often confess it). This corruption type means a business has to go to long-run relations with the public officials to function and pay a certain percentage as a sort of “unofficial salary” to the decision-makers. This part concerns the business corruption for running companies only; it doesn’t cover such cases as, for example, paying % of salary to have a specific position in some company either to public officials or to some company management. We considered this type of corruption is the most difficult to evaluate. This could be a topic for further researches in the field of corruption.

✓ **Corruption type 4 (corrtype4)** is related to foreign companies who potentially may want to enter the local country market if they should be ready to face corruption for this. One of the proxies taken for this estimation is informal gifts and payments needed to obtain import licenses. The primary question was: “application to obtain an import license submitted over the last 2 years?”. If the reply is no – then we did not consider it as a company that may suffer this type of corruption, because either a company has no relation to imports or too much time left to recall the details. If the reply was yes, then we looked at such questions as:

1) “when you applied for an import license, was an informal gift requested?”;
2) “average number of days for imported goods to clear customs in last fiscal year”;
3) “unofficial payments or gifts: to deal with customs or imports”.

For people who denied answering this question we used “0*”; “0” means people replied no; “1” means there are payments but they are relatively low (comparing to other countries); “2” – other (means firms pay grafts for functioning and often confess it).

✓ **Corruption type 5 (corrtype5)** is related to the possibility to face bribery at normal functioning by inspection officers (like fire inspectors, health inspectors, labor security inspectors, tax inspectors) which can mean additional charges for having smooth business processes. The question sounds like: “in any of these inspections was a gift or informal requested?”. For people who denied answering this question we used “0*”; 0 means people replied no; 1 means there are payments but they are relatively low (comparing to other countries); 2 – other (means firms pay grafts for functioning and often confess it).

✓ **Corruption type 6 (corrtype6)** is a variable created to estimate the total intensity in dealing with public authorities in terms of corruption based on the scoring system of the expected level of corruption modeled in the survey. Connecting the time variable and corruption type we received the estimation of trends of corruption represented on charts in the presentation to this survey. We understand all types of corruption may be somehow interrelated and even country corruption may be interrelated too – better research data is needed to estimate these effects. This variable is a dummy that is 0 if no clues for estimating any types of corruption, 1 otherwise.

✓ Firm size – the size of the companies (small-medium-big);

✓ Industry – enlarged industry indicator representing the coded values for classified in this analysis industries (72 industries in total that were compressed to 10 macro-industries in the later analysis).

✓ N – number of observations with the related responses.

As we have factor variables in most cases with meaning more than 2 (so, the classical dummy approach is not a remedy), we cannot use the classical regression analysis. We suggest using the MLT clustering to estimate the corruption patterns better.

A couple of strange things in the data were found: All the respondents are either function at point populated with X or work at some sector: 2 surveys combined. Less than 52% of firms have the only owner; 88.8% of firms are domestic (not foreign); 1/3 of firm owners are females. It means, the BEEPS combines several surveys at once. It is better to cluster such data rather than analyze it with the standard methods. 15.27% of firms used corruption to receive electricity supply; Only 40% of firms had no significant problems with a power outage (no more than 2 hours). Corruption exists in every country in Survey: the correlation between the answers when people told they know and replied and the people who told they don't know is – 94%. Thus, in countries with potentially low corruption patterns, we see a greater number of people refusing to answer yes or no. This may mean the countries where we do not observe corruption – still can have a significantly higher level of corruption than can be directly estimated from the BEEPS survey.

Methodology

We see that on many parameters, even in such qualitative studies as BEEPS, we lack data. Consequently, using classical approaches to the analysis of corruption may not be applicable even for very high-quality data. Classical approaches involve the use of regression methods or a rating scale and their further interpretation. It is very difficult to interpret when there is any missing data. It may happen that data on different subjects of different types of corruption may be present for some parameters and absent for others. As a result, we can get incomparable data with the classical approach. This can reduce the quality of the analysis. It is proposed to use machine-learning technologies (MLT) to create clusters of corruption within the framework of different factors and types of corruption. It is expected this approach to be beneficial. It will be enough to split the data into clusters.

At first, it was assessed the data of the survey. There were many variables and many contradictory data in the survey. Even countries were not always easy to identify. For example, country code 46 represents Sweden which has a number of observations, although, it should not be in the database, according to the instructions provided to the data. Some data represent unique case studies of only a few countries. For example, there was a survey devoted to Ukraine and Uzbekistan only. Secondly, there were noticed indicators that could be useful in estimating the potential corruption levels in the countries represented in the survey. It was paid attention that there could be separated into 2 types of corruption. The first type of corruption makes business processes more smooth, for example, firms can provide a gift to receive a better electricity supply or better phone connection (such data can be found in the survey). The second type of corruption should give advantages to the business units, like, for example, receiving a governmental grant. There is a common pattern of the survey. The data is devoted to developing the post-communist world (circa 98% of the data) and some closely related countries (circa 2% of the dataset). We can divide all the countries into several regions:

✓ region related to the post-communist countries;

- ✓ Region Europe (it has some overlaps with the previous region);
- ✓ region Vietnam,
- ✓ region South Korea
- ✓ strange regions (Sweden persists in the country codes 46; Switzerland persists as a country code 41 – both countries are unusual for the BEEPS survey and, maybe, data errors or input errors into the primary data).

This division may show us if there are some regional patterns in the data. For example, in Turkey "baksheesh" (kickback) is not related to a bribe; in Korea, gifts may be considered a normal phenomenon. Some geographic regions can have a unique way of development that can show different results in the functioning of the firms.

The three main methods used in this analysis are descriptive statistics, which summarize data from the BEEPS survey using indexes such as means, correlations, standard deviations, and inferential statistics, which draws conclusions from BEEPS data that are subject to random variation (e.g., observational errors (like with data from „Sweden"); and methods related with machine learning techniques for classification of the BEEPS panel data called hcluster [industrial standard method for MLT]).

We see that the most number of observations in the survey is devoted to Russia and some other populated countries; the correlation between the number of observations in the survey and the country population is 91%; however, if we estimate the number of population in the country divided by the number of observations in the BEEPS survey, we will find, that, for example, Russia is double underestimated comparing to the other countries (less number of observations per limited number of population).

The BEEPS data was joined with the 5 databases to estimate the main parameters (source: DBNomics and national statistical offices). We were interested in economic indicators like population, inflation, GDP, GDP deflator, and currency exchange rate changes. These data were necessary to compare the BEEPS data with the macroeconomic data we receive to compare the results from the BEEPS questioning if they do not contradict the macroeconomic data. We found no big contradictions with the macroeconomic indicators of what the respondents were answering. Moreover, we could visually compare the corruption pattern and the macroeconomic indicators like output; the quantitative estimation of them can be a topic for future research.

We suggest MLT k-means and hclusters are sufficient for estimating the different patterns of corruption. They are sufficiently effective for this job. The MLT for clustering uses dots to join the data. Usually, the first step is putting a random dot at the multidimensional field. The number of dimensions is connected with the number of parameters we put into the model. Then the model assigns the second and third multidimensional dots and estimates the potential distance between them. The model continues to put the dots unless no further changes are optimal. MLT gives a huge contribution in minimizing residuals, hence, maximizing the correctness and efficiency of the model. Changes are optimal only when we can put an observational dot so that the distance between the dots inside some cluster is the nearest; at the same time, each dot from each cluster should be as far as possible from other clusters (otherwise additional clustering still makes sense and the model will continue processing the data). Thus, we find the 3 or 4 clusters (optimal for our data) that are maximum different from each other and maximum close to each other. If, for example, we do not have sufficient data for a country named "X" – we can say (judging by the survey responses we have) that it is maximum close to country Z where we have sufficient data. Hence, we can say that the corruption pattern in country "X" is, most probably, similar to the corruption pattern of country Z; at the same time, both X and Z are maximally distant from any countries from different clusters and maximally close to the countries within their cluster. Some data for some countries may look weird. For example, people deny to reply they pay bribes, but additional questions reveal there should be a good basis for corruption. Therefore, such weird patterned countries will, most probably, turn to another cluster that will collect all such weird patterns and allow the researchers to pay more attention to them. We suggest MLT k-means and MLT hclusters methods to be sufficient to spread the countries by groups.

Main Strategy Used

There are performed 3 types of analysis.

1) Analysis of answers where people replied they know about corruption (without mentioning the sums and so on); this was used as a proxy for greasing the wheels corruption – when companies use corruptions to make their business smoother;

2) Analysis of answers where firms replied they gave a certain percentage of the contract to receive the contract – it was used as a proxy for having an advantage corruption – when companies use corruption to beat all their competitors and receive important contracts and deals;

3) Machine learning cluster analysis of corruption (both types of corruption) by the industry patterns, company size patterns (Shemetev, 2012), and corruption types (among these two mentioned) patterns. The results of all the 3 methods were very similar. The results of cluster analysis revealed: Corruption is mostly related to Corporate Size: Neither industry nor geography has a strong impact on participating in corruption as corporate size. 2 previous types of analysis were used to check if MLT output has nothing strange or unusual at these or that assumptions.

Empirical Results and Discussion

Thus, we analyzed the population in each country, GDP, GDP per capita; data from local statistical offices were used to check if the data presented in the BEEPS corresponds with the national statistical offices' data. This data is used to verify the data patterns in the BEEPS survey: if the data we see in the BEEPS survey correlates with the statistics taken from other sources. For example, Russia, as a country with the maximum number of observations in the survey, is chosen as one of the most important sample countries to compare. The results coincided within 20% confidence intervals between the data in BEEPS and local statistical offices. This reveals the BEEPS survey is more or less reliable for analysis.

Figure2. Corruption Trends Revealed From The BEEPS Data

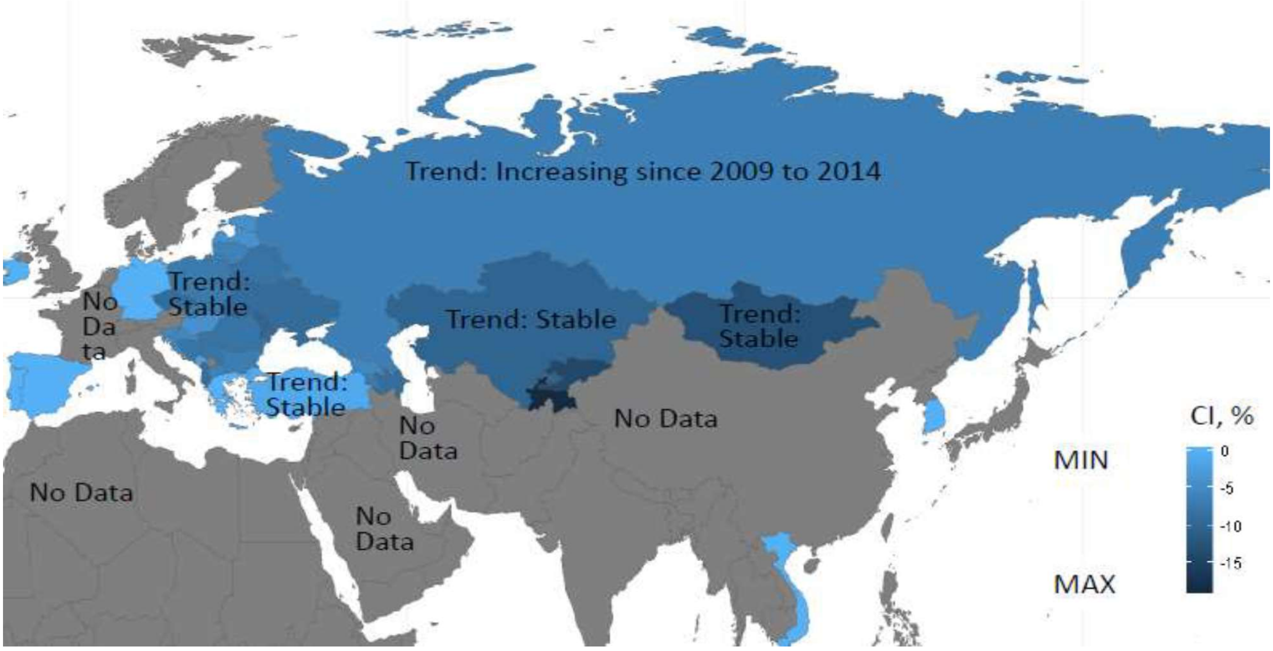


Figure 2 is created by the author in the statistical programming language R. CI is the corruption index on the map representing the total percentage of cases where corruption was mentioned; it has a negative ("-") sign on a scale as a proxy for the penalty for the output. The closer this index to 0 – the better it is for the economy. The formulas for estimating the index are stated below:

$Y*(1+ (CI))^{t}=rGDP$ (1),

$CI_{i} = -100*\Sigma C_{ij}/n_{i}$ (2)

Notes:

i – country or region within a country;

j – a specific observation with complex questioning;

Y is an effective output (like effective GDP that could potentially be produced if there would be minimum corruption),

CI – is the corruption index that is negative in our case. For example, if effective GDP is 100 currency units produced for a country – corruption index can reduce it to some number (like 10% (if CI = -10); 15% (if CI = -15)). CI should combine the effective output of a country with the real GDP produced by the country in reality,

C_{ij} – corruption control cluster for a country based on the replies to the most effective estimation questions that are represented below. We used MLT k-means to estimate it. It shows the correction term for a specific country for the corruption level (depending on the cluster the country is in). The more positive answers to any of the most important corruption questions we receive, – the higher will be the value. It is recommended to have data from top-managers of not less than 300 companies of different sectors and sizes for having objectivity. We estimated there is no sense to ask direct questions about the size of profit company pays – we estimated the represented below questions generate sufficient proxy for clustering for estimating the corruption penalty within the express-method;

t – period (1 year for the CI presented in figure 1),

n_i – the number of observations for a specific country,

rGDP – is the real GDP produced by the country (it is estimated by the national statistical offices and the international bodies like the World Bank or IMF). rGDP can be replaced with rGRP. This is the real gross regional product. If we estimate corruption on the level of regions if the data is available – CI then should be estimated for a region in the same way as for a specific country.

Calculating CI – is an express-method for getting a proxy for the estimation of the corruption (may be used k-means or hclust methods described above). It is important to follow the next algorithm to create this index for a country. The first step is creating a dummy variable for each observation that is equal to 1 if any of the corruption cases persists and is confirmed by the positive answer of top-manager or business-owner (the list of express-method questions is represented below). This can be any positive answer by any type of corruption question for a specific observation. The research suggested the next 5 questions to be the most effective and even sufficient for estimating the corruption for express-method:

- ✓ What % of senior management time was spent dealing with government regulations? Explanation: This is a control proxy – if time is too high, - this should cause suspicions.
- ✓ Was an informal gift or payment requested in any of these inspections?
- ✓ What percentage of contract value average firm pays in informal gifts to the government to secure a contract?
- ✓ When you applied for an operating license was an informal gift requested?
- ✓ Was an informal gift/payment expected or requested for a construction-related permit?

This method is quite easy and, at the same time, effective to estimate the corruption at the country-level, if questionaries are made by respected companies like Nielsen for a country. This method generates a proxy that estimates corruption losses from indexes. It can be used as an express method to estimate the countries for corruption patterns. This research revealed these questions are the most effective (out of 1672 questions) to estimate the corruption level. These questions may potentially be slightly modified. Then, even simple k-means in R with the simplest MLT package could perform this express analysis without losing much time and effort. At the same time, such an approach can easily save time and money at planning and performing researches about corruption. Any researcher can perform this questionary on condition if he will be able to represent his questioning at a similarly high level as it would do a respected international agency or company.

It is considered 3 to 4 to be an optimal k-value for dividing the clusters for the countries represented in the BEEPS survey for the clustering like k-means (we prove this hypothesis below). Researchers may use $\theta = 1$ to calibrate their model.

Our model performed the next summary statistics for the corruption penalty in terms of CI. Minimum is 0%; 1 quintile is -3.43%; median is -6.5%; mean -6.1%; 3 quintile -8.8%; max -18.73%. Our data is skewed in distribution. Generally, the countries interviewed in BEEPS more tend to suffer corruption patterns closer to minimum values rather than to maximum (the most popular values), although, we have some significant corruption outliers that shift the median away from the mean. We think corruption takes a part of the output as a penalty. The research suggests one may use this express method with express questions to estimate the approximate level of the corruption penalty within a region or a country. One only needs to guarantee the quality of questioning, including the need for the research to be representative.

Corruption index is developed in this research penalty for corruption that country or region has each period. Each year country loses part of the output on corruption. Corruption brings negative effects on the economy. If this is "grease wheels" corruption or "taking advantage" corruption, – this all grants an unfair advantage and turns the competition from the main market principles. This always is less effective than if the competition would be fair.

We think CI suggested in this research has an advantage over the international corruption indexes, because international indexes are made, to some extent, by expert conclusions and are like a black box for the external observer. Our index has a computation basis for estimation. Improvement of our CI can be a topic for future research.

The paper suggests a more complex cluster analysis using the majority of the questions from the BEEPS (not just a few like in the previous research for the most effective questions). It is important to notice the 4 cluster groups can be revealed in the BEEPS data using machine learning technologies. Clusters C1-C3 are ranged by the machine-learning clustering technology according to the score of the expected corruption that business uses. In "Minimal corruption cluster" (C1) – corruption is more a rare phenomenon and is expected to be usually used for making business operations smoother and, usually, bigger firms may have more deals with corruption in this cluster. "Heavy corruption cluster" (C3) reveals the second type of corruption is maximum often to be expected, like, using gifts for having contracts; we may expect that firms in this cluster pay up to 100% to grafts for having contracts and for running a business. "Intermediate Corruption Cluster" (C2) is an intermediate cluster between C1 and C3 that has signs of both types of corruption at a higher level than in C1 and lower level than in C3 countries; firms pay grafts, usually from 20 to 50% of the contracts. "Potentially high corruption cluster" (CX) is related to the countries where machine learning could not establish a clear pattern related to cluster C2 or C3. Although, the pattern is different than in C1 as well; therefore we may suggest in such countries corruption level can be potentially high, but human confession is relatively low that is why the BEEPS survey is insufficient to give a perfect picture in these countries. These countries cannot be directly categorized from the BEEPS data. The main significant factor uniting all the clusters is the company size – big companies more often use corruption of the second type to have an advantage, while small ones try to use it to have more smooth business processes (like having fewer losses from inspection). Interesting results can be seen from the cluster analysis: Georgian managers say they have a similar level of corruption to Germany; Turkey has a unique corporate corruption pattern that doesn't look similar to any other country (according to the replies). Vietnam, although provided little data can have similar cluster corporate corruption pictures with Macedonia; Romania; Kazakhstan; Azerbaijan; Armenia. Such results were received by seeing the output of the machine learning technologies for clustering.

Figure3. Corruption Clusters Revealed From The BEEPS Data

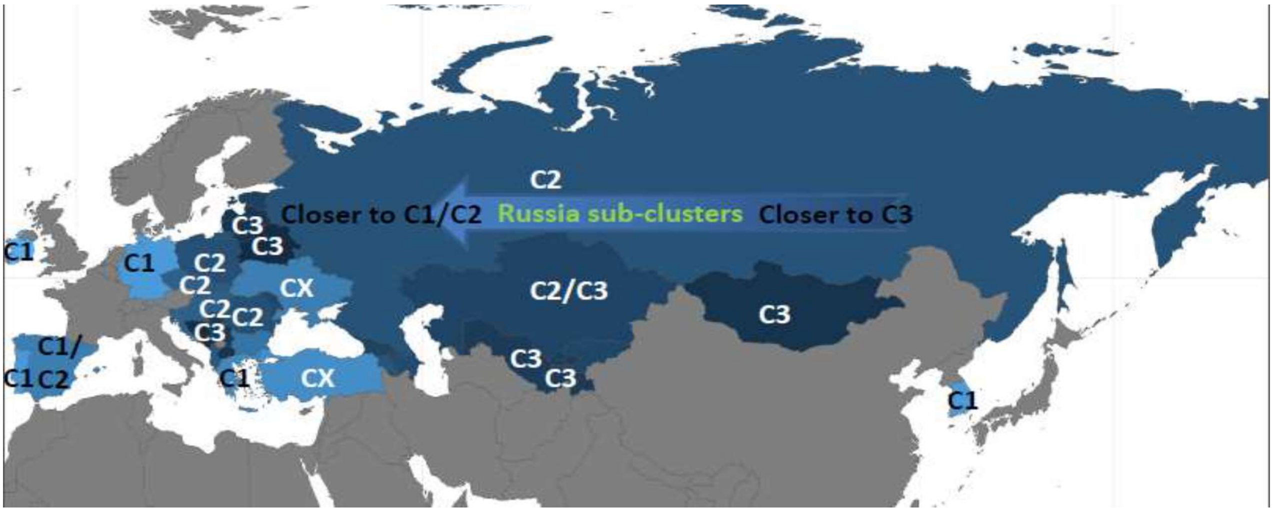


Figure 3 is an important key to understanding our research. We see corruption has 4 clusters and no direct geographical pattern. CX cluster is the cluster with the weird pattern. One can observe enhanced corruption patterns in the post-communist world. One might assume this could be the consequences of the regimes of the past. Most countries have positive patterns of corruption which means corruption decreases. It is analyzed internal corruption only, that is, corruption within each particular country or region (there performed a regional analysis for Russia that revealed the highest corruption patterns are closer to the central and eastern parts of the country in percentages, but not in absolute numbers). We understand that analyzing Russia by its regions could be a topic for future researches. We saw that higher levels of corruption correlate with lower levels of the GDP. We pretend the idea of the method for calculating the penalty for corruption for the output is new (compared to the other literature we studied).

Conclusion

We prove the categorization of corruption that partially corresponds with prominent researches. For example, corruption types 2 and 4 corresponds with a subtype of legal harassment (Jain, 2001; Kaufmann and Vicente, 2011); however, we split this type of corruption for foreign companies entering the local markets (type 4) and type 2 – for the domestic firms. We think such splitting provides more precise information for businesses. Foreign companies entering the local markets (or foreign investors) want to know how to be better prepared for the corruption they may potentially face entering some local market. Domestic companies would like to know corruption barriers from competitors abroad and barriers for their potential foreign investors they may face; type 4 would be important for them. Type 2 is important for local companies to understand the situation better. It becomes even more important if we are able to receive regional data. Companies may be interested to know what to be prepared to enter the markets in other regions of the same country. Thus, splitting corruption that makes difference with (Jain, 2001; Kaufmann and Vicente, 2011) creates practical application benefits, this study suggests.

In addition, some researchers (Thompson, 2018; Silver and Rand, 1978) suggested that corruption in taking advantage of the competitors is the most devastating form of corruption. It maximizes the losses of business and society. It is very difficult to estimate this form of corruption from the data, because entrepreneurs may tend to deviate from direct answering „yes“ to this question. They may not want to confess. They may want to preserve the competitive advantage they received at the price of corruption. We suggested a proxy that is easier to get and simpler to implement. Corruption type 1 (in addition to corruption types 2 and 4) is designed to catch this. We suggested if this form of corruption is widespread – it should touch the primary branch necessary for any type of business (electric supply). The easiest way to prevent competitors from entering the market is to prevent them from getting electricity. This guarantees the maximum advantage. Small businesses are less likely to be powerful enough to use this corruption type. Thus, they can reply fair to the questions about the difficulties with the electric supply with higher probabilities. Businesses of all sizes may contribute to answering these questions. Small business units may be fairer. Thus, this is a proxy obtained from a wider range of sources that make more opportunities to get a piece of qualitative information. If there is a problem with electric supply, there are higher probabilities there are difficulties with the type 1 corruption as well, this study suggests.

Corruption type 3 is another type of information. It could be so that the business pays interest from its revenue or profit to some public officials for having absolute benefits. We decided to split this „taking advantage“ into 4 types of corruption. And since „greasing the wheels“ and „taking advantage“ are not expected to be corner solutions (when businesses aim at only one purpose at once), at some measure these types of corruption may measure both of the mentioned in scientific literature types (legal and bureaucratic harassments). This creates additional new points in this research compared to the scientific literature.

Corruption type 5 is a suggested proxy for “greasing the wheels” corruption (Méon and Weill, 2010). Unlike the study of (Méon and Weill, 2010), we suggest using a proxy to have a better opportunity to view this type of business. Such an approach creates a practical application benefit, – it becomes easier to evaluate and estimate this type of corruption. Small businesses will more likely use this type of corruption and more likely answer these questions about the inspections they face.

Another interesting pattern is revealed that each company confesses in a maximum of only 1 type of corruption (0 cases when it is not true) in all countries in all periods in the survey. No firm paid, say, gift for more than 1 thing like if a company paid for electricity supply – it did not pay gift for anything else. It means BEEPS merged several questionaries in one dataset. It makes more benefit in applying our approach that can deal with such a pattern and still effectively evaluate corruption from the cluster analysis. The free software (programming language R) used to receive these results.

We see that big firms tend to use “taking advantage” of corruption, while small firms tend “to grease their wheels”. We suggest small firms are difficult to act together as a one team. Hence, each small firm has few opportunities to gain money for “taking advantage” through corruption. Big companies are more interesting for public officials. They tend to pay more attention to these types of businesses. Therefore, big businesses might have an incentive to deviate from fair market practices and apply corruption to bit their competitors and secure the market positions. Firm becomes closer to an oligopoly position at the market when securing the market position. A country may lose rather than win from corruption. This is visible from the concept of the penalty for corruption. Our suggested method for estimating the CI may be improved in future researches. It may become more precise.

Summary

There are several types of corruption. Two of them create two main patterns of corruption. Corruption of the first type can be called “greasing the wheels corruption”. It means that business participates in it to make some business processes smooth and make bureaucratic procedures easier.

Corruption of the second type is connected with the legal issues and has, thus, benefits over their competitors. Both types of corruption are devastating to the economy. Corruption creates incentives for fair agents (like politicians or firms) to become unfair ones. Bigger firms have higher incentives for participating in corruption procedures. On the other hand, unfair public officials are more interested in working with bigger companies because they can pay more. We can say that company size is the main factor that influences corruption.

References:

Babecký, Jan, and Nauro F. Campos (2011), Does Reform Work? An Econometric Survey of the Reform-Growth Puzzle. *Journal of Comparative Economics*, doi:10.1016/j.jce.2010.11.001.

Babecky, Jan, and Tomas Havranek (2014), Structural Reforms and Growth in Transition. *Economics of Transition*, doi:10.1111/ecot.12029.

Čábelková, Inna, and Jan Hanousek (2004). The Power of Negative Thinking: Corruption, Perception and Willingness to Bribe in Ukraine. *Applied Economics*, 2004, doi:10.1080/0003684041000167430З.

DBnomics Official Site. *Statistical Databases of the population and GDP*. Retrieved from URL: https://db.nomics.world/ (date of access: 22.12.2019)

Eurpoean Bank for Reconstruction and Development (2014*), Business Environment and Enterprise Performance Survey (BEEPS) [Dataset V]*.Retrieved from URL: https://www.ebrd.com/ (date of access: 22.12.2019)

Hanousek, Jan, and Anna Kochanova (2016). Bribery Environments and Firm Performance: Evidence from CEE Countries. *European Journal of Political Economy*, 2016, doi:10.1016/j.ejpoleco.2016.02.002.

Jain, Arvind K. (2001), Corruption: A Review., *Journal of Economic Surveys*, doi:10.1111/1467-6419.00133.

Катц, М. (2020). Последствия коррупции: почему это касается каждого retrieved from: https://youtu.be/uyiudptWAAY (date of access: 2020/12/22)

Kaufmann, Daniel (1997), Corruption: The Facts. *Foreign Policy*, doi:10.2307/1149337.

Kaufmann, Daniel, and Pedro C. Vicente (2011), Legal Corruption. *Economics and Politics*, doi:10.1111/j.1468-0343.2010.00377.x.

Lindgreen, A., & Lindgreen, A. (2004). Corruption and unethical behavior: report on a set of Danish guidelines. *Journal of Business Ethics*. https://doi.org/10.1023/B

Mauro, Paolo (1995), Corruption and Growth. *Quarterly Journal of Economics*, doi:10.2307/2946696.

Méon, Pierre Guillaume, and Laurent Weill (2010), Is Corruption an Efficient Grease? *World Development*, doi:10.1016/j.worlddev.2009.06.004.

Méon, Pierre Guillaume, and Khalid Sekkat (2005). Does Corruption Grease or Sand the Wheels of Growth? *Public Choice*, doi:10.1007/s11127-005-3988-0.

Russian Governmental Statistical Comittee Official Site. *Statistical Database of the Russian Federation*. Retrieved from URL: https://www.gks.ru/ (date of access: 22.12.2019)

Sallaberry, Jonatas Dutra, et al. (2020) Measurement of Damage from Corruption in Brazil. *Journal of Financial Crime*, doi:10.1108/JFC-04-2020-0057.

Shemetev, Alexander (2010), Complex Financial Analysis and Bankruptcy Prognosis and Also Financial Management-Marketing Self-Taught Book. *Book (in Russian)*. 1st ed., *Polygraphist*, https://books.google.cz/books?id=dq0sUIpiGuAC&hl=ru&source=gbs_similarbooks. (date of access: 29.12.2020)

Shemetev, Alexander (2012). Complex Financial Analysis and Bankruptcy Prognosis and Also Financial Management-Marketing Manual for Self-Tuition Book. 1st ed., Zodchy, https://books.google.cz/books/about/Complex_financial_analysis_and_bankruptc.html?id=iViOsALV23QC&redir_esc=y. (date of access: 29.12.2020)

Shemetev, Alexander (2020). R Package alexandershemetev. *GitHub Repository*, https://github.com/Alexandershemetev/alexandershemetev (date of access: 29.12.2020)

Silver, Carole Kupferman, and Ayn Rand (1978), Atlas Shrugged. *The English Journal*, doi:10.2307/814754.

Slobodyan, Sergey, and Raf Wouters (2012), Learning in a Medium-Scale DSGE Model with Expectations Based on Small Forecasting Models. *American Economic Journal: Macroeconomics*, doi:10.1257/mac.4.2.65.

Smith, Adam (1776), *An Inquiry into the Wealth of Nations*, Strahan and Cadell, London, 1776.

Thompson, Dennis F. (2018), Theories of Institutional Corruption. *Annual Review of Political Science*, 2018, doi:10.1146/annurev-polisci-120117-110316.

World Bank (1997), "Helping Countries Combat Corruption: The Role of the World Bank." *Poverty Reduction and Economic Management*.