Determinants of economic development: SEM approach.

Oleg Lugovoy[a,b], Vladimir Potashnikov[a] and Andrey Zubarev[a]

[a] Russian Presidential Academy of National Economy and Public Administration

[b] Environmental Defence Fund

January 31, 2017

CONFERENCE DRAFT PAPER – PLEASE DO NOT CITE OR CIRCULATE WITHOUT THE AUTHORS' PERMISSION

Abstract

In this paper, we propose not only a new approach for finding main determinants of economic development and defining development itself but also a new methodology for verifying theoretical models on big raw datasets. This methodology includes working with database with the help of regular expressions and picking out clusters of variables with the purpose of reducing the set of variables to potentially important ones. Using SEM technique we estimated many models where economic development was treated as a latent variable and characterised by GDP and quality of health and education systems. Quality of institutions and involvement in international trade were found to be main determinants of development what comports with many theoretical models. We also showed how constructed models may be used for forecasting.

Keywords: SEM, economic development, multi-country panel, institutions.

JEL Codes: O11, O43, O47, C3.

1 Introduction

In this paper, we propose a new approach for finding main determinants of economic development and defining development itself. In the economic literature there are many theoretical models and schemes that define main factors of economic development. The most important question is how to verify or falsify these models.

To deal with this issue one can find huge datasets with a vast number of variables. It may be a challenge to look them through and choose appropriate variables for the research. Problems could arise from inappropriate measurement of variables or simply separating suitable variables by topic.

Two more important points should be noted. First, not all variables from theoretical model have good analogues or proxies in the real data. Second, researchers often have to deal with unobservable objects (what economic development in particular is). This leads us to the conclusion that the new methodology is needed to solve these problems. This methodology should be unified and automatic as much as possible. This goal was mainly inspired by the paper of Sala-i-Martin (1997) in which the author looked for determinants of economic growth using relatively small dataset (consisted of 60 variables) and proposed ad hoc criteria for checking the robustness of each variable. In this research we want to extend the methodology of Sala-i-Martin (1997) in accordance with issues pointed above.

In short, we present some kind of methodology that at the first stage deals with big datasets and constructs "correlation baskets" (to be described later in the main part of the paper) consisting of potentially important variables and cluster of variables for explaining economic development. At the second stage we use SEM approach for modelling development as a latent variable characterised by different important factors contrary to many other studies where GDP or income level were used as a proxy for economic development. The concept of latent variable in SEM framework gives us some flexibility in capturing what we understand by the term "economic development".

Before trying to solve the stated problem it is worth discussing some theoretical and empirical background of economic growth and development and application of SEM technique.

In the paper Rodrik et al. (2004) authors estimate system of simultaneous equations using theoretical scheme of the determinants of economic development. Their scheme allows for different channels of influence between geography (which is exogenous), integration (could be treated as involvement in world trade), institutions and income level, the latter one is treated as proxy for development. The authors found some evidence that the quality of institutions is the most important factor for income level. They also found that geography has less important but still significant direct and indirect (through the quality of institutions) effect on income level. Controlling for institutes they didn't find any evidence in favour of hypothesis of direct effect of integration on development.

There are many other papers which explore multi-country regression analysis to find main determinants of differences in economic growth and development. Most important seminal works are Levine and Renelt (1992), Barro (1996), Sala-i-Martin (1997). In the latter work, to explain economic growth, the author introduced original method of variable choice based on the distribution of the variable's coefficients from different equations. Barro (1996) estimated system of equation with instruments and found that rule of law and democracy are important factors for economic growth, albeit positive effect from democracy on growth is valid only up to some level of democracy. The author also found that growth is stimulated by higher initial level of some education indicators and terms of trade (the latter goes in line with hypothesis in Rodrik et al. (2004)). The paper by Acemoglu et al. (2000) provides some evidence that differences in the quality of institutions account for about three-quarters of differences in economic growth.

Results of empirical studies sometimes contradict each other. This arises from differences in data used, pool of countries, time periods, specifications and methodology. With the help of new methodology that extends that of Sala-i-Martin (1997) we want to test the results obtained in Barro (1996) employing theoretical schemes proposed by Rodrik et al. (2004). Endogenous and unobservable factors also complicate estimation process and results interpretation. SEM technique helps partly circumvent these problems. Therefore, it seems reasonable to consider some works which incorporate SEM approach and are directly or indirectly associated with economic growth and development.

Paper by Shen and Williamson (2005) doesn't deal with economic growth and development explicitly but studies determinants of democracy and corruption control using SEM approach. Their model allows for direct effect of democracy on corruption control. Both variables are latent. The authors found that not only democracy affects corruption control, but economic freedom (openness) and government strength also affect it. Corruption does impede economic development so we could use some variables from the paper above in our study.

Corruption is also included into the model as a single latent variable in the paper by Dreher et al. (2007). In fact, the authors proposed a concept of the corruption index. The work by Macias and Cazzavillan (2010) is dedicated to the estimation of main factors that affect the size of informal sector of economy (it is strongly related to the level of economic development) which is included into the model as a latent variable.

In some sense, the combination of the last three models described above is presented in the work of Buehn and Schneider (2009), where the authors examine the relationship between shadow economy and corruption. Some variables from this model could be potentially useful for our study, for example fiscal regulation, transfers and subsides, bureaucracy costs, rule of law, judicial independence, etc. It was found that there is a positive correlation between corruption and shadow economy, but the effect of corruption on shadow economy is stronger than the reverse one.

The paper is structured as follows. Section 2 provides description of the dataset. Section 3 starts with dataset screening and then provides the technique of construction of correlation baskets which give us potentially important variables for economic development explanation. In Section 4, we present model specification and estimation results. Section 5 concludes.

2 Model specification

In this study, we want to apply SEM technique for finding main determinants of economic development. Most studies concerning economic development treat it as a high level of GDP per capita or income. But having high income only is not enough for country to be named as developed. When we are talking about a developed country we assume that some other important characteristics (such as health and education standards, quality of institutions, etc.) are of the high level. So it is quite convenient to aggregate all these characteristics somehow.

According to SEM principles latent variables should be characterised by their indicators which should be correlated strongly enough. As a main indicator we employ GDP PPP per capita. We also decided to use two more indicators of development besides GDP that should characterise the quality of health and education systems. The list of explanatory variables should somehow comport with the theory mentioned above. Graphical representation of the model is as follows

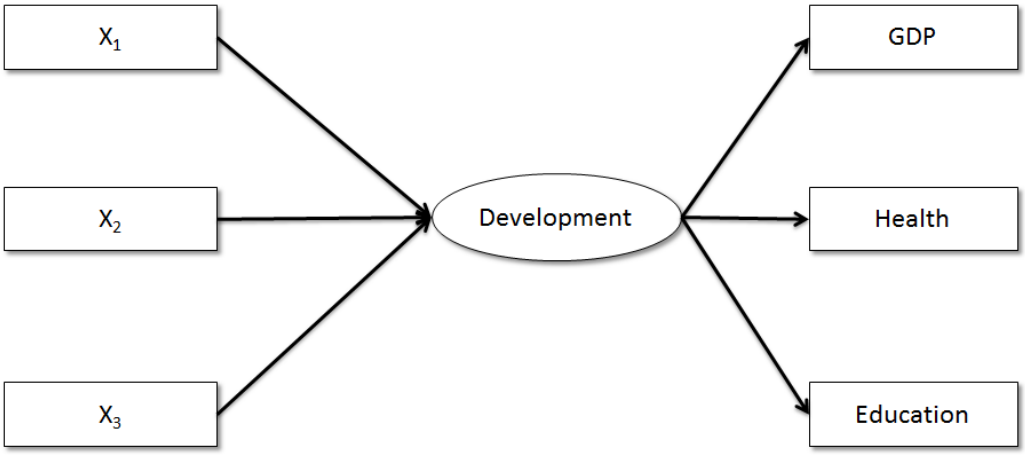

Figure 1: Model structure

where X_i stands for exogenous variable that may explain development. It may be some institutional characteristics, indicator of involvement in international trade, etc. The number of exogenous variable may also vary.

In fact, we follow Sustainable Development Goals[1] when choosing indicators of development. Unfortunately, it is quite difficult to take into account all the sustainable development goals because some of them are poorly measured in data and finding appropriate proxies may be a challenge. The problem of model estimation (algorithm convergence) in case of too many indicators of latent variable also arises. That is why we decided to concentrate only on main goals according to our view.

The very important question is how to find potential candidates not only for explanatory variables but also for indicators of development, because there are a lot of variables that describe health and education systems. At the same time, it is hard to estimate all possible models if one have thousands variables to choose from because in this case the estimation procedure will last far too long. That is why we need some kind of methodology to handle this problem. We describe this methodology in the next three sections.

3 Data description

To get a relatively big bunch of variables for our study we gather them from several reliable databases. Finally we have a multi-country panel filled with yearly data. Let's provide some details about the datasets we employ.

The database of the World Bank, WDI (World Development Indicators), contains information about most types of economic activities of countries. WDI also includes data on demography, health and education. It consists of more than 1,300 variables for 247 regions (over 170 countries) starting from 1960. Variables WDI database are grouped in three different levels of nested topics. Hereinafter we will address them as clusters at level i, where $i = 1, 2$ or 3 (clusters at level 1 may consist of several subclusters at level 2, etc.)

PWT (Penn World Table) database contains information about economic activities of countries based on the methodology SRNA (System of Real National Accounts). According to the PWT's authors, this methodology allows for a more correct comparison of observations for different countries compared with the SNA (System of National Accounts) (see Summers and Heston (1991)).

We also use Economic Freedom of the World Data (The Fraser Institute) database and The Worldwide Governance Indicators (WGI). WGI database contains 6 aggregates of the efficiency of public administration, such as «Voice and Accountability», «Government Effectiveness», «Control of Corruption», etc. Economic Freedom database consists of 41 variables which measure how policies and

[1] http://www.un.org/sustainabledevelopment/sustainable-development-goals/

institutions in countries support economic freedom.

SEM is based on the covariation analysis. The closer variables' distributions are to the normal one, the more valid the inference is. For this reason we apply some transformations to the variables to make them look more like normally distributed.

There are several standard transformations that lead to infinite support and moments close to those of the normal distribution. We use logistic, Fisher and logarithmic transformations, and the choice between them depends on the properties of the initial density function. Box-Cox transformation is more universal and gives better results but we want to have a more unified system of transformation for better interpretation of the results. Solvers that maximise likelihood can work improperly if variables differ in scale far too much (more than by 10^3 times). For handling this problem we use normalisation for some variables.

4 Preliminary variable selection and correlation baskets

The model we are going to estimate should be based on the theory. However, the model may contain unobservable or poorly measured variables. At the same time, there are a lot of variables which can substitute for those of our interest. In particular, we need to find not only candidates for development (which is a latent variable) indicators but also explanatory variables. Manual search for variables (including substitutes for indicators) can be time consuming and lead to unavoidable mistakes. But finding as many of these variables as possible in a large database is crucial for testing the robustness of the results to the choice of the substitutes.

To overcome this problem, we decided to develop a new partly automatic procedure which searches potential candidates to include in our model. The algorithm is designed for the search of variables that are close (in some sense) to a "base" variable, which is supposed to be in the model as a main indicator of the latent variable (it also might be the case that instead of this variable its close substitutes are included into the model). In our case GDP PPP per capita was chosen as that "base" variable.

In the procedure we need to define a metric for measuring how close variables are to each other. As far as we decided to use SEM technique as a main instrument in our analysis it seems quite intuitive to use correlation as a metric. That is why we decided to call the following procedure "correlation basket" construction.

Let's describe the algorithm. At first stage we use regular expressions to pick out variables that are inappropriately measured for our study. These variables have to be normalised by the population, and/or GDP PPP, and/or price level to get rid of the scale of economy and nominal values. Nominal variables are transformed into the real ones through the search by "current" and "LCU". Variables that measure energy or mass are normalised by GDP using the following regular expressions: "kt", "Mt", "kg", "energy", "kWh", "toe", "tce", "kg of oil equivalent", "PJ", etc. Some variables' names include more than one of the expressions used so we applied special priority rules.

At the second stage some "base" variable has to be chosen (as we sad before it is GDP PPP per capita in our case). Then we find the most correlated variables with the "base" one. We bound the number of these variables by 400 which is quite a naive choice based on the dimension of the whole dataset and it definitely may vary.

Some constraints on observations should also be introduced. In particular, we consider only those variables which have at least 800 common observation points with the "base" variable and include 11 periods (years) for at least 50 countries in order to treat observation selection as representative.

Some variables that are not highly correlated with the "base" variable may still be very important. They could explain relatively small part of information that was left unexplained after regressing the "base" variable on those highly correlated with it. So at the second stage, to take into account these second order effects, we look for variables, that are highly correlated with the unexplained variance at the first step for each of the 400 variables. In practice, there is no difference between looking at correlation with unexplained variance or just at R^2 of the hole regression. Then the procedure may be continued for as many steps as one needs.

Finally we get the rooted tree (using the notation of graph theory) representing our regressions. GDP PPP per capita is the root of the tree. There are no more than 400 vertices (nodes) that have depth $= 1$. The number of children at each these vertices is bounded by 50. So the largest possible number of vertices that have depth $= 2$ equals $400 * 50 = 20000$. As far as depth $= 3$ is concerned, each vertex that has depth $= 2$ should have no more than 5 children, this makes the bound for the number of vertices with depth $= 3$ be equal to $20000 * 5 = 100000$ (actual number of vertices may be less than 100000 due to constraints described above). They are actually leafs of the tree because we decided to bound the depth of the tree by 3. Structure of the tree is depicted in Figure 2.

If the observation is the same for all variables then the search for variables on each step that are highly correlated with the unexplained variance from the previous step is equal to the search of an additional regressor which gives the highest value of R^2. But the number of observations that are used for correlation

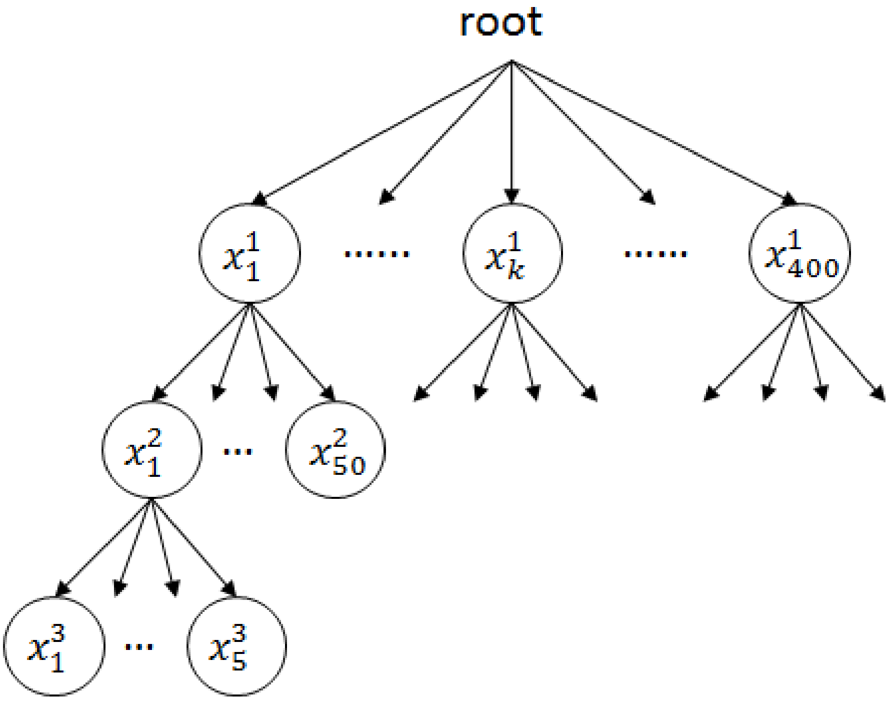

Figure 2: Regression tree

estimation on each step may decrease due to data incompleteness. As a result this is not the explanatory power of a new variable but the change in the observation set that may raise R^2. To control for this potential problem we introduce additional criteria (constraints). Here they are:

1. Explained variance of the variables previously included (into regression) shouldn't fall by more than 10% with the addition of a new variable due to reduction of the observation set. Formally, this could be written as

$$R^2(y|x_1, ..., x_n)|_{\text{domain}(x_{n+1})} > 0.9 * R^2(y|x_1, ..., x_n)$$

   where conditioning on domain(x_{n+1}) means reducing the observation set to that which variable x_{n+1} is defined on.

2. R^2 at each stage should increase at least by 1%:

$$R^2(y|x_1, ..., x_{n+1}) - R^2(y|x_1, ..., x_n) > 1\%$$

3. Additional regressor should raise R^2 on the same domain:

$$R^2(y|x_1, ..., x_{n+1}) - R^2(y|x_1, ..., x_n)|_{\text{domain}(x_{n+1})} > 1\%$$

After all regressions are estimated and tree is constructed, the variables that were picked out should be grouped by clusters (and/or subclusters) which they are assigned to in the dataset. To extract the necessary information from the results, we use such graphical tools as histograms for the increase in R^2 for each value of depth of the tree and for each cluster. At the Figure 3, three histograms for the cluster WGI are depicted.

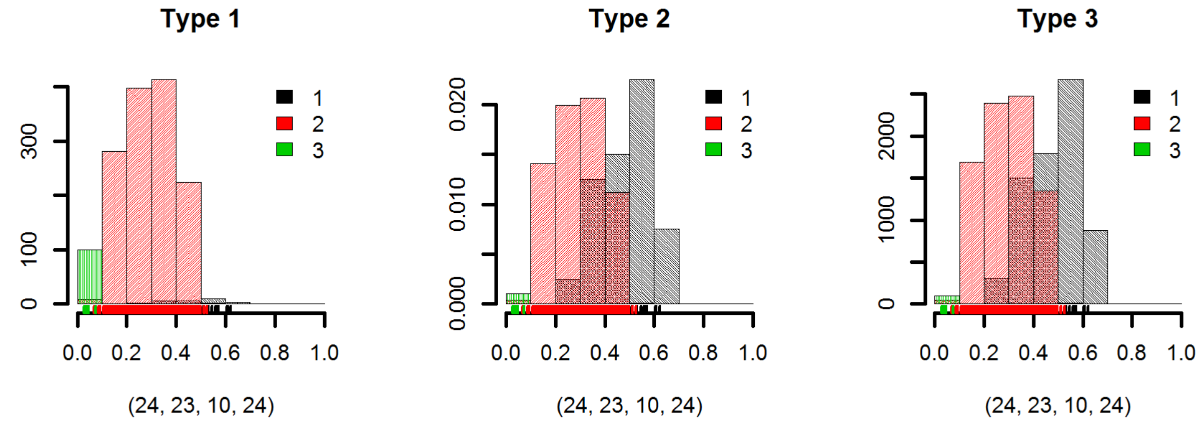

Figure 3: Correlation basket for WGI cluster

Horizontal axis represents the increase in R^2 for particular cluster with the step of 10%. Different colours correspond to different values of depth which variables are considered from. Vertical axis displays how many times variables have value of R^2 increase in each decile. Numbers in parenthesis represent the quantity of unique variables at vertices that have depth 1, 2 and 3 and in total correspondingly.

There are much more vertices of depth = 3 than vertices of depth = 1, so the higher value of depth, the higher the probability of the variable to appear in vertices of this depth. That is why the bars that correspond to vertices with depth = 1 could be hardly seen. The naive solution in this case is the adjustment of bars' height for each value of depth to the number of vertices. So bars of black, red and green colour should be adjusted by 400, 400*50, and 400*50*5 correspondingly. Then the bars' height is equal to the proportion of variables' appearance (variables from particular cluster) in the total number of vertices with particular value of depth. That is how the central histogram in Figure 3 was depicted.

The second option is to count vertices and all descendants (children, children of children, etc.) for each cluster and each value of depth[2]. This corresponds to

[2] E.g. if variable at vertex with depth = 1 have 3 children, and these 3 vertices have the total number of children of 5, then we get $1 + 3 + 5 = 9$.

the left histogram at Figure 3.

After "correlation basket" construction, there are 513 unique variables left that are assigned to 18 clusters. Each cluster includes 3 to 140 variables. The biggest cluster, *"Economic Policy & Debt"*, is divided into 6 subclusters. Consideration of relatively small clusters of variables gives us an opportunity to split variables into three groups: exogenous, indicators and those that are ill-suited or unimportant for our analysis.

Charts depicted in Figure 4 give us a clue about which variables are ill-suited or unimportant. E.g. if any cluster appears at vertices with depth $= 1$, but doesn't appear at vertices with depth $= 2$ or gives small value of increase in R^2 (or appears only at depth $= 3$). In such manner, clusters *"Price levels, expenditure categories and capital", "Exchange rates and GDP price levels", "National accountsbased variables", "Shares in CGDPo"* may be picked out as inappropriate.

We found many clusters such as *"Infrastructure", "Real GDP, employment and population levels", "Price levels, expenditure categories and capital", "Financial Sector", "Exchange rates and GDP price levels", "Current price GDP, capital and TFP", "Social Protection & Labour", "Environment", "Shares in CGDPo", "National accountsbased variables"* to be ill-suited for the model as exogenous variables due to mismatch with the theory. For example, *"Infrastructure"* consists of proxies for capital, while our goal is not to estimate production function's parameters but find conditions which may help to achieve desirable level of capital. We also think that cluster *"Social Protection & Labour"* looks more like indicator of development, but in the very beginning we decided to restrict ourselves with three indicators in this study.

Clusters *"Health"* and *"Education"* contain indicators of development according to our a priori view. *"WGI"* and *"Economic freedom"* consist of proxies for institutions which are important an part of theoretical models. So we decided to combine them and put in one cluster of exogenous variables in the SEM specification called *"Institutions"*. Another cluster the importance of which stems from theory is *"Private Sector & Trade"* because it characterises involvement in international trade. So we also use it as a separate cluster in the SEM model (called *"Trade"*). *"Economic Policy & Debt", "Poverty" and "Public Sector"* are combined in the third cluster of exogenous variables called *"Leftover"*.

5 Estimation results

In the previous section, we picked out several clusters of variables that are potentially important for explaining development. Each cluster includes only those

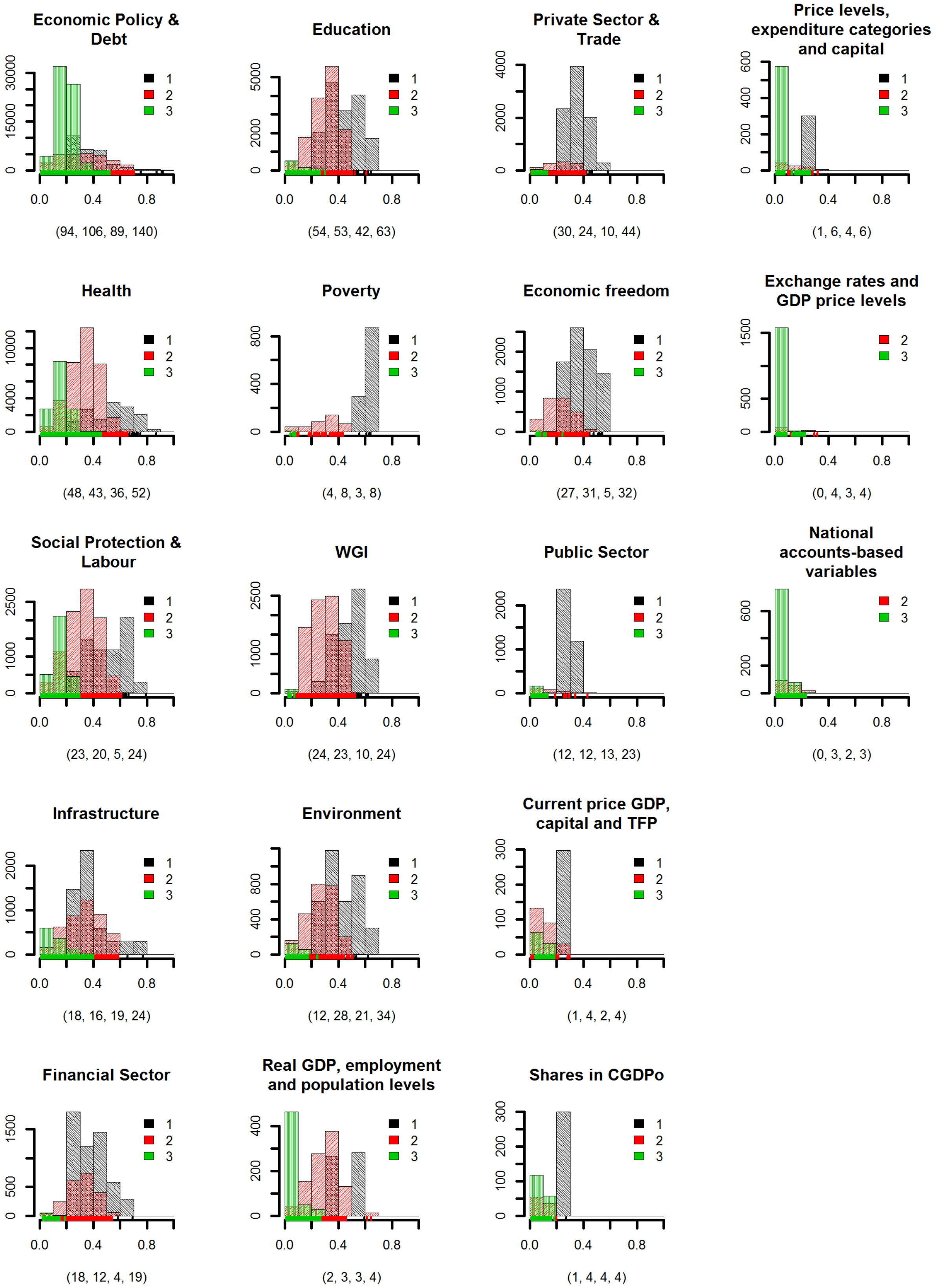

Figure 4: Correlation baskets (of type 3)

variables that appeared in the "correlation baskets". We also found out which variables that characterise education and health system are highly correlated with our "base" variable so they could be used as potential indicators of development along with GDP PPP per capita. Three clusters were chosen to represent exogenous variables in the SEM model specification: *"Institutions", "Trade"* and *"Leftover"*.

After "correlation baskets" construction each cluster contains the following number of variables: *"GDP"* – 1, *"Health"* – 18, *"Education"* – 7, *"Institutions"* – 38, *"Trade"* – 41, *"Leftover"* – 29. Combinatorial calculus gives us 5 692 932 models in total. When applying criteria for the minimum number of observations (mentioned above) and setting threshold for absolute value of correlation between indicators ($\geq$ 0.6), the number of models reduces to 2 254 152. Finally, only 168 262 of them were estimated properly (solver's algorithm[3] didn't converge for others).

Let's now consider how we can analyse these estimation results looking at specific plots depicted in Figure 5 for one particular variable. In this case it is "Personal remittances, received (% of GDP)". Grey points above zero line represent all estimated models. One of the key features of these plots is the reflection around the horizontal axis: points above zero line correspond to models where the coefficient associated with the variable of interest is positive, while points under zero line correspond to a negative sign of coefficient. So one shouldn't treat negative part of the vertical axis as a negative domain: the domain is still positive but points correspond to a negative sign of coefficient. So grey points under zero line are just a reflection of those above zero. Points highlighted with the main colour (here it is a blue one) shows models which the variable is included to and in which it is significant. Auxiliary colour (purple here) corresponds to the models with insignificant coefficient for this variable. Horizontal axis itself shows the proportion of explained variation of latent variable. Vertical axis of the left plot represents RMSEA-value[4] of the estimated model. To make it look more transparent, we added right plot where vertical axis shows p-value of RMSEA. Satisfactory models have low value (usually between 0 and 0.05) of this statistics and the right plot makes the cloud of these models less dense.

In this particular plot, we see that the coefficient not only loses significance but also has different signs in some models. This leads us to the idea that this particular variable isn't robust. Another important point is that just a few models have both high value of explained variation of latent variable and p-value of

[3] For SEM estimation we used *lavaan* package in R.

[4] One of the fit measures for SEM models.

RMSEA. So the set of satisfactory models shouldn't include such a variable.

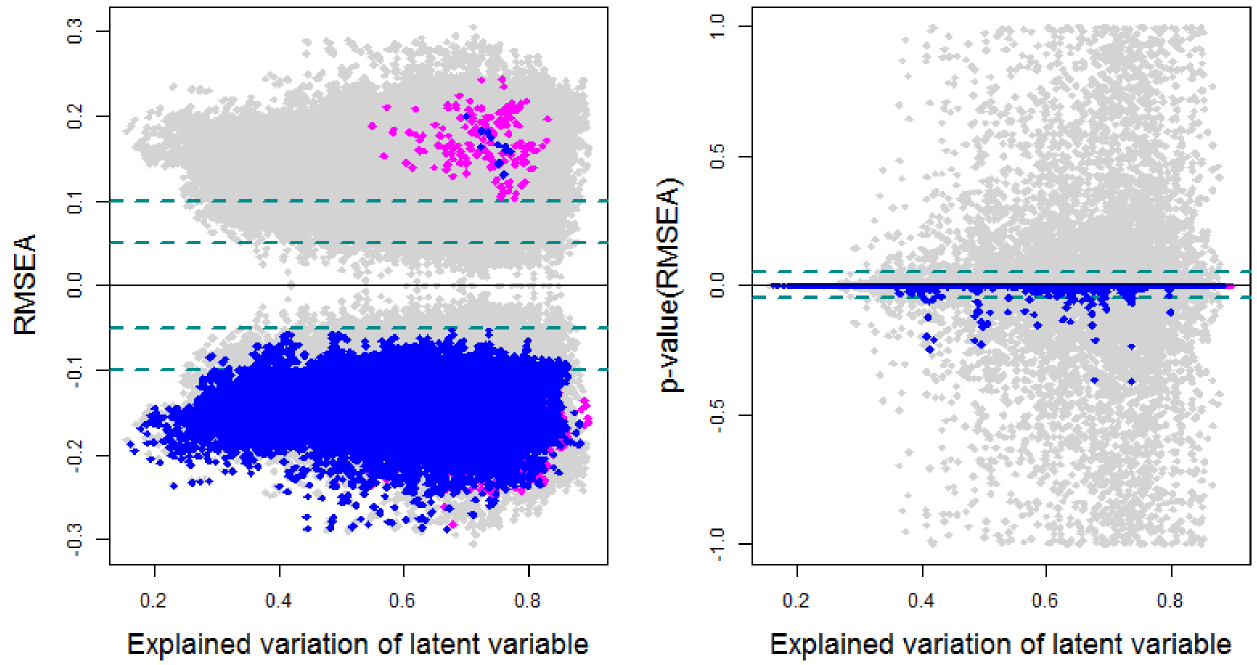

Figure 5: Estimation results for "Personal remittances, received (% of GDP)"

As follows from the description above, the most reliable models correspond to points in upper or lower left corner of the right plot with high p-value of RMSEA and proportion of explained variation of latent variable. Let's now switch to some variables that are included in one of these models that will be presented below. Figures 6 and 7 depict models that include variables from the clusters *"Health"* and *"Institutions"*, particularly "Mortality rate, neonatal (per 1,000 live births)" and "Voice and Accountability" correspondingly. From the plots we could see that there are quite many reliable models with high p-value of RMSEA and proportion of explained variation of latent variable. Coefficients associated with these variables also are of a constant sign and do not loose significance. So we may treat the behaviour of these variables as a robust one. The same could be said about four other variables from the final model. Similar plots for these variables can be found in Appendix.

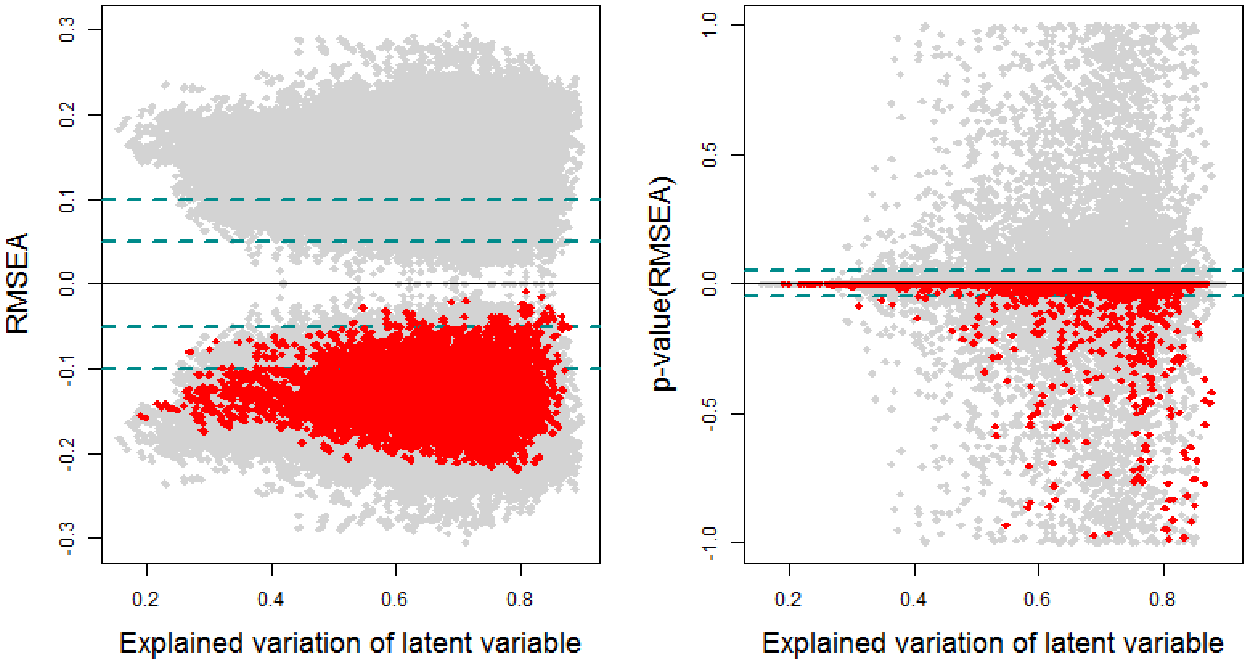

Figure 6: Estimation results for "Mortality rate, neonatal (per 1,000 live births)"

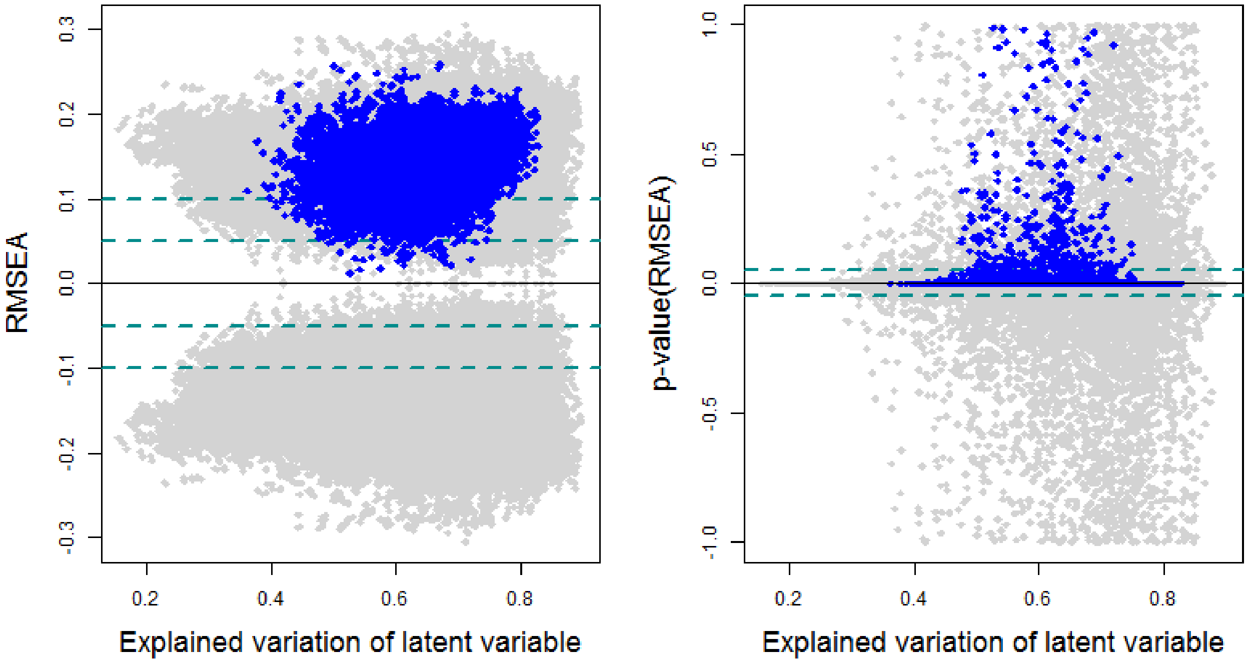

Figure 7: Estimation results for "Voice and Accountability"

The number of models that have p-value(RMSEA) $\geq 10\%$ and proportion of the explained variation of latent variable higher than 2/3 is 2035. We provide estimation results for one (representative) of these models. The model includes the following variables: "GDP PPP per capita", "Mortality rate, neonatal (per 1,000 live births)", "School enrollment, secondary, male (% gross)", "Voice and Accountability"[5], "Merchandise exports by the reporting economy (% of GDP)" and "Revenue, excluding grants (% of GDP)"[6]. Parameter estimates are presented in Table 1.

Table 1: Parameter estimates.

Measurement part

	Unstd.est.	Std.err.	P-value	Std.est.	Std.all
GDP	1	-	-	0.599	0.942
Health	-0.800	0.015	0.000	-0.480	-0.940
Education	0.288	0.007	0.000	0.172	0.887

Structural part

	Unstd.est.	Std.err.	P-value	Std.est.	Std.all
Institute	39.131	1.868	0.000	65.305	0.604
Leftover	0.250	0.032	0.000	0.417	0.232
Trade	0.577	0.071	0.000	0.963	0.285

Variances

	Unstd.est.	Std.err.	P-value	Std.est.	Std.all
GDP	0.046	0.005	0.000	0.046	0.113
Health	0.031	0.002	0.000	0.031	0.117
Edu	0.008	0.000	0.000	0.008	0.214
y	0.112	0.009	0.000	0.311	0.311

The column with unstandardised estimate corresponds to the the case where the first parameter (this does not limit generality) is set to unity. We see that all the coefficients are significant. Signs of all coefficients are as anticipated. In the measurement part, it is natural to understand higher development as higher GDP

[5] Voice and accountability captures perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media.

[6] Revenue is cash receipts from taxes, social contributions, and other revenues such as fines, fees, rent, and income from property or sales. Grants are also considered as revenue but are excluded here.

per capita, higher school enrolment and lower neonatal mortality rate. As far as structural part is concerned, higher values of quality of institutions, involvement in international trade and the ratio of revenue (excluding grants) to GDP enhance development.

The column with standardised estimates (with the variance of the latent variable normalised to unity) provides us with additional information. We can see that proportion of unexplained variance of the latent variable is about 31%, which means that variables from the structural part of the model explain about 69% of the variance of the latent variable. The last column shows parameter estimates in the case when variances of all variables are normalised to unity.

We could check the quality of the model looking at some descriptive statistics from Table 2. ML and Robust columns correspond to basic MLM estimation and one with the correction for possible non-normality in the data. Chi-square statistics has high p-value what speaks in favour of good quality of the model. CFI and TLI tests have statistics higher than 0.9 which states that our model improves baseline model (independence model). But actually this result is not a very impressive one. Looking at RMSEA is more important to understand the quality of the model. Based on the value of the lower bound of the confidence interval (lower than 0.05) it doesn't reject close-fit hypothesis. At the same time with the value of the upper bound of the confidence interval lower than 0.1 the poor-fit hypothesis is rejected.

Table 2: Model estimate statistics.

Main Chi-square statistics

	ML	Robust
Minimum Function Test Statistic	9.809	9.015
Degrees of freedom	6	6
P-value (Chi-square)	0.133	0.173
Scaling correction factor for the Satorra-Bentler correction	1.088	

User model versus baseline model

	ML	Robust
Comparative Fit Index (CFI)	0.999	0.999
Tucker-Lewis Index (TLI)	0.998	0.999
Robust Comparative Fit Index (CFI)	0.999	
Robust Tucker-Lewis Index (TLI)	0.999	

Root Mean Square Error of Approximation

	ML	Robust
RMSEA	0.024	0.021
90 Percent Confidence Interval	(0.000;0.049)	(0.000;0.046)
P-value RMSEA $\leq$ 0.05	0.954	0.973
Robust RMSEA	0.022	
90 Percent Confidence Interval	(0.000;0.049)	(0.000;0.046)

**Number of observations 1128*

As far as endogeneity problem is concerned, it couldn't be fully handled but there are some hints that help us see how critical it is for the model. Moving different variables between structural and measurement parts of the model back and forth and looking at the quality of the model could help us understand whether there is a substantial endogeneity for each variable or for the whole model. We constructed many models in such a way and found no improvement while most of the models were of much worse quality. These results speak in favour of the adequacy of the original specification.

One may find application of this results to forecasting being rather interesting. The concept of specifying development as a latent variable in the SEM framework gives us an opportunity to extract data on this variable from the model. Trying to answer the question where Russia stands in terms of its development we made a quite transparent plot depicted in Figure 8. Two dashed lines correspond to the actual levels of development for Russia and Germany as the model predicts them. We omitted the scale of horizontal axis because it makes not much sense for a latent variable. As we can see Germany has higher development level which is not very surprising. There are also two distribution densities depicted that represent predicted density function of development distribution based on exogenous variables only (indicators of development are not taken into account). The grey coloured PDF reflects the case where all of the exogenous variables are assigned actual values for Russia. The purple coloured density function is based on the same levels of all variables except for institutional characteristic, which is assigned the actual value for Germany.

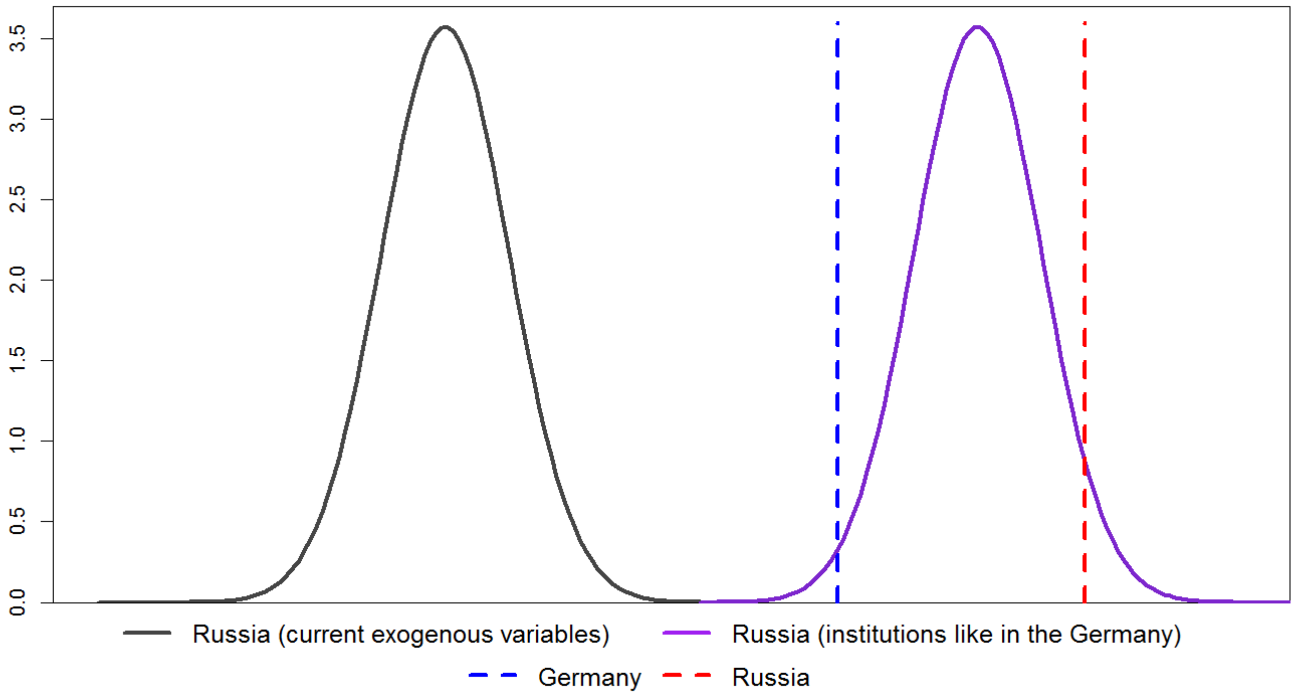

Figure 8: Development predictions

Two results that follow from this chart are of particular interest. First, had Russia the same quality if institutions as that of Germany, then Russian level of development would be quite close to the development of Germany. Another interesting point is that countries with exogenous characteristics similar to Russian ones (which is represented by the grey coloured PDF) have far worse level of development compared to the actual level of Russia's development. There is a logic behind it: observable levels of health and education systems' indicators (which in fact are very inert) are quite high in Russia due to the USSR's heritage, and GDP per capita level may be relatively high due to oil revenues. But assuming the decreasing pattern for oil revenues in the future, we may conclude that not improving institutional environment in Russia will lead to a substantial drop in the level of development that will be expressed as a decline in the quality of education and healthcare.

6 Conclusion

This study is devoted to finding the main determinants of economic development. Along with this issue, we proposed new unified and automatic methodology for dealing with with huge datasets in order to find main factors that affect any unobservable variable. This methodology may be used for verifying theoretical models. The dataset used in this study was checked and transformed with the help of regular expressions. Then we significantly reduced the number of potentially important variables and clusters of variables by construction of "correlation baskets". At the last stage, SEM methodology was used for finding determinants of economic development.

We treated development itself as a latent variable and characterised it by such indicators as level of GDP PPP per capita and quality of health and education systems. We found that institutional characteristics and involvement in international trade are the most important factors of economic development, more then two thirds of the variation of which is explained by our models.

We also applied our results to forecasting. Based on model predictions we concluded that not improving institutional environment in Russia may lead to a substantial drop in the level of development expressed by a decline in the quality of education and healthcare.

As the prospective future directions in this study we consider are the expansion of the database used and the introduction of geography and other important characteristics into the model.

References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2000). The colonial origins of comparative development: An empirical investigation. Technical report, National bureau of economic research.

Barro, R. J. (1996). Determinants of economic growth: a cross-country empirical study. Technical report, National Bureau of Economic Research.

Buehn, A. and Schneider, F. (2009). Corruption and the shadow economy: a structural equation model approach.

Dreher, A., Kotsogiannis, C., and McCorriston, S. (2007). Corruption around the world: Evidence from a structural model. *Journal of Comparative Economics*, 35(3):443–466.

Levine, R. and Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American economic review*, pages 942–963.

Macias, J. B. and Cazzavillan, G. (2010). Modeling the informal economy in mexico. a structural equation approach. *The Journal of Developing Areas*, 44(1):345–365.

Rodrik, D., Subramanian, A., and Trebbi, F. (2004). Institutions rule: the primacy of institutions over geography and integration in economic development. *Journal of economic growth*, 9(2):131–165.

Sala-i-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review*, pages 178–183.

Shen, C. and Williamson, J. B. (2005). Corruption, democracy, economic freedom, and state strength a cross-national analysis. *International Journal of Comparative Sociology*, 46(4):327–345.

Summers, R. and Heston, A. (1991). The penn world table (mark 5): An expanded set of international comparisons, 1950-1988. *Quarterly Journal of Economics*, 106(2):327–386.

Appendix A Estimation results

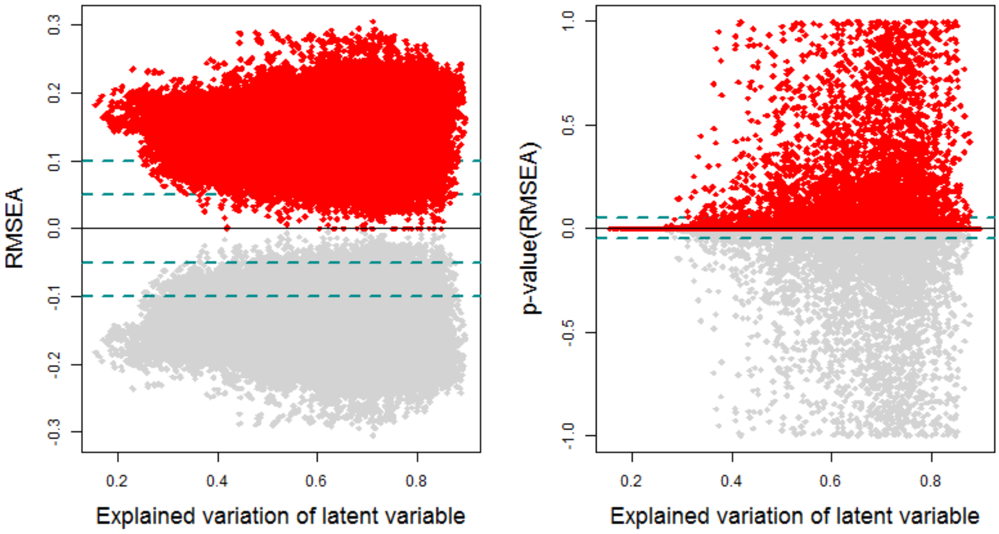

Figure A1: Estimation results for "Expenditure-side real GDP atchained PPPs per capita (2005US$pcap)"

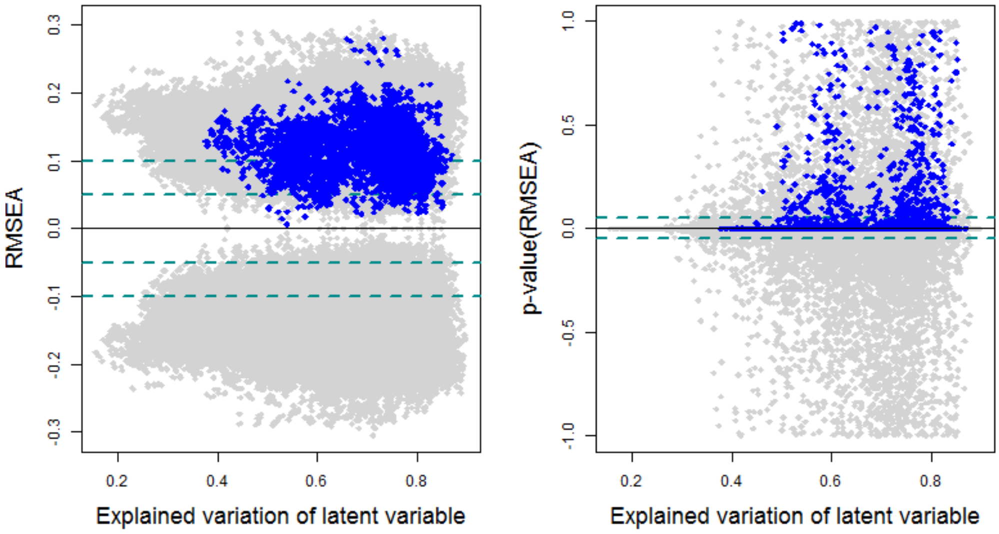

Figure A2: Estimation results for "Leftover Revenue, excluding grants (% of GDP)"

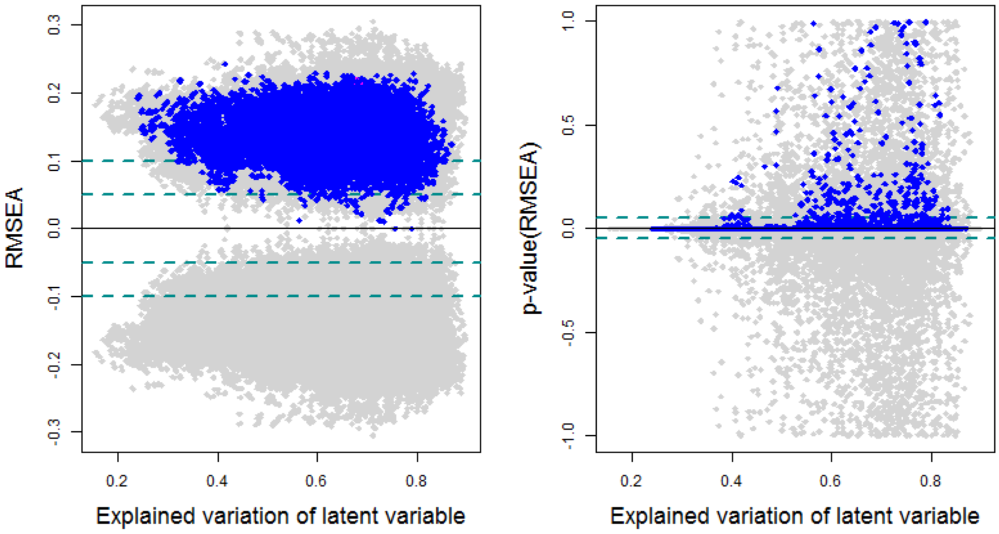

Figure A3: Estimation results for "Trade Merchandise exports by the reporting economy (current US$)"

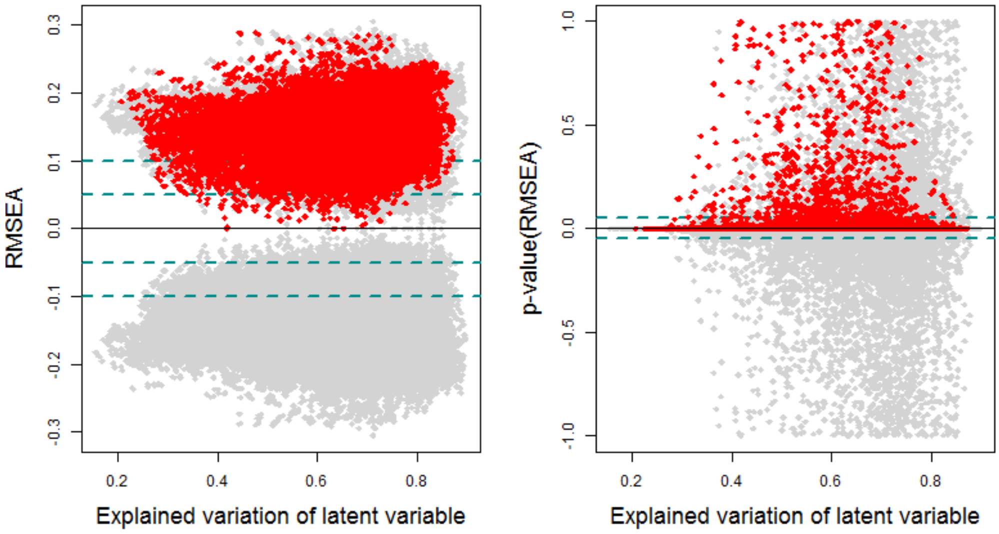

Figure A4: Estimation results for "School enrollment, secondary, male (% gross)"