EPJ Web of Conferences **173**, 03013 (2018) https://doi.org/10.1051/epjconf/201817303013
Mathematical Modeling and Computational Physics 2017

Data Mining in Institutional Economics Tasks

Igor Kirilyuk[1,⋆], *Anna* Kuznetsova[2,⋆⋆], and *Oleg* Senko[3,⋆⋆⋆]

[1] *Institute of Economics of RAS, Moscow, Russia*
[2] *Emanuel Institute of Biochemical Physics, Moscow, Russia*
[3] *FRC "Informatics and Control" of RAS, Moscow, Russia*

Abstract. The paper discusses problems associated with the use of data mining tools to study discrepancies between countries with different types of institutional matrices by variety of potential explanatory variables: climate, economic or infrastructure indicators. An approach is presented which is based on the search of statistically valid regularities describing the dependence of the institutional type on a single variable or a pair of variables. Examples of regularities are given.

1 Introduction

The theory of institutional matrices proposed by S.G. Kirdina ([1, 2]) supposes that two types of basic public institutions (or two types of institutional matrices in other words) may be identified. The first type may be characterized by a redistributive economy, a unitary political system, and communitarian ideology. This basic type will be referred to as X type. The second type that will be referred to as Y type is complementary to the first type. It is characterized by the market economy, the federal political system and the ideology of individualism. The countries of the world may be divided in two groups: X countries with dominance of X type of the institutional matrices and Y countries with dominance of Y type of the institutional matrices. Our goal is to check the institutional matrices theory and also to describe the relationship between the economic development characteristics and the dominating institutional type. In this regard we try to find statistically significant regularities allowing to discriminate between two types of countries by macroeconomic variables, climate and infrastructure indicators. The presence of valid patterns in the data provides strong evidence of the correctness of the institutional matrices theory. In [3] the relationship between the dominating type of institutional matrices and the climate or infrastructure variables was studied. In [4] the optimal valid partitioning (OVP) method was applied to compare joint dynamics of macroeconomic variables in groups of X and Y countries. A specially developed technique for evaluating multiple testing effects was used. A statistically significant relationship was found to hold between the institutional type and mutual dynamics of the gross domestic product (GDP), domestic credit and the governmental spendings.

⋆e-mail: igokir@rambler.ru
⋆⋆e-mail: azfor@yandex.ru
⋆⋆⋆e-mail: senkoov@mail.ru

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

2 Statistical tools

Today different data mining tools are available that may be useful in tasks of valid regularities search. There are several peculiarities that influence the choice of analytical tools. We must compare rather small groups of countries where experts are sure about the dominant type of institutional matrices. The full number of such countries varies from 25 to 35. Climate indicators include average monthly temperatures and precipitation amounts, their average year minimum, maximum and variations. Infrastructure indicators include road and railway density, irrigation levels and others. Average values of the macroeconomic variables over some time periods are less informative than various parameters describing joint dynamics of macroeconomic variables. The number of such parameters may achieve several tens. Hence the total number of factors involved in analyses may significantly exceed the total number of countries. Isolated factors often are not sufficiently informative and so it is important to analyse their combinations. There is no evidence that the studied factors may be distributed normally. In our research we used the OVP method [3]. OVP technique is aimed to explore the dependence of a target value on the set of explanatory variables. The OVP method includes the search of one-dimensional and two-dimensional regularities. A one-dimensional regularity is described by a threshold for the corresponding factor that separates in the best way the compared groups. Two quadrants to the left and to the right of the threshold are formed. Example of one-dimensional regularity is given at figure 1.

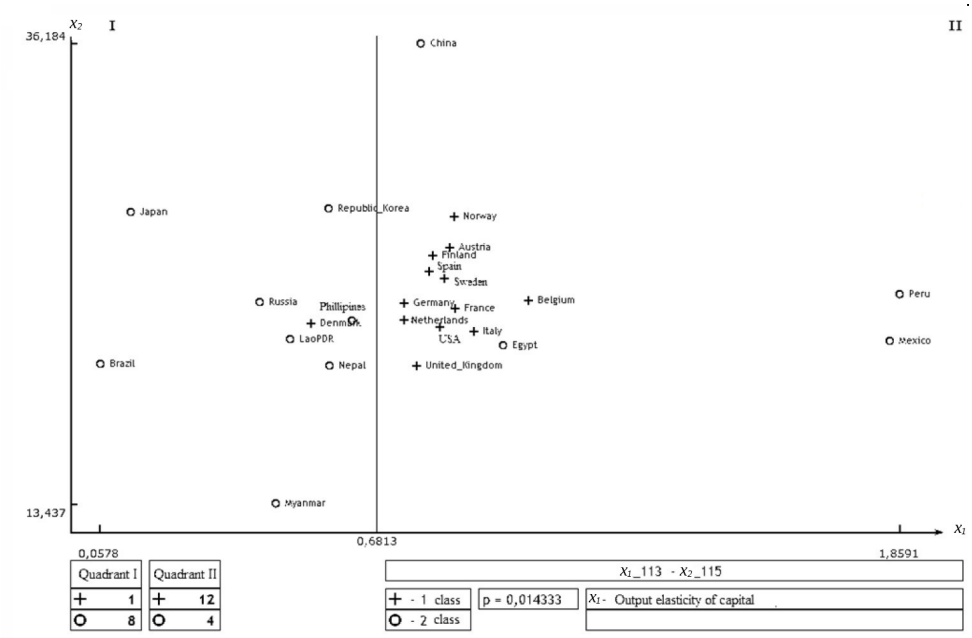

Figure 1. One-dimensional regularity describing the dependence of the institutional type on the output elasticity of the capital

A two-dimensional regularity is described by two boundaries that are parallel to the coordinate axes. The goal is the maximal possible separation of two compared groups. Four quadrants are formed as it is shown in figure 2.

Optimal boundaries providing the best separation of compared groups are searched through the optimization of the quality functional Q calculated by the dataset $\widetilde{S}_{\text{tr}} = \{(y^{(1)}, \mathbf{x}^{(1)}), \ldots, (y^{(m)}, \mathbf{x}^{(m)})\}$, where $y^{(1)}, \ldots, y^{(m)}$ are values of the target y and $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ are vectors of explanatory variables

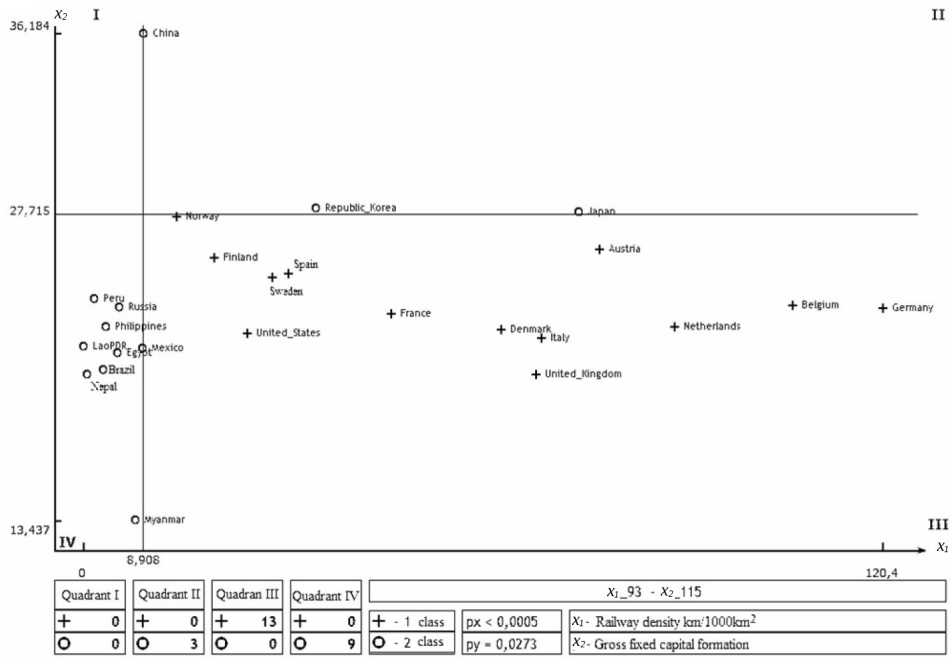

Figure 2. Two-dimensional regularity describing the dependence of the institutional type on the railway density and GFCF

$x_1, \ldots, x_n$ at objects from data set $\widetilde{S}_{\text{tr}}$: $Q = \dfrac{1}{D_y} \sum_{i=1}^{r} m_i \cdot (\hat{y}_i - \hat{y}_0)^2$, where $\hat{y}_0$ denotes the average of the target values y at objects of $\widetilde{S}_{\text{tr}}$, $\hat{y}_i$ denotes the average of the target values y at objects of $\widetilde{S}_{\text{tr}}$ from the quadrant Q_i, m_i is the number of $\widetilde{S}_{\text{tr}}$ objects from Q_i, r denotes the number of quadrants. The validity of the regularities is evaluated with the help of a random permutation test (RPT). The values of the functional Q on the data set $\widetilde{S}_{\text{tr}}$ are compared with the values of the functional Q calculated on a set of random samples generated from $\widetilde{S}_{\text{tr}}$. The positions of the values of the target y in the random sample are randomly permuted relatively to the fixed positions of the explanatory variables vectors. The proportion of random samples for which Q does not exceed the value calculated on $\widetilde{S}_{\text{tr}}$ is used as the p-value estimate. It was shown [4] that the RPT tests fulfill the null hypothesis that the target y is independent on the corresponding explanatory variable.

In the case of two-dimensional regularities the combination of two null hypothesis is tested. Let the dependence of the target y on two explanatory variables x_1 and x_2 is studied. At the first step a first null hypothesis is tested: the dependence of y on x_1 and x_2 is exhaustively described by a simple one-dimensional regularity with the threshold for the variable x_2 coinciding with the corresponding border value in the tested two-dimensional regularity. To test this hypothesis comparison is made with the values of the functional Q calculated on a set of random samples generated from $\widetilde{S}_{\text{tr}}$. The procedure coincides with the technique used when one-dimensional regularity is verified but permutations are allowed only inside two subsets of $\widetilde{S}_{\text{tr}}$ with x_2 values to the left and to the right of the threshold. Thus a p-value evaluating the validity of x_1 contribution to the two-dimensional regularity is calculated. In the same way a p-value evaluating the validity of x_2 contribution to the two-dimensional regularity is calculated. The two-dimensional regularity is assumed to be valid only if contributions of both variables are valid at some fixed accuracy level. It must be noted that the RPT does not ask for a priori assumptions about the probability distributions. There is no any requirement on sample sizes.

The permutation test evaluates the validity of isolated regularities. But in tasks where the number of tested data patterns is large multiple testing problem exists: the probability that a really random patterns will be incidentally identified as a valid regularity is also large. For example in tasks with a large number of explanatory variables the probability that an empirical distribution for at least one variable occasionally corresponds to a valid one-dimensional regularity is high.

Usually a Bonferroni correction [4] is used to receive more realistic validity estimates: the calculated p-values are multiplied by the number of tested explanatory variables. The multiple testing problem is more difficult when two-dimensional regularities are searched because it is necessary to multiply the calculated p-values by a very large number of tested pairs of variables. The Bonferroni correction is too pessimistic when it is applied to two-dimensional regularities search. An alternative approach was discussed in [4]. Two dimensional regularities found in a true data set were compared with two-dimensional regularities found in random samples that were obtained from the true data set by target permutation. Such procedure is more correct than the inferences got from the standard Bonferroni procedure.

3 Results

In this paper we present new results of recent researches. A statistically significant ($p = 0.014$) relationship was found to hold between institutional type and the output elasticity of the capital (OEC) parameters in the Cobb-Douglas production function [5]. This regularity is shown in figure 1. It is seen that the OEC is less than the threshold 0.6813 for 8 X countries (Russia, Japan, Republic of Korea, Phillipines, Laos, Nepal, Brazil, Myanmar) and only for one Y country – Denmark. On the contrary, the OEC is larger than the threshold for 12 Y countries and only for 4 X countries. The OEC for Mexico and Peru are extremely high.

From the figure 2 a two dimensional significant regularity follows which describes the relationship between the institutional type and two variables: the gross fixed capital formation (GFCF) ($p < 0.0005$) and the railway density ($p = 0.0273$). The quadrant IV corresponding to low railway density and relatively low GFCF contains 9 X countries, the quadrant III corresponding to high railway density and relatively low GFCF contains all the 13 countries of Y type, the quadrant II corresponding to high railway density and high GFCF contain 3 countries of X type: Japan, China, Republic of Korea.

The existence of regularities shown in figures 1 and 2 show that the parameters of the production functions or GFCF may be related to the dominating type of the institutional matrices. Data from [6] was used.

Acknowledgement

The work was supported by RFBR grant 17-02-00207.

References

[1] S. Kirdina, Journal of Economic Issues **47**, 309–322 (2014)
[2] S. Kirdina-Chandler, Journal of Economic Issues **51**, 476–485 (2017)
[3] I. L. Kirilyuk, A. I. Volynsky, M. S. Kruglova, A. V. Kuznetsova, A. A. Rubinstein, and O. V. Senko, Computer Research and Modeling **51**, 923–939 (2015) [in Russian]
[4] I. L. Kirilyuk, A. V. Kuznetsova, O. V. Senko, and A. M. Morozov, Pattern recognition and image analysis **27**, 94–104 (2015)
[5] C. W. Cobb and P. H. Douglas, American Economic Review **51** (18), 139–165 (1928)
[6] R. C. Feenstra, R. Inklaar, and M. P. Timmer, American economic review **105** (10), 3150–3182 (2015)