Models and algorithms for human capital reproduction intellectual analysis

Igor Kartsan[1,2,*], *Aleksandr* Zhukov[3,4], and *Sergey* Pronichkin[3,5]

[1]Marine Hydrophysical Institute, Russian Academy of Sciences, 2, Kapitanskaya str., 299011 Sevastopol, Russia
[2]Reshetnev Siberian State University of Science and Technology, 31, Krasnoiarskii Rabochii prospekt, 660037 Krasnoyarsk, Russia
[3]Expert and Analytical Center, 33, Talalikhina str., 109316 Moscow, Russia
[4]Institute of Astronomy of the Russian Academy of Sciences, 48, Pyatnitskaya str., 119017 Moscow, Russia
[5]Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, 40, Vavilov str., 119333 Moscow, Russia

Abstract. The managerial decisions making tasks in human capital reproduction complex systems are solved on the basis of models built on experimental data. It is problematic to take into account all the factors affecting the human capital reproduction. Existing approaches are not focused on building models for the human capital reproduction with incomplete information. Algorithms for inductive modeling are developed for the human capital reproduction systems characteristics functional description. The software is developed to implement the proposed algorithms for the human capital reproduction intellectual analysis based on the metric spaces of multisets.

1 Introduction

Currently the main sustainable development factor is human capital, which not only affects materialized capital, but also manages it. This is explained by the fact that in recent years the economy has turned into an information system, where the main aspect of its competitiveness is not fixed assets, but human abilities, skills and competence [1, 2]. As human capital accumulates, marginal benefits decrease and marginal costs increase. Therefore, it is necessary not only to form, but also to reproduce human capital.

Human capital reproduction is the formation of a person's productive abilities through investments in specific processes of an individual's activity - namely, in education and health promotion, which contribute to the human capital development. The whole process of human capital reproduction consists in a gradual transition from one phase to another:

- formation, where the certain knowledge accumulation directly takes place, which later a person uses in the production process;
- distribution, when a person begins to use the accumulated knowledge in certain areas and sectors of the economy;

* Corresponding author: kartsan2003@mail.ru

- exchange, where a certain intellectual base of any business entity is exchanged for remuneration for its activities;
- consumption, where, firstly, there is a productive use of human capital, and secondly, the basis for its further improvement is formed.

Human capital reproduction different phases necessitate taking into account a multitude of accompanying processes. The expected results from the use of certain means can be unpredictable as a result of the action of random external factors. When external factors are strictly defined or known, then the uncertainty can be taken into account and, accordingly, it is possible to propose ways to handle them. In the tasks of human capital reproduction system analysis, there are three main types of uncertainty: uncertainty of goals; situational uncertainty; informational uncertainty.

One of the human capital reproduction information uncertainty features is the uncertainty caused by the incompleteness of the data. There is a need to recover missing data based on the intelligent algorithms selection by which they will be recovered. This task is important for processing small sample sizes, when an incorrect assessment of the human capital reproduction is highly undesirable and can lead to errors in the predictive models construction.

2 Materials and methods

Human capital reproduction data analysis models. The study of the human capital reproduction processes using mathematical models allows to investigate the quantitative relationships between the input and output variables, as well as the factors affecting the output variables. This allows to study the processes behavior at any time intervals. Mathematical models used for these purposes should take into account the peculiarities of the interaction of quantitative and qualitative variables with possible consideration of real time based on simulation [3, 4].

The choice of the mathematical model structure is not an easy task that must be solved interactively. First, the model structure is estimated approximately based on the patterns study, correlation functions analysis, as well as visual data analysis. In this case, several of the most probable structures are selected. Then the candidate models parameters estimates are calculated and the optimal ones are selected using the corresponding statistical characteristics of the models quality. Different methods and approaches can be used to predict incomplete data, depending on the reasons for these uncertainties.

Currently, many exhaustive and iterative models have been developed for analyzing human capital reproduction data [5, 6]. Enumerated models are effective as a means of structural identification, but only with a limited number of arguments. Iterative approaches are computationally efficient with a large number of arguments, but the specificity of their architecture does not guarantee the construction of an adequate structure model to fill in the gaps in data on the human capital reproduction.

The use of any means to fill the gaps can bias the sample design that will be obtained from existing incomplete data, which can distort the real distribution of observations in the sample and reduce the actual significance of the results obtained.

When choosing a specific model to fill in the gaps, one should take into account the possibilities of its application, which significantly depend on the data analysis method that is supposed to be used in the future. There are various approaches to handle missing information, such as EM-estimation, Hot Deck, Zet, Barlet's method, Resampling, ZET braid [7, 8].

Expectation maximization estimation is an iterative procedure designed to solve optimization problems for a certain functional through an analytical search for the extremum of a function. It allows not only to reproduce missing values using a two-step iterative

algorithm interface, but also to estimate mean, covariance and correlation matrices for quantitative variables.

The Hot Deck method uses substitution instead of the missing value of the closest info item. The missing data can be selected both from the entire set of complete observations, and from some subgroup (cluster) to which the target object belongs. To fill the gap for the selected characteristic of the target object, the value of this characteristic is used for the object closest to the target. The type of distance function for determining the missing value is selected based on the type of data being studied, as well as ideas about the nature of the relationship between the variables.

The ZET method consists in selecting each value to fill the gap not over the entire set of observations, but from some part of it, which is called the component matrix, made up of component rows and columns. The componentness of a certain line is a value inversely proportional to the Cartesian distance along the target line (incomplete observation with a gap) in the space, the axes of which are the variables (the characteristics of objects) [9, 10]. Based on the component matrix data, in the future, a functional dependence of the predicted value on the corresponding value in the component matrix is built, on the basis of which the value of the missing data is then predicted.

Bartlett's method consists of two stages: substitution instead of skipping the initial generated values in the first stage; at the second stage, the target variable covariance analysis and the construction of a dichotomous indicator of the observations completeness for the target variable.

Resampling is an iterative method that involves changing rows with missing data with randomly selected rows from a full observations matrix, and then constructing a regression equation to predict the missing value. Regression modeling procedures are repeated several times, after which the values of the obtained regression coefficients are averaged and the final value is obtained, which gives the missing value maximum forecast accuracy [11, 12].

The ZET Braid method contains a mechanism for objective selection of the competent matrix dimension. A sequential selection of competent rows and columns is carried out, and each time a new competent matrix is formed. Then, according to a given criterion, its effectiveness in predicting gaps is determined [13, 14].

It is proposed to consider each of these methods as the opinion of a separate expert. It is proposed to use multisets as a mathematical model for the expert assessments presentation [15, 16].

3 Results

Algorithms for human capital reproduction data analysis/ The tasks of managerial decisions making in human capital reproduction complex systems are solved on the basis of models built on experimental data. Problematic is both the accounting of all factors affecting the human capital reproduction in specific conditions and the complexity of collecting reliable information [17].

There are many algorithms [18-22] that are used in problems of human capital reproduction modeling, but not all of them are focused on building models of complex systems in conditions of incomplete information. It is advisable to use the methods and tools of inductive modeling, designed primarily for the functional description of the systems characteristics for the human capital reproduction.

Iterative algorithms have been developed that solve the problem of constructing models based on a sample data $V = (X\ k)$ of n observations of m input (Expectation maximization, Hot Deck, ZET, Bartlett's method, Resampling, ZET Braid) and one output variable k is multiplicity of the multiset. The following types of algorithms for the intellectual analysis of the human capital reproduction have been implemented.

A multi-row algorithm, where in the process of calculations at each iteration (selection series) intermediate partial models are formed from all possible pairs of arguments, where L are the best outputs of the previous series t:

$$k_s^{t+1} = f(k_i^t, k_j^t),\ i = 1, \ldots, L,\ j = i+1, \ldots, L. \quad (1)$$

Relaxation algorithm, in which pairs are formed from intermediate and initial arguments:

$$k_s^{t+1} = f(k_i^t, x_j). \quad (2)$$

A combined algorithm, where pairs are formed from both intermediate and initial arguments, that is, it combines the two previous algorithms:

$$k_s^{t+1} = f(k_i^t, k_j^t) \vee f(k_i^t, x_j). \quad (3)$$

A generalized algorithm, where pairs are formed as in the combined algorithm, and combinatorial optimization of the complexity of all private models is also applied:

$$k_s^{t+1} = f^*(k_i^t, k_j^t) \vee f^*(k_i^t, x_j). \quad (4)$$

The three previous types of algorithms are special cases of the generalized one. In this algorithm, combinatorial optimization of the particular models complexity consists in the fact that on each row, models of the following form are considered:

$$f(k_i^t, k_j^t) = a_0 b_1 + a_1 b_2 k_i^{t-1} + a_2 b_3 k_j^{t-1}, \quad (5)$$

where b_i are the elements of the binary structure vector b taking the value 1 or 0, depending on the inclusion or exclusion of the corresponding argument. This optimizes the particular model complexity. All algorithms look for the optimal model as a solution to the optimization problem:

$$f^* = arg \min_{f \in \Phi} r(k, f(X, \hat{p}_f)), \quad (6)$$

where r is the regularity criterion $\left\| k_{V_2} - \hat{k}_{V_2|V_1} \right\| = \left\| k_{V_2} - X_{V_2} \hat{p}_{V_1} \right\|$, which is based on dividing the sample V into two parts V_1 and V_2 with the volume n_{V_1} and n_{V_2}, $n_{V_1} + n_{V_2} = n$, $\hat{p}_{V_1}$ is the estimate of the parameters on the subsample V_1.

A client-server software has been developed that implements the proposed algorithms for the intellectual analysis of the human capital reproduction. Client-server interaction in the software package determines the functional distribution between the client and server parts into the so-called "operational levels":

- user interface responsible for data presentation and timely response to user commands;
- server that is responsible both for the level at which the information received from the user is processed and for the human capital reproduction data management level, ensuring their storage and access to them.

The software package consists of several blocks: data storage unit, which stores both input and output data and intermediate results; task formation block, in which control parameters

are set; block for solving the problem, in which the modeling process can be performed in three modes (two automatic and one interactive).

Using the data storage unit, having an initial sample of human capital reproduction data, it is possible to split data into projects, store intermediate calculations for further continuation of the modeling process, save the final calculation results and use the resulting models for new data.

The system operates simultaneously with three databases: the initial database, the calculation database and the results database. After the initial sample is generated or obtained, the block for generating the problem is used. At this stage, the sample is divided into two parts - training and testing: the model coefficients are estimated on the training sample, the best models of human capital reproduction are selected on the test sample based on the regularity criterion. In the process of generating a data sample, it is possible to set: the type of sample splitting, the noise level, the external criterion, the modeling algorithm. Further, depending on the use of various modifications of the algorithm (1) - (5), different complexity models are generated, for each of which the criterion value is calculated by which they are selected for the next series.

In the software package, the process of modeling the human capital reproduction can be implemented in three modes (two automatic and one interactive).

Automatic (the process of self-organization of models is performed automatically) the mode is implemented in two versions: standard - the same type of private description is set for all rows without exception; planned - the process of self-organization of models is performed automatically according to a given plan, that is, when the type of private description is set different for the series.

Interactive when it is possible to directly participate in the models self-organization process:

- include or not include modifications on any row;
- change the complexity of private description models;
- choose a different number of models that will move to the next row;
- use different criteria for selecting the best models.

The developed program interface is a set of tools through which the user can control the modeling process. The following interface features have been implemented: the self-organization process can be stopped at any stage of the calculation, and then at any time the calculations can be extended, while all intermediate calculations will be saved; on any row, it is possible to include or not to include different modifications, change the complexity of models of a private description, choose a different number of models that will go to the next row, change the criteria for choosing the best models.

The database stores raw data, calculation data, and results data. Having an initial sample, it is possible to split the data into projects, store intermediate calculations for further continuation of the modeling process, save the calculation results, and also use the resulting models on new data.

Depending on the use of various optimization options, models of different complexity are generated, for each of which the criterion value is calculated by which they are selected for the next series. Such a structural arrangement allows experimenting with the input data at each stage, thereby interactively changing the algorithm structure for analyzing data on the human capital reproduction.

4 Discussion

In the tasks of intelligent processing of human capital reproduction data, the greatest difficulty remains the need to classify uncertainties of different types and the resulting gaps and imprecise values. Efficient algorithms are needed for handling uncertainties and

associated missing values that are specific to this area. The main goal of the analysts' work is precisely the identification and development of such management decisions that may be typical for solving various problems of increasing the efficiency of human capital reproduction.

A step-by-step solution to the problem of filling in the missing data of the human capital reproduction involves analyzing the essence of the process described by a certain sequence of data, selecting the structure of the model, choosing adequate data mining methods to fill in the missing data, and implementing these methods with modern tools.

5 Conclusions

The developed software allows working with different data sets, performing planned computational experiments and solving practical problems of intellectual analysis of human capital reproduction. The constructed models are provided by the system for graphical and meaningful analysis and are stored in the database for further use. The information system has been implemented, in which the modeling process can be performed in automatic and interactive modes.

Acknowledgments

The work was carried out within the framework of the state task of the Ministry of Education and Science of the Russian Federation on the topic "Conceptual modeling of the information and educational environment for the reproduction of human capital in the digital economy" № 121102600069-2 (code FNRN – E).

References

1. M. Wang, M. Xu, S. Ma, Structural Change and Economic Dynamics **59**, 427–441 (2021) DOI: 10.1016/j.strueco.2021.09.018
2. S. Managi, M. Jimichi, C. Saka, Economic Analysis and Policy **72**, 268–275 (2021) DOI: 10.1016/j.eap.2021.08.013
3. S. Bosi, T. Lloyd, K, Nishimura, Mathematical Social Sciences **112**, 145–158 (2021) DOI: 10.1016/j.mathsocsci.2021.03.013
4. L. Zhang, D. Godil, M. Anser, Science of The Total Environment **774**, 145553 (2021) DOI: 10.1016/j.scitotenv.2021.145553
5. I.V. Kovalev, A.S. Andronov, I.N. Kartsan, M.V. Karaseva, IOP Conference Series: Materials Science and Engineering **862(4)**, 042055 (2020) DOI: 10.1088/1757-899X/862/4/042055
6. U. Pata, A, Caglar, Energy **216**, 119220 (2020) DOI: 10.1016/j.energy.2020.119220
7. Z. Hou, M. Jin, S, Kumbhakar, European Journal of Operational Research **287**, 317–330 (2020) DOI: 10.1016/j.ejor.2020.04.039
8. P. Conti, F. Mecatti, F. Nicolussi, Computational Statistics & Data Analysis **167**, 107366 (2022) DOI: 10.1016/j.csda.2021.107366
9. D. Sullivan, R. Andridge, Computational Statistics & Data Analysis **82**, 173–185 (2015) DOI: 10.1016/j.csda.2014.09.008
10. R. Krause, M. Huisman, C. Steglich et al, Social Networks **62**, 99–112 (2020)

11. H. Lin, F. Zhou, Q. Wang et al, Journal of Econometrics **205**, 363–380 (2018) DOI: 10.1016/j.jeconom.2018.03.017
12. S.V. Efremova, I.N. Kartsan, A.O. Zhukov, IOP Conference Series: Materials Science and Engineering **1047(1)**, 012068 (2021) DOI: 10.1088/1757-899X/1047/1/012068
13. J. Xiao, Y. Wang, J. Chen et al, Information Sciences **569**, 508–526 (2021)
14. S. Arciniegas, M. Pena, P. Rodrigues, Computers and Electronics in Agriculture **176**, 105617 (2020) DOI: 10.1016/j.compag.2020.105617
15. M.G. Semenenko, I.V. Kniazeva, L.S. Beckel et al, IOP Conference Series: Materials Science and Engineering **537(3)**, 032095. (2019) DOI: 10.1088/1757-899X/537/3/032095
16. S. Faisal, G. Tutz, Information Sciences **570**, 500–516 (2021) DOI: 10.1016/j.ins.2021.04.009
17. R. Gopalakrishnan, C. Guevara, M. Akiva, Transportation Research Part B **142**, 45–57 (2020) DOI: 10.1016/j.trb.2020.10.002
18. J. Li, X. Yu, Discrete Mathematics **344**, 112487 (2021) DOI: 10.1016/j.disc.2021.112487
19. Z. Lin, J. Ma, Y. Zhou et al, Advances in Applied Mathematics **129**, 102206 (2021) DOI: 10.1016/j.aam.2021.102206
20. M. Rafi, M. Naseef, S. Prasad, Energy Economics **101**, 105427 (2021) DOI: 10.1016/j.eneco.2021.105427
21. M. Gillman, Economic Modelling **99**, 105470 (2021) DOI: 10.1016/j.econmod.2021.02.011
22. M. Alvarez, H. Strulik, Journal of Economic Behavior & Organization **181**, 211–240 (2020) DOI: 10.1016/j.jebo.2020.11.034