CONTENTS

Modeling of social and economic systems
T.K. Bogdanova, A.R. Kamalova, T.K. Kravchenko, A.I. Poltorak
Problems of modeling the valuation of residential properties........ 7

Zh. N. Zenkova, W. Musoni
The economic order quantity taking into account additional information about the known quantile of the cumulative distribution function of the product’s sales volume........................................... 24

Yu. F. Telnov, V.A. Kazakov, V.M. Trembach
Developing a knowledge-based system for the design of innovative product creation processes for network enterprises......................................................... 35

Decision making and business intelligence
V.N. Kazmin, A.B. Menisov, I.A. Shastun
An approach to identifying bots in social networks based on the special association of classifiers ................................. 54

Data analysis and intelligence systems
Yu. A. Zelenkov, E.V. Lashkevich
Fuzzy regression model of the impact of technology on living standards ............................................................... 67

Information systems and technologies in business
R.A. Zhukov
An approach to assessing the functioning of hierarchical socio-economic systems and decision-making based on the EFRA software package .............................................. 82
Business Informatics is a peer reviewed interdisciplinary academic journal published since 2007 by National Research University Higher School of Economics (HSE), Moscow, Russian Federation. The journal is administered by HSE Graduate School of Business. The journal is published quarterly.

The mission of the journal is to develop business informatics as a new field within both information technologies and management. It provides dissemination of latest technical and methodological developments, promotes new competences and provides a framework for discussion in the field of application of modern IT solutions in business, management and economics.

The journal publishes papers in the areas of, but not limited to:

- data analysis and intelligence systems
- information systems and technologies in business
- mathematical methods and algorithms of business informatics
- software engineering
- internet technologies
- business processes modeling and analysis
- standardization, certification, quality, innovations
- legal aspects of business informatics
- decision making and business intelligence
- modeling of social and economic systems
- information security.

The journal is included into the list of peer reviewed scientific editions established by the Supreme Certification Commission of the Russian Federation.

The journal is included into Web of Science Emerging Sources Citation Index (WoS ESCI) and Russian Science Citation Index on the Web of Science platform (RSCI).


Editor-in-Chief: Dr. Yevgeni A. Koucheryavy.
ABOUT THE HIGHER SCHOOL OF ECONOMICS

Consistently ranked as one of Russia’s top universities, the Higher School of Economics (HSE) is a leader in Russian education and one of the preeminent economics and social sciences universities in Eastern Europe and Eurasia. Having rapidly grown into a well-renowned research university over two decades, HSE sets itself apart with its international presence and cooperation.

Our faculty, researchers, and students represent over 50 countries, and are dedicated to maintaining the highest academic standards. Our newly adopted structural reforms support both HSE’s drive to internationalize and the groundbreaking research of our faculty, researchers, and students.

Now a dynamic university with four campuses, HSE is a leader in combining Russian educational traditions with the best international teaching and research practices. HSE offers outstanding educational programs from secondary school to doctoral studies, with top departments and research centers in a number of international fields.

Since 2013, HSE has been a member of the 5-100 Russian Academic Excellence Project, a highly selective government program aimed at boosting the international competitiveness of Russian universities.
The School of Business Informatics is one of the leading divisions of HSE’s Faculty of Business and Management. The School offers students diverse courses taught by full-time HSE instructors and invited business practitioners. Students are also given the opportunity to carry out fundamental and applied projects at various academic centers and laboratories.

Within the undergraduate program, students participate each year in different case-competitions (PWC, E&Y, Deloitte, Cisco, Google, CIMA, Microsoft Imagine CUP, IBM Smarter Planet, GMC etc.) and some of them are usually as being best students by IBM, Microsoft, SAP, etc. Students also have an opportunity to participate in exchange programs with the University of Passau, the University of Munster, the University of Business and Economics in Vienna, the Seoul National University of Science and Technology, the Radbound University Nijmegen and various summer schools (Hong Kong, Israel etc.). Graduates successfully continue their studies in Russia and abroad, start their own businesses and are employed in high-skilled positions in IT companies.

There are four graduate programs provided by the School:

- Business Informatics
- E-Business;
- Information Security Management;
- Big Data Systems.

The School’s activities are aimed at achieving greater integration into the global education and research community. A member of the European Research Center for Information Systems (ERCIS), the School cooperates with leading universities and research institutions around the world through academic exchange programs and participation in international educational and research projects.
Problems of modeling the valuation of residential properties

Tatiana K. Bogdanova
E-mail: tanbog@hse.ru

Albina R. Kamalova
E-mail: akamalova@hse.ru

Tatiana K. Kravchenko
E-mail: tkravchenko@hse.ru

Anastasia I. Poltorak
E-mail: poltoraknasty@gmail.com

Abstract
The solution of the housing problem for many decades has been and remains one of the most important tasks facing the nation. The problem of modeling the value of residential properties is becoming more and more urgent, since a high-quality forecast makes it possible to reduce risks, both for government bodies and for realtors specializing in the purchase and sale of residential properties, as well as for ordinary citizens who buy or sell apartments. Predictive models allow us to get an adequate assessment of both the current and future situation on the residential property market, to identify trends in the cost of housing and the factors influencing these changes. This involves both the qualitative characteristics of the particular property, and the general condition and the dynamics of the real estate market. Russia is characterized by significant differences in the level of development of regions, therefore, by differences in trends of supply and demand prices for real estate. Valuation of residential properties at the regional level is particularly important, since all of the above determines the social and economic importance of this problem. This article presents a comprehensive model for estimating the value of residential properties in the secondary housing market of Moscow using decision tree methods and ordinal logistic regression. A predictive model of the level of housing comfort was built using the CRT decision tree method. The results of this forecast are used as input information for an ordinal logistic regression model for estimating the value of residential properties in the secondary market of Moscow. Testing the model on real data showed the high predictive ability of the model we generated.
Introduction

The residential property market plays an important role in the development of the social and economic spheres of the country. An unfavorable situation on the real estate market may lead to crises such as the subprime crisis in 2008, when there was a “collapse” in the real estate market due to the sharp increase in delinquencies / defaults on mortgages and mass alienation of property for the benefit of creditors.

According to paragraph 1 of Article 16 of the Housing Code of the Russian Federation, residential properties include: a residential building, a part of a residential building, an apartment, a part of an apartment, a room [1]. At the same time, the residential property market can be conditionally divided into the primary and secondary housing market. The primary market is real estate that appears on the market as a commodity for the first time, and that was not previously owned by someone. The secondary market is real estate that was already used and was in private or municipal ownership. It is important to note that the primary and secondary markets are inextricably linked, since an increase in supply in the secondary market inevitably causes a decline in demand and, consequently, decline in prices in the primary market. If construction costs rise, then prices for residential properties in the primary market, which determine the volume of real estate created, will also rise. In turn, this will lead to an increase in prices in the secondary housing market. The residential property market has low price elasticity.

In the primary property market, property is sold directly by developers or agents. At the same time, it is possible to purchase at different stages of construction, including by agreement on a cost-sharing arrangement. In the secondary market, the buyer always buys a completed apartment. The market value of a real estate property depends on various factors, such as external factors (changes in the procedure for obtaining a building permit, the key rate, conditions for mortgage lending, limitations on price increases, tax incentives, licensing of developers and agents, the economic situation in the country / region, demand for housing etc.) and internal factors (the residential property’s surface area, number of rooms, wall material, floor, state of repair, balcony, parking, distance from transport, metro, distance to the city center, etc.).

Experts (appraisers) for real estate valuations in practice mainly use traditional methods: the comparative method, the cost method and the income method. In general, the estimates obtained by these methods may vary. In addition, in most cases a limited set of factors is taken into account, and the final price of a given property is ultimately determined by an expert who is guided by his own experience, the current situation in the real estate market, and available information on the impact of macroeconomic and demographic factors. However, the determination of the objective value of real estate is necessary in a variety of cases: for the purposes of purchase and sale, for tax purposes, for mortgage lending when transferring property as collateral and setting the loan amount by banks, when considering property

Key words: model; value estimation; residential properties; secondary housing market; ordinal logistic regression; decision tree; CRT method; value forecast.

DOI: 10.17323/2587-814X.2020.3.7.23
disputes, for determining the price of insurance by insurance companies, as well as making investment decisions.

In this regard, the problem of the accuracy and objectivity of estimating the value of residential properties becomes more and more urgent, including the issues of modeling the price of residential real estate, the study of factors affecting the cost, and the use of analytical methods of valuation.

In Russia, since the 1990s, a series of scientific works has been published devoted to the development of economic and mathematical methods of real estate valuation [2–4]. The classification proposed in [5] is fundamental for many studies. This paper presents the pricing in the housing market in Russia and highlights the pricing factors, which are divided by levels in relation to supply and demand. In [6–10], methods of neural network modeling were used to estimate the value of real estate.

The relevance of estimating the value of real estate is also noted in the works of several foreign authors [11–14].

Analysis of publications by foreign authors in the Scopus database for the period from 2000 to 2020 (Figure 1) also confirms the relevance of this issue, both for the Russian and foreign residential property markets.

As shown in Figure 1, interest in this issue has grown significantly in recent years. If in 2002 only one work was published, then 15 and 16 works, respectively, were devoted to estimating the value of real estate in 2018 and 2019.

A significant share of the papers is devoted to the analysis and comparison of various methods for estimating the value of residential properties, both from a theoretical and practical point of view. Some authors focused their attention on the problem of estimating the impact of the value of residential properties on the mortgage crisis of 2008 in the context of mortgage lending [15]. The applicability of the concept of “sustainable real estate” in Poland and the issues of valuation of such real estate are considered in [16]. The authors of the work [17] studied the behavior of appraisers and buyers in the residential property market using a survey.

\[\text{Scopus}: \text{http://scopus.com}\]
method. In studies [17–19], an expert system was developed to estimate the value of residential properties using decision support methods.

Many authors proposed their own models based on analytical methods for estimating the value of residential properties and tested them on real data. At the same time, the models had differences, both in terms of the methods used and the variables.

In [20], a multiple regression model was proposed that uses the tax rate, floors, number of bathrooms, total area, living space, number of garages, number of rooms, number of bedrooms, age of the house as independent variables. Work [21] estimated the influence of such factors as distance to the city center, availability of parking, ecology, floor on which the apartment is located, total area, noise, maintenance costs, availability of central heating, proximity to kindergartens and schools, wall material, age of the house, proximity to cultural and historical sites, the presence of a balcony or loggia.

A model for estimating the cost of apartments and residential properties, as a combination of an artificial neural network model and a geographic information system, based on data on the cost of residential properties in Albacete for 2002, was presented in [22]. The explanatory variables were the type of housing, location, age of the house, number of bedrooms, the presence of an elevator, a balcony, a parking space, a heating system, condition of the property and distance to the city center. In this paper, one of the most important variables was the distance to the central business center.

A significant number of publications consider the theoretical aspects of modeling the valuation of residential properties. They are devoted to the generalization of the models used, the analysis of the methods used, the purposes of the valuation and other characteristics of the spatial structure that determine the category of the proposed models. In work [23] models of real estate appraisal are analyzed depending on the purposes of the valuation (sale and purchase of property, mergers and acquisitions, taxation, mortgage lending). In [24], the models of real estate valuation are considered in the context of the methods used, which can be divided into two groups: traditional and advanced (analytical).

The traditional methods for estimating the value of residential properties [24–26] include the income method, the comparative method, the investment method and the cost method. Analytical methods include hedonic price models (HPM), artificial neural network (ANN), expert system (ES), fuzzy logic system (FLS), spatial analysis methods and Autoregressive Integrated Moving Average (ARIMA). At the same time, the ANN, ES and FLS methods are considered by the authors as estimation methods based on artificial intelligence technology.

In the study [27], the authors proposed a classification of analytical models for estimating the value of real estate and divided them into the following: models based on artificial intelligence methods, models based on a geographic information system and mixed models.

The purposes of this work are the following:

- to propose a spatial structure of residential property valuation models that allows classifying models from different points of view, namely: the estimated value, modeling methods, the number of valuation objects, modeling objectives, the modeling object, the type of object, the real estate market, the subject of the assessment and the level of modeling;
- to provide a classification of factors affecting the cost of one square meter of residential property;

---

2 Preliminary results of the study were presented in the graduate work by Anastasia I. Poltorak performed at the HSE Faculty of Business and Management in 2019
to develop a comprehensive model for estimating the value of residential properties in the secondary housing market in Moscow using the decision tree method to predict the level of comfort in housing and ordinal logistic regression to predict the estimation of the value of residential properties, using the results of the model for predicting the level of comfort in housing as input information;

to test the developed complex model on real data and estimate its predictive ability.

1. Spatial structure of residential property valuation models

The usually proposed classification of models assumes their division depending on the valuation methods used. Summarizing the above, Figure 2 represents the spatial structure of models for estimating the value of residential properties depending on various characteristics, which include:

- estimated value (models for assessing market, residual, cadastral, investment, liquidation, book value);
- modeling methods (models based on traditional and analytical methods);
- number of objects (models of individual and mass appraisal of the value of residential properties);
- the goals of modeling (models for prices forecasting, analyzing pricing factors, managing the value of residential properties, assessing the impact of external factors, taxation);
- object of modeling (models for evaluating one square meter of residential property, residential property in general);
- type of property (models for estimating residential buildings, apartments, rooms);
- type of market (models for evaluating residential properties in the primary and secondary markets);
- subject of appraisal (models of appraisal of residential properties for a seller, buyer, investor);
- level of modeling (models for evaluating residential properties at the regional level, federal level, at the level of a district, county, residential property).

2. Analysis of the residential property market in Moscow

According to Rosstat, among the subjects of the Russian Federation in 2019, the largest volumes of housing construction were carried out in the Moscow region and Moscow, where, respectively, 10.5% and 6.3% of the total area of housing commissioned in Russia was located. According to the Bank of Russia, the largest volume of loans in the total volume of housing loans in the Russian Federation was issued in Moscow (13.8%) and in the Moscow region (8.5%). At the same time, according to preliminary data, the real disposable cash income of the general population of Russia in 2019 increased by 1.0% compared to 2018.

The weighted average interest rate on residential mortgage lending, according to the Bank of Russia, decreased from 9.87% in January to 9.00% in December 2019. In 2020, the downward trend in mortgage lending rates continued. In addition, in the period from April 17 to November 1, 2020, a preferential mortgage was introduced for the purchase of an apartment in new buildings with a rate of no more than 6.5%, while the loan amount is limited (up to 8 million rubles in Moscow and the Moscow region) [28].

---

4 Central Bank of the Russian Federation: https://www.cbr.ru/
5 Central Bank of the Russian Federation: https://www.cbr.ru/
Fig. 2. Spatial structure of residential properties valuation models
The residential property market in Moscow and the Moscow region is characterized by a wide variety of categories of properties, both in the primary and secondary markets. Depending on various factors, a certain price can be set for a particular property.

Among the significant variety of categories of residential properties, the majority in the market of Moscow and the Moscow region are residential properties located in apartment buildings. Therefore, to analyze the value of residential properties, it is appropriate to choose this category.

It should be noted that the cost of housing in Moscow and the Moscow region is influenced by various factors, which can be conditionally divided into environmental factors (exogenous) and internal factors (endogenous). The classification of factors affecting the cost of one square meter of an apartment is shown in Figure 3.

Among the exogenous factors, we can highlight the impact of factors of the world, country, and regional levels, and among the endogenous — factors that relate to the local level, and the factors of the given property itself.

The world-level factors affecting the cost of one square meter include the coronavirus pandemic, sanctions against Russia, changes in oil prices, exchange rates and other factors. Their influence is rather difficult to directly assess in value terms, but the indirect influence of these factors on the value of residential properties undoubtedly is felt. The impact of these factors is manifested, first, in a direct impact on the state of the country’s economy, which leads to changes in the residential property sector.

Country-level factors include changes in mortgage lending rates, the introduction of escrow accounts, taxation, inflation and benefits. The influence of these factors can be directly determined. For example, according to the analysis carried out by the staff of the Institute of Economic Forecasting of the Russian Academy of Sciences, as a result of a decrease in the lending rate from 10 to 6.5%, the value of the Housing Affordability Index (HAI) in the primary market will increase from 104.7% to 135.5% [29]. These measures were introduced by the state in order to support the construction sector and maintain the price level in the housing market at the expense of demand. The introduction of an escrow account mechanism, an increase in the VAT rate to 20% in 2019, and an increase in inflation (the level of which, according to the Central Bank of the Russian Federation, is expected to be in the range of 3.8—4.8% at the end of 2020) have a significant impact on the value of residential properties.

Regional factors affecting the value of residential properties include population size, income level of the population, climatic conditions, renovation, supply and demand.

The value of a property is also influenced by internal factors that directly characterize the property itself and its location in relation to infrastructure. Among such factors at the local level, we can identify transport accessibility, infrastructure development, district/county, environmental situation, the presence of a parking space, a guarded yard, a concierge, an elevator, etc. At the level of the building itself, we can highlight the factors that exert an influence, depending on whether this building is under construction or ready. Factors affecting the value of a secondary market property include year of construction, floor, total / living area, number of rooms, presence of a balcony, wall material, view from the window, ceiling height. The factors affecting the cost of the primary market building, in addition to the above factors, include the cost of construction, the stage of construction, the rating of the developer,

5 Central Bank of the Russian Federation: https://www.cbr.ru/
Factors affecting the cost of one square meter of residential property

**Exogenous**
- World level
  - Coronavirus pandemic
  - Sanctions
  - Oil price
  - Exchange rate
- Country level
  - Key rate
  - Introduction of escrow accounts
  - Inflation
  - Benefits
  - Taxation
- Regional level
  - Population
  - Employment rate
  - Population income level
  - Climatic conditions
  - Renovation
  - Supply and demand
- Local level
  - Transport accessibility
  - Infrastructure
  - District / County
  - Ecological situation
  - Parking
  - Guarded yard
  - Concierge
  - Influencing factors in the secondary market
- Object level
  - Primary building object
  - Construction stage
  - Developer rating
  - Construction cost
  - Marketing
  - Object in the secondary market
    - Year of construction
    - Floor
    - Total / living area
    - Number of rooms
    - Balcony
    - Wall material
    - Ceiling height
    - View from the window

**Endogenous**
- Primary building object
  - Construction stage
  - Developer rating
  - Construction cost
  - Marketing
- Object in the secondary market
  - Year of construction
  - Floor
  - Total / living area
  - Number of rooms
  - Balcony
  - Wall material
  - Ceiling height
  - View from the window

*Fig. 3. Proposed classification of factors affecting the cost of one square meter of residential property*
marketing. For example, the closer the delivery term, the more expensive the residential property is, and vice versa. When people buy residential properties at the construction stage, they pay special attention to the rating and experience of the developer, therefore, the cost of buildings with the same indicators from a more reliable developer is usually higher.

It should be noted that pricing in the primary and secondary residential property markets is significantly different. Therefore, the asset pricing model built for the primary market will not be relevant for the secondary market.

3. A comprehensive model for predicting the value of residential properties in the secondary market

The problem of predicting the value of real estate in the secondary housing market has been and remains relevant for several decades. A high-quality forecast helps to reduce risks, both for government and for realtors specializing in the purchase and sale of housing, as well as for ordinary citizens who buy or sell apartments. The developed complex model for predicting the value of residential properties in the secondary market includes two sub-models – a model for predicting the level of comfort of housing based on local data, and a model for predicting the value of a unit of residential property based on building factors and input variables that are the result of predicting the comfort model of housing.

3.1. Model for predicting the comfort level of residential properties in the secondary market in Moscow

To predict the level of comfort in residential properties, a sample of 304 observations was formed, including 14 factors taken from the CIAN database.

It is assumed that the dependent variable “comfort” can take three values: 1 – a minimum level of comfort; 2 – average level of comfort; 3 – a high level of housing comfort. The independent factors are: presence of an elevator, the year of construction, the district, the county, the presence of a garbage chute, presence of parking, “house for demolition,” the distance from the metro, the type of house (0 – panel; 1 – brick; 2 – block; 3 – monolithic), presence of a concierge, number of floors, distance from the Moscow Ring Road, number of entrances, assessment of infrastructure (1 – minimally developed infrastructure; 2 – moderately developed infrastructure: there are shops and pharmacies, but no entertainment centers, medical and educational institutions; 3 – the most developed infrastructure).

The frequency distribution of the sample values according to the comfort level categories is as follows: the highest number of sample elements fell into the category of high comfort level – 117, the average comfort level category – 101, and the minimum comfort level category – 86.

To predict the level of housing comfort in Moscow, 12 territorial-administrative districts values were randomly included in the sample and grouped into five enlarged districts according to the principle of the proximity of average housing prices (Table 1).

A new variable “district” was created, consisting of five combined districts and taking the following values:

- Central;
- North, East, South-West, North-West;
- North-East, South-East, South;
- Western;
- Novomoskovsky, Zelenogradsky, Troitsky.

7 CIAN: https://www.cian.ru/
The histogram of the frequency distribution of the values of the “okrug” variable in the sample is shown in Figure 4.

As can be seen from the histogram, the largest number of values fell into category 2, which is explained by the fact that this category covers four of the 12 territorial-administrative districts of Moscow at once.

To predict the level of comfort, the CRT decision tree method was used (tree depth – , number of nodes – 25). The significant variables include: “distance from the metro,” “infrastructure,” “distance from the Moscow Ring Road,” “age of the house,” “district,” “number of levels of the house,” “presence of an elevator,” “type of house,” “number of entrances,” “presence of a garbage chute,” “house for demolition,” “availability of parking” and “presence of a concierge.”

Table 1.

<table>
<thead>
<tr>
<th>No</th>
<th>District</th>
<th>Average price in rubles</th>
<th>Number of observations</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Central</td>
<td>20 634 899.40</td>
<td>25</td>
<td>16 110 812.30</td>
</tr>
<tr>
<td>2</td>
<td>West</td>
<td>13 425 600.00</td>
<td>25</td>
<td>6 336 071.39</td>
</tr>
<tr>
<td>3</td>
<td>East</td>
<td>10 571 250.00</td>
<td>24</td>
<td>7 849 054.95</td>
</tr>
<tr>
<td>4</td>
<td>North</td>
<td>10 417 986.20</td>
<td>25</td>
<td>6 341 765.62</td>
</tr>
<tr>
<td>5</td>
<td>South–West</td>
<td>9 797 307.69</td>
<td>26</td>
<td>4 014 771.39</td>
</tr>
<tr>
<td>6</td>
<td>North–West</td>
<td>9 471 791.92</td>
<td>24</td>
<td>4 063 054.40</td>
</tr>
<tr>
<td>7</td>
<td>South</td>
<td>8 560 000.00</td>
<td>25</td>
<td>2 340 539.47</td>
</tr>
<tr>
<td>8</td>
<td>North–East</td>
<td>8 091 280.00</td>
<td>25</td>
<td>2 754 360.66</td>
</tr>
<tr>
<td>9</td>
<td>South–East</td>
<td>7 535 199.96</td>
<td>25</td>
<td>2 385 758.27</td>
</tr>
<tr>
<td>10</td>
<td>Novomoskovsky</td>
<td>6 437 160.00</td>
<td>25</td>
<td>2 232 530.14</td>
</tr>
<tr>
<td>11</td>
<td>Zelenogradsky</td>
<td>6 124 800.00</td>
<td>25</td>
<td>2 243 342.22</td>
</tr>
<tr>
<td>12</td>
<td>Troitsk</td>
<td>4 287 666.67</td>
<td>30</td>
<td>1 103 601.26</td>
</tr>
<tr>
<td></td>
<td><strong>Average total</strong></td>
<td><strong>9 612 911.82</strong></td>
<td><strong>304</strong></td>
<td><strong>4 814 638.51</strong></td>
</tr>
</tbody>
</table>

Fig. 4. The structure of the variable “district”
As the analysis showed, buildings with the maximum level of comfort, first of all, are characterized by proximity to the metro—the division into groups occurs at a level of more or less 16 minutes to the metro. Node 1 got 109 values from the sample corresponding to the maximum comfort level, which amounted to 66% of the values of node 1. At the same time, node 2 received only 8 values from the sample with the maximum comfort level, which is only 5.8% of the values from node 2.

Furthermore, there is a separation by distance from the Moscow Ring Road. If the property is near the metro, but outside the Moscow Ring Road, then, perhaps, its territorial location is not as attractive as buildings that are further from the metro, but closer to the center. At the same time, real estate that is not within walking distance of the metro is further divided according to the availability of parking, since their residents use their own vehicles. If there is a parking space, a greater number of node values belong to comfort groups 2 and 3, while in the absence of a parking space, the object will most likely belong to comfort group 1.

The classification of independent variables in terms of significance is shown in Figure 5. The share of correctly predicted values was 71.7%. The predicted values of the comfort level and the resulting nodes are used to build a predictive model for the comfort of residential properties.

3.2. Forecasting the cost of a unit of residential property in Moscow based on an ordinal logistic model

An ordinal logistic regression model was used to predict the cost of a unit of residential property in the secondary market of Moscow. All residential property prices were grouped into six price categories (Table 2).

### Table 2. Compliance with the category and price range of the apartment value

<table>
<thead>
<tr>
<th>Category number</th>
<th>Price range (RUB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>less than 5 500 000</td>
</tr>
<tr>
<td>2</td>
<td>from 5 500 001 to 6 450 000</td>
</tr>
<tr>
<td>3</td>
<td>from 6 450 001 to 7 500 000</td>
</tr>
<tr>
<td>4</td>
<td>from 7 500 001 to 8 950 000</td>
</tr>
<tr>
<td>5</td>
<td>from 8 950 001 to 13 490 000</td>
</tr>
<tr>
<td>6</td>
<td>more than 13 490 000</td>
</tr>
</tbody>
</table>

Analysis of the constructed model coefficients showed that the following variables are significant: “total area” (totalarea), “ceiling height” (ceilingheight), “number of rooms” (rooms), “repairs,” “balcony,” “district” and “predicted comfort” (Table 3).

The constructed model looks like:

\[ y = 0.133 \cdot \text{totalarea} + 1.492 \cdot \text{ceilingheight} - 8.738 \cdot \text{rooms} - 1.250 \cdot \text{repairs} - 0.586 \cdot \text{balcony} + 9.920 \cdot \text{district1} + 5.608 \cdot \text{district2} + 4.959 \cdot \text{district3} + 5.652 \cdot \text{district4} - 2.950 \cdot \text{comfort1} - 3.132 \cdot \text{comfort2}, \]

where district1 — Central District;
district2 — union of four districts: North, East, South-West, North-West;
district3 — the union of three districts: North-East, South-East, South;
district4 — Western District;
district5 — the union of three districts: Novomoskovsky, Zelenogradsky and Troitsky;
**Table 3.** Ordinal logistic model of the cost of 1 sq. m of residential property

<table>
<thead>
<tr>
<th>No</th>
<th>Variable</th>
<th>Estimation</th>
<th>Standard error</th>
<th>Wald</th>
<th>Degree of freedom</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>totalarea</td>
<td>0.133</td>
<td>0.021</td>
<td>41.602</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>ceilingheight</td>
<td>1.492</td>
<td>0.554</td>
<td>7.250</td>
<td>1</td>
<td>0.007</td>
</tr>
<tr>
<td>3</td>
<td>rooms</td>
<td>-8.738</td>
<td>0.374</td>
<td>546.982</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>repairs = 0</td>
<td>-1.250</td>
<td>0.289</td>
<td>18.73</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>repairs = 1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>balcony = 0</td>
<td>-0.586</td>
<td>0.307</td>
<td>3.652</td>
<td>1</td>
<td>0.056</td>
</tr>
<tr>
<td>7</td>
<td>balcony = 1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>district1</td>
<td>9.920</td>
<td>0.905</td>
<td>120.245</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>9</td>
<td>district2</td>
<td>5.608</td>
<td>0.507</td>
<td>122.387</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>10</td>
<td>district3</td>
<td>4.959</td>
<td>0.495</td>
<td>100.487</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>11</td>
<td>district4</td>
<td>5.652</td>
<td>0.654</td>
<td>74.702</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>12</td>
<td>district5</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>comfort1</td>
<td>-2.950</td>
<td>0.413</td>
<td>51.124</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>14</td>
<td>comfort2</td>
<td>-3.132</td>
<td>0.419</td>
<td>55.737</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>15</td>
<td>comfort3</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Connection function: logit

* This parameter is set to zero because it is redundant

**Fig. 5.** Classification of independent variables by significance

---

**MODELING OF SOCIAL AND ECONOMIC SYSTEMS**

---

**BUSINESS INFORMATICS** Vol. 14 No 3 – 2020
comfort1 — minimum housing comfort;
comfort2 — average housing comfort;
comfort3 — high housing comfort.

The variables “district” and “comfort” are dummy variables. In an equation, the number of dummy variables is always 1 less than the dummy variables themselves, since one of them is always the base variable. Thus, in this model, the variables district5 and comfort3 are selected as basic variables.

The quality of the constructed model is evidenced by a decrease in the log likelihood with included variables by almost two times, high values of the pseudo R-square of Cox and Snell (0.849) and Nigelkirk (0.873), and the proportion of correctly predicted values (71%).

To test the complex model we developed, a new sample of 30 apartments in the Moscow secondary market was collected, taken randomly from the CIAN database for 2019. Comparative data on approbation of the ordinal logistic model, including the actual and predicted category, as well as the average relative forecast error, are presented in Table 4.

As can be seen from table 4, the average forecast error was 10%. This indicates a high predictive ability of the model. It should be noted that the model did not make gross errors: the values for only three observations were incorrectly predicted. At the same time, in the case of observation 8, the actual price of the apartment is close to the lower border of the 6th price range and the model predicted falling into this price category. In the case of observation 24, the actual price of the apartment is close to the upper border of the 4th price range or to the lower border of the 5th price range, and the model predicted entry to the 5th price range. The most serious forecast error was made in the case of observation 14, for which the actual price lies in the 5th price range, and the forecast price category turned out to be the 6th.

Thus, the approbation of the developed complex model for predicting the cost of a unit of residential property showed that the average relative forecast error does not exceed 10%. At the same time, errors lie within neighboring price categories and take place only when the real price of residential property is close to the border of the price range. The complex model we developed makes it possible to estimate the price range in which the value of the property is located based on its initial characteristics.

Conclusion

Within the framework of this study, the existing Russian and foreign models and methods for predicting the value of residential properties have been analyzed. On this basis, the spatial structure of the models for estimating the value of residential properties is proposed, and the pricing factors influencing the value of residential properties in Moscow are identified, investigated and analyzed.

A model has been developed for predicting the comfort of residential properties based on a random sample of properties (apartments) in all administrative districts of Moscow based on CRT classification trees in which the dependent variable of the value of residential property was transformed into a categorical variable characterizing the price category. The predicted accuracy of the CRT classification tree method was 72%. A complex model for predicting the cost of residential properties has been developed. It consists of two sub-models — a model for predicting the comfort of residential property and an ordinal logistic regression model for predicting the cost of a unit of residential property, including factors characterizing the building (apartment) and the predicted value of the level of comfort, which is the result of the model for predicting the comfort of residential objects. The ordinal
### Table 4.
Forecast of the cost per unit of residential property in the secondary market of Moscow

<table>
<thead>
<tr>
<th>No</th>
<th>Predicted category</th>
<th>Predicted probability of falling into a category</th>
<th>Price range for a category, RUB</th>
<th>Actual value, RUB</th>
<th>Actual category</th>
<th>Category entry check</th>
<th>Average relative forecast error, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.50</td>
<td>6 450 000 – 7 500 000</td>
<td>6 800 000</td>
<td>2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.94</td>
<td>&gt; 13 490 000</td>
<td>15 200 000</td>
<td>6</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.52</td>
<td>7 500 000 – 8 950 000</td>
<td>7 800 000</td>
<td>4</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0.65</td>
<td>8 950 000 – 13 490 000</td>
<td>12 500 000</td>
<td>5</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.46</td>
<td>7 500 000 – 8 950 000</td>
<td>8 200 000</td>
<td>4</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.58</td>
<td>6 450 000 – 7 500 000</td>
<td>7 000 000</td>
<td>3</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>0.58</td>
<td>6 450 000 – 7 500 000</td>
<td>6 650 000</td>
<td>3</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>0.89</td>
<td>&gt; 13 490 000</td>
<td>13 200 000</td>
<td>5</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0.54</td>
<td>6 450 000 – 7 500 000</td>
<td>6 500 000</td>
<td>2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>1.00</td>
<td>&gt; 13 490 000</td>
<td>22 999 000</td>
<td>6</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.60</td>
<td>5 500 000 – 6 450 000</td>
<td>6 040 000</td>
<td>1</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>0.58</td>
<td>7 500 000 – 8 950 000</td>
<td>8 500 000</td>
<td>4</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>0.55</td>
<td>5 500 000 – 6 450 000</td>
<td>5 700 000</td>
<td>2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>6</td>
<td>0.54</td>
<td>8 950 000 – 13 490 000</td>
<td>9 300 000</td>
<td>5</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>0.59</td>
<td>7 500 000 – 8 950 000</td>
<td>7 950 000</td>
<td>4</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>0.55</td>
<td>5 500 000 – 6 450 000</td>
<td>5 700 000</td>
<td>2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>0.54</td>
<td>5 500 000 – 6 450 000</td>
<td>6 200 000</td>
<td>2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>0.58</td>
<td>8 950 000 – 13 490 000</td>
<td>10 300 000</td>
<td>5</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0.90</td>
<td>5 500 000 – 6 450 000</td>
<td>4 950 000</td>
<td>1</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>0.47</td>
<td>5 500 000 – 6 450 000</td>
<td>5 900 000</td>
<td>2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>0.53</td>
<td>5 500 000 – 6 450 000</td>
<td>6 800 000</td>
<td>2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>4</td>
<td>0.53</td>
<td>7 500 000 – 8 950 000</td>
<td>8 750 000</td>
<td>4</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>0.55</td>
<td>5 500 000 – 6 450 000</td>
<td>5 200 000</td>
<td>2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>5</td>
<td>0.64</td>
<td>7 500 000 – 8 950 000</td>
<td>8 700 000</td>
<td>4</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>0.56</td>
<td>6 450 000 – 7 500 000</td>
<td>7 350 000</td>
<td>3</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>0.50</td>
<td>5 500 000 – 6 450 000</td>
<td>6 100 000</td>
<td>2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>4</td>
<td>0.55</td>
<td>7 500 000 – 8 950 000</td>
<td>7 800 000</td>
<td>4</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>2</td>
<td>0.43</td>
<td>5 500 000 – 6 450 000</td>
<td>5 850 000</td>
<td>2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>2</td>
<td>0.49</td>
<td>5 500 000 – 6 450 000</td>
<td>5 850 000</td>
<td>2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>5</td>
<td>0.62</td>
<td>8 950 000 – 13 490 000</td>
<td>11 300 000</td>
<td>5</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>
logistic regression model showed a predictive accuracy of 71%. The comprehensive model we developed was also tested on real data on the cost of residential properties in the secondary market of Moscow for 2019. The relative error in the forecast of the price category of the value of residential properties was 10%.

References


**About the authors**

**Tatiana K. Bogdanova**
Cand. Sci. (Econ.); Associate Professor, Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia; E-mail: tanbog@hse.ru ORCID: 0000-0002-0018-2946
Albina R. Kamalova  
Doctoral Student, Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;  
E-mail: akamalova@hse.ru

Tatiana K. Kravchenko  
Dr. Sci. (Econ.); Professor, Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;  
E-mail: tkravchenko@hse.ru  
ORCID: 0000-0002-6479-6250

Anastasia I. Poltorak  
Leading Analyst, Seldon 2 LLC, 24 build. 1, Radio Street, Moscow 105005, Russia;  
E-mail: poltoraknasty@gmail.com
The economic order quantity taking into account additional information about the known quantile of the cumulative distribution function of the product’s sales volume

Zhanna N. Zenkova
E-mail: zhanna.zenkova@mail.tsu.ru

Wilson Musoni
E-mail: wmusoni@uok.ac.rw

National Research Tomsk State University,
Address: 36, Lenin Street, Tomsk 634030, Russia

Abstract

In modern logistics and supply chain management, the task of inventory management is paramount. The total costs of the enterprise and consequently, its profit, directly depend on the accuracy of calculating the volumes and terms of orders. In this work, the problem of increasing the accuracy of calculating the economic order quantity for a product was solved by involving additional information about the known quantile of a given level of the distribution function of the volume of product’s demand. The quantile information was used to recalculate the annual demand for the product, based on a modified estimator of the sales expectation for the period. The modified estimator is asymptotically unbiased, normal, and more accurate than the traditional sample mean in the sense of mean squared error. New formulas for calculating the economic order quantity and its confidence interval were presented and tested on real data on the monthly sales volumes of goods of a large retail store network over two years. It is shown that the classic way of mean calculation led to an underestimation of the volume of the economic order quantity, which in turn increased the risk of a shortage, and hence a drop in the quality of logistics services. The new calculation method also showed that the period between orders should be one day shorter. The work is practically significant; according to its results, recommendations are given to the enterprise.
Introduction

In the struggle for the attention of consumers, the quality of the service provided to them and an increase in the level of their satisfaction [1], modern enterprises are increasingly focused on improving the accuracy of forecasting future demand for a product, while reducing the risks of shortages and surpluses, trying to introduce pulling production systems throughout the supply chain and building all production processes based on the final need for the product [2]. Knowing quite accurately the volume of future needs, an enterprise can ensure the supply of raw materials, details and goods with minimal costs and minimal risks of disruption of the production process due to a shortage of goods or overstocking of warehouses, moreover, almost exactly on time [3, 4]. Note that in this case trade is also considered as production. Therefore, in modern logistics, in order to search for the best estimate of potential demand, all kinds of mathematical models of various levels of complexity [5–8] are used, including those that involve additional information to reduce the forecasting error and improve the quality of assessment [9–11] and product classification [6, 12].

This paper proposes a new, more accurate method for calculating the economic order quantity size and its confidence intervals using additional information about the quantile of the sales volume cumulative distribution function (CDF). The source of additional information can be logisticians, marketers, analysts of the company, who have a fairly broad knowledge of the specifics of the enterprise and the specifics of organizing the supply and storage of goods.

Restrictive parameters are also taken into account, for example, the maximum volume of the warehouse and the observed frequency of cases of its overflow. Usually such knowledge is ignored, although it can be used to advantage, since it has long been known that additional information helps to improve the quality of statistical procedures [6–13]. It is also worth noting that the use of information about the quantile has already been considered in a number of works [13–18], using the method of projecting the estimate of the distribution function into an a priori class.

In our case, such a modified estimate of the mean is used, for which it is not required to first construct the modified empirical cumulative distribution function (EDF) and calculate the estimate of the mean value of the indicator by substituting the EDF into the integral of the mathematical expectation, which somewhat simplifies the calculations.

As a result, it should be emphasized that in this study, additional information made it possible to significantly reduce the mean squared error in assessing the annual demand for materials, and, therefore, to increase the accuracy when calculating the economic order quantity.

Key words: economic order quantity; sample mean; additional information; quantile; modified mean estimator; modified confidence interval; assessment accuracy; mean squared error.

to the minimization of the total costs of its placement and storage, and ultimately affects not only the company’s profit, but also its market value [19]. The size EOQ directly affects many logistics processes of the company and the logistics chain as a whole [20], up to the choice of packaging [21] and the mode of transport for delivery [22]. Inaccuracies in EOQ calculations can lead to the so-called bull-whip effect [23, 24], which often leads to colossal losses, as was first identified by Procter & Gamble when selling diapers [25, 26].

The classic formula for EOQ, taking into account the losses associated with the freezing of working capital, is obtained by minimizing the function of the annual total cost of purchasing and placing an order by $X$:

$$TC(X) = \frac{M}{X} \cdot k + \frac{X}{2} \cdot P \cdot (l + z),$$  \hspace{1cm} (1)

where $M$ is the annual demand for a product (stock);

$k$ – is fixed costs for placing one order;

$P$ – is the average annual cost of a unit of the considered product (stock);

$z$ – is the share of the price not received due to the freezing of working capital in the stock of goods during a year (as a minimum value, we can consider the current refinancing rate or the size of the minimum rate on deposit accounts of commercial banks);

$l$ – is the rate of storage costs during a year, a share of the price $P$.

In fact, the first term in the cost function (1), equal to $\frac{M}{X} \cdot k$ is the cost of placing orders in the amount of $X$ during a year, while the number of orders per year $r = \frac{M}{X}$. The second term $\frac{X}{2} \cdot P \cdot (l + z)$, is the average cost of storing the stock during a year. It is easy to prove that the minimum value of the function of annual total costs is achieved if

$$TC'(X) = -\frac{M}{X^2} \cdot k + \frac{1}{2} \cdot P \cdot (l + z) = 0,$$  \hspace{1cm} (2)

so

$$X_o = \sqrt{\frac{2 Mk}{P(l + z)}}.$$  \hspace{1cm} (3)

Note that formula (3) can be applicable only for a sufficiently stable demand for a product [27], i.e., actually for goods from group X according to the classification by the method of XYZ analysis [3, 4, 6]. In other words, the formula is applicable for such goods for which the coefficient of variation (CV) of a number of demand (sales) values does not exceed 10%. In this case, CV is expressed by the formula

$$CV = \frac{\sqrt{S^2}}{\bar{X}} \cdot 100\%.$$  \hspace{1cm} (4)

In the formula (4)

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i,$$  \hspace{1cm} (5)

represents the average level of demand (sales),

$$S^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2,$$  \hspace{1cm} (6)

is the sample variance [28], $X_1, X_2, ..., X_N$ are sales (demand) volumes for a product during the year, $N$ – the number of considered periods during the year.

Suppose that the business experts of the company know that for a certain period the overall level of demand for the product did not exceed a certain threshold level $x_q$ in $q \cdot 100\%$ cases. In fact, assuming that the volume of demand (sales) $X$ is a random variable (RV) with CDF $F(x) = P(X \leq x)$, then such information can be presented in the form

$$F(x_q) = q,$$  \hspace{1cm} (7)

where $x_q$ is the known quantile of the CDF of a known level $q$. 
We use additional information about the CDF quantile to improve the quality of EOQ estimation (3). To do this, assume that \( \{X_1, X_2, \ldots, X_N\} \) is a sample of size \( N \) with independent, equally distributed RVs with CDF \( F(x) \). Then you can find a more accurate estimate of the annual demand for a product (stock) using a modified estimate of the average level of demand, taking into account additional information of the form (7), according to the formula:

\[
M^q = m \cdot \bar{X}^q,
\]

where \( m \) is the number of periods in a year (for example, \( m = 12 \), if the average monthly demand is calculated),

\[
\bar{X}^q = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i}^{N} X_i \cdot \left(1 - \frac{I_{(X_i \leq x_q)} - q}{q(1-q)} \right)
\]

is a modified estimate of the average demand for the period [12]. This estimate is asymptotically unbiased and normal. Its variance is determined by the formula [12]:

\[
Var\{\bar{X}^q\} = \sigma^2 - E^2 \left(1 - \frac{I_{(X_i \leq x_q)} - q}{q(1-q)} \right) + O\left(\frac{1}{N}\right),
\]

where

\[
\sigma^2 = N \cdot Var\{\bar{X}\} = Var\{X\}.
\]

Moreover, it is easy to verify using formula (10) that

\[
\sigma^2_q = \lim_{t \to +} N \cdot Var\{\bar{X}^t\} = \sigma^2 - \left(1 - \frac{q}{q} \cdot \int_{-\infty}^{\chi} x dF(x) - \frac{q}{1-q} \cdot \int_{\chi}^{\chi_q} x dF(x) \right)^2.
\]

Hence, it is obvious that \( \sigma^2_q \leq \sigma^2 \), and thus due to asymptotic unbiasedness for sufficiently large volumes of observations, the use of additional information about the known quantile leads to a decrease in the mean squared error (MSE) normalized to \( N \):

\[
N \cdot MSE\{\bar{X}^q\} = N \cdot E\{\bar{X}^q - a\}^2 \leq \sigma^2 = \frac{N \cdot E\{\bar{X} - a\}^2}{N},
\]

and hence, to improve the accuracy of estimating the average volume of demand. Here \( a = EX \).

Note that the estimate of the mean with allowance for the quantile [13] obtained using the projection method [9] has a similar asymptotic variance (12). The advantage of estimate (9) is that it does not require any preliminary actions, namely, constructing an estimate of the EDF and its projection into the a priori class with the subsequent substitution of the modified estimate of the EDF into the integral for the mathematical expectation.

Figure 1 shows a graph of dependence \( \sigma^2_q \) from \( q \) for \( F(x) = U(0,1) \) (x) is uniform in (0,1) CDF, for which \( \sigma^2 = \gamma^2 \). Figure 2 shows a similar graph for \( F(x) = N(0,1) \) – standard normal CDF with zero mean and variance \( \sigma^2 = 1 \). Figure 3 shows the corresponding graph for the exponential distribution \( F(x) = 1 - e^{-x}, x \geq 0 \), with parameter \( \lambda = 1 \), \( \sigma^2 = \gamma^2 = 1 \). It can be seen from the graphs that taking into account additional information about the quantile made it possible to significantly improve the quality of estimation for all three cases, for sufficiently large volumes of observations \( N \).

Knowing the asymptotic normality allows one to obtain confidence intervals with a confidence level \( \gamma \) for the modified mean (here \( L \) is the lower bound, \( H \) is the upper bound):

\[
L^q = \bar{X}^q - \frac{Z_{\gamma} \cdot \sigma^2_q}{\sqrt{N}},
\]

\[
H^q = \bar{X}^q + \frac{Z_{\gamma} \cdot \sigma^2_q}{\sqrt{N}},
\]

Figure 1 shows a graph of dependence \( \sigma^2_q \) from \( q \) for \( F(x) = U(0,1) \) (x) is uniform in (0,1) CDF, for which \( \sigma^2 = \gamma^2 \). Figure 2 shows a similar graph for \( F(x) = N(0,1) \) – standard normal CDF with zero mean and variance \( \sigma^2 = 1 \). Figure 3 shows the corresponding graph for the exponential distribution \( F(x) = 1 - e^{-x}, x \geq 0 \), with parameter \( \lambda = 1 \), \( \sigma^2 = \gamma^2 = 1 \). It can be seen from the graphs that taking into account additional information about the quantile made it possible to significantly improve the quality of estimation for all three cases, for sufficiently large volumes of observations \( N \).

Knowing the asymptotic normality allows one to obtain confidence intervals with a confidence level \( \gamma \) for the modified mean (here \( L \) is the lower bound, \( H \) is the upper bound):

\[
L^q = \bar{X}^q - \frac{Z_{\gamma} \cdot \sigma^2_q}{\sqrt{N}},
\]

\[
H^q = \bar{X}^q + \frac{Z_{\gamma} \cdot \sigma^2_q}{\sqrt{N}},
\]
where $\sigma_q = \sqrt{\sigma^2_q}$, $\sigma^2_q$ is calculated by the formula (12).

Thus, a more accurate modified formula for calculating the economic order quantity can be found using the formula:

$$X^*_q = \frac{2M^q \cdot k}{P(l + z)}$$  \hspace{1cm} (16)

where $M^q = m \cdot \bar{X}^q$, $\bar{X}^q$ is the average level of demand for a period, $m$ is the number of periods in a year. Confidence interval with confidence level $\gamma$ for modified economic order quantity are as follows:

$$L_{\text{eqor}} = \frac{2mk}{P(l + z)} \left( \bar{X}^q - z_\gamma \cdot \frac{\sigma^2_q}{\sqrt{N}} \right);$$  \hspace{1cm} (17)

$$H_{\text{eqor}} = \frac{2mk}{P(l + z)} \left( \bar{X}^q + z_\gamma \cdot \frac{\sigma^2_q}{\sqrt{N}} \right);$$  \hspace{1cm} (18)

where $z_\gamma$ is the quantile of standard normal distribution with level $\frac{1 + \gamma}{2}$.

Confidence level $\gamma$ can be used as a measure for regulating the level of risk, it is established by the company’s logisticians from the following considerations:

1. Since for highly profitable goods of group A (here AX), the deficit is critical and leads to direct losses [29], then it is recommended to take the upper bound $H_{\text{eqor}}$ as the basis for calculating the parameters of the inventory management system, the risk of a deficit is

$$R_D = \frac{1 - \gamma}{2}$$

For example, if you need to provide a level of logistics service at which $R_D$ is allowed at a level no more than 2% then $\gamma = 0.96$; for $R_D = 5\%$ $\gamma = 0.9$. However, it should be noted that the lower the risk $R_D$, the more $\gamma$, the wider the confidence interval, the greater the value $H_{\text{eqor}}$, therefore, it is necessary to increase
investments in stocks (this is a “payment” for a high level of service). Depending on the firm’s strategy $H_{KQ}$, is also used for calculating the parameters of the replenishment system for goods of group B (BX), but with a higher level $R_B$.

2. Lower bound $L_{KQ}$ is advisable to be used for low-profit products of group C (CX), for which it is extremely important to reduce the risk of surplus $R_S = \frac{1 - \gamma}{2}$.

That is, if the company chooses a strategy of maximum customer satisfaction, then a higher level is chosen $R_S$, for example, $R_S = 10\%$, then you need to calculate the confidence limit $L_{KQ}$ with $\gamma = 0.8$. If the enterprise is aimed at reducing costs, then it is necessary to reduce $R_S$, for example, at $R_S = 1\%$ $\gamma = 0.98$, while an increase in $\gamma$ is equivalent to a decrease in $L_{KQ}$, and, therefore, a reduction in stocks. Depending on the firm’s strategy $L_{KQ}$ is also used for calculating the parameters of the replenishment system for goods of group B (BX), but with a lower level $R_S$.

Note that the equality of the levels of deficit risks $R_B$ and surplus $R_S$ is due to the symmetry of the standard normal distribution: the probability of “hitting” to the right of the upper boundary of the confidence interval is equal to “hitting” to the left of the lower boundary. Recall that the consideration here only of the groups AX, BX and CX, obtained according to the classification by the method of joint ABC–XYZ analysis [4, 6], is due to the fact that the calculation of the economic order quantity $EOQ$ is permissible only for goods with stable demand from group X.

We also note that modern databases allow you to track the demand for a product (stock) with its detailing not only with accuracy of weeks or days, but also to hours. Thus, the number of periods can be considered large enough to use a more accurate formula (16).

2. Calculation of the economic order quantity based on real data

The modified calculation method proposed in the work was tested on real data on sales of a large retail chain. The monthly sales volumes for 2017–2018 were considered (a more detailed consideration here is inappropriate due to the loss of visualization of the application of the technique). Table 1 and Figure 4 show monthly sales values for two years and a graph of their dynamics. In order to preserve trade secrets, the name of the enterprise and the product is not specified, but it should be noted that the product is not perishable; its shelf life exceeds three years.

### Table 1.

<table>
<thead>
<tr>
<th>Year</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>9,081</td>
<td>7,953</td>
</tr>
<tr>
<td>February</td>
<td>7,578</td>
<td>6,267</td>
</tr>
<tr>
<td>March</td>
<td>7,578</td>
<td>7,700</td>
</tr>
<tr>
<td>April</td>
<td>8,044</td>
<td>7,747</td>
</tr>
<tr>
<td>May</td>
<td>8,490</td>
<td>7,780</td>
</tr>
<tr>
<td>June</td>
<td>7,587</td>
<td>7,110</td>
</tr>
<tr>
<td>July</td>
<td>7,031</td>
<td>7,392</td>
</tr>
<tr>
<td>August</td>
<td>7,560</td>
<td>7,328</td>
</tr>
<tr>
<td>September</td>
<td>8,258</td>
<td>7,776</td>
</tr>
<tr>
<td>October</td>
<td>7,609</td>
<td>8,586</td>
</tr>
<tr>
<td>November</td>
<td>7,501</td>
<td>8,302</td>
</tr>
<tr>
<td>December</td>
<td>7,898</td>
<td>8,999</td>
</tr>
</tbody>
</table>
Based on the given data, the average sales level \( \bar{X} \) per month is \( \bar{X} = 7383.75 \) units, and the sample variance \( S^2 = 402546.98 \) units in square, so the coefficient of variation \( CV = 8.10\% \), which allows us to conclude that the demand for this product is stable (the product belongs to group X). It is interesting to note that the data under consideration turned out to be normally distributed, which was confirmed by the Shapiro–Wilk test [30] with the calculated probability \( p\)-value = 0.691. Figure 5 shows the EDF \( F_N(x) \) graph (for convenience, the function is displayed as dots) and the corresponding normal CDF \( N_{(7383.75; 402546.98)}(x) \).

The average annual demand for a product is \( M = 94029 \) units. The company indicated the average price of the product for the period \( P = 110 \) rubles/unit, estimating storage costs at 50\% of the price, i.e. \( l = 0.5 \) and let \( z = 0.06 \). The cost of placing one order is \( k = 5000 \) rubles. As a result

\[
X_o = \sqrt{\frac{2 \cdot 94029 \cdot 5000}{110 \cdot 0.56}} = 3906.97 \text{ units}.
\]

As this product is indivisible, we will find the optimal whole \( X_o \) as one of the nearest integer arguments delivering the minimum of the total cost function:

\[
TC(3906) = 240669.623 > TC(3907) = 240669.616 \text{ rubles/year}.
\]

Eventually, \( X^* = 3907 \) units, therefore, the number of orders per year will be at least
times with the frequency between orders \( \left[ \frac{365}{24.07} \right] \) = 15 days for category AX products and \( \left[ \frac{24.07}{365} \right] \) = 16 days for CX. For category BX, the choice of the number of days between deliveries (15 or 16) depends primarily on the strategic goals of the company.

When carrying out the calculations, the company provided additional information that for quite a long time (more than two years) in 95% of cases the monthly sales of this product did not exceed 9000 units per month, i.e. \( F(x) = q = F(9000) = 0.95 \). This allows a more accurate method of calculating the economic order size (16).

For the correct use of the modified statistics, the statistical hypothesis of the mutual independence of the data was tested by the series criterion \([31]\), which confirmed the independence with the calculated probability \( p \)-value = 0.53161. As a result, it was found that \( \bar{X} = 8.162.93 \) units per month, \( M = 12.8 \times 162.93 = 97,955.16 \) units per year. Then \( X^t = 3,987.71 \) units per month, the minimum cost function is achieved at \( X^t = 3,988 \) units per month, the frequency of orders was \( r^t = 24.56 \), or at least 24 times a year. At the same time, the period between orders changed by 14.86 days. This means that if, according to the joint ABC-XYZ analysis, the product in question belongs to the AX group, then it must be ordered once every 14 days, for BX and CX – once every 15 days. Figure 6 shows graphs of the dependencies of the total cost of restocking \( TC(X) \) and \( TC^*(X) \).

Note that the use of the modified method showed that the volume of the order was underestimated by 2.07% that could lead to a shortage of goods and to a loss of profit. As a result, the estimated total costs of the enterprise increased by 5,492.25 rubles / year, however, an improvement in the accuracy of estimating the annual demand for a product could potentially reduce logistics risks and compensate for losses due to higher profits, the increase of which is due to the higher quality of logistics services. We use the knowledge of the distribution of sales volumes of the initial data in order to obtain confidence intervals with the level of confidence \( \gamma \) for assessing the monthly average demand for a product (here the variance \( \sigma^2 \) is assumed to be known):

\[
L = \bar{X} - \frac{z_{\gamma} \cdot \sigma}{\sqrt{N}}; \quad (19)
\]

\[
H = \bar{X} + \frac{z_{\gamma} \cdot \sigma}{\sqrt{N}}; \quad (20)
\]

where \( z_{\gamma} \) is the \( \frac{1+\gamma}{2} \) - level quantile of standard normal distribution; \( \sigma = 634.47 \) units per month.

To construct confidence intervals for the modified mean, using formula (12) at \( F(x) = N(\bar{X} = 8.162.93, \sigma^2 = 503.16) \), it was determined that \( \sigma_q = \sqrt{\sigma^2} = 503.16 \) units per month. The results are shown in Table 2.
For AX category products, it is recommended to use it as an economic order quantity $H_{EoQ}$. If the company seeks to meet the demand as much as possible, then large values of $\gamma$ should be chosen, for example, 0.98. In this case, the volume of delivery will be 4 046 units, and the risk of a shortage $R_D$ is no more than 1%. If the enterprise is aimed at reducing costs, then it needs to set higher values of $\gamma$, for example, 5%. Then it has to choose $\gamma = 1 - 2 \times 0.05 = 0.9$, as a result, the economic order quantity must be 4 029 units. If our product belongs to the CX group, then it is advisable to minimize the permissible risk of surplus $R_S$, setting it at the level $1 - \frac{\gamma}{2}$. Then, for example, at a 10% risk level $\gamma = 1 - 2 \times 0.1 = 0.8$ and the volume will be added to the order $L_{EoQ} = 3 955$ units. At the same time, it is expected that in two cases out of ten the demand for this product will be only partially satisfied, since it will be completely sold out and there will not be enough for all buyers.

It should also be noted that the confidence intervals with additional information are significantly narrower (by almost 22%) than the intervals without the quantile. This is due to the fact that the modified mean has an asymptotic variance $\sigma^2 \leq \sigma^2$, so at the same level of deficit risks $R_D$ and surplus $R_S$ there can be found more accurate confidence limits — values of the economic order quantity for goods from different groups AX, BX or CX.

### Conclusion

In this paper, we propose a new method for calculating the economic order quantity, taking into account additional information about the known quantile of a given level of the cumulative distribution function of sales (demand) volumes, as well as confidence intervals for it. We show that for a sufficiently large number of observations, the new method gives a more accurate value of the economic order quantity and narrower confidence intervals since it is based on an asymptotically unbiased estimate of the average level of demand for a product with a smaller mean squared error. The new calculation method was tested on real data on monthly sales of a large retail chain for two years. The company was given recommendations on the choice of parameters for the inventory management system of the product in question.

### Table 2

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$z_\gamma$</th>
<th>$L$</th>
<th>$H$</th>
<th>$L^*$</th>
<th>$H^*$</th>
<th>$L_{EoQ}$</th>
<th>$H_{EoQ}$</th>
<th>$L_{EoQ^*}$</th>
<th>$H_{EoQ^*}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.98</td>
<td>2.326</td>
<td>7534</td>
<td>8138</td>
<td>7923</td>
<td>8402</td>
<td>3831</td>
<td>3982</td>
<td>3928</td>
<td>4046</td>
</tr>
<tr>
<td>0.95</td>
<td>1.960</td>
<td>7581</td>
<td>8090</td>
<td>7961</td>
<td>8365</td>
<td>3842</td>
<td>3970</td>
<td>3938</td>
<td>4037</td>
</tr>
<tr>
<td>0.9</td>
<td>1.645</td>
<td>7622</td>
<td>8049</td>
<td>7993</td>
<td>8332</td>
<td>3853</td>
<td>3960</td>
<td>3945</td>
<td>4029</td>
</tr>
<tr>
<td>0.8</td>
<td>1.282</td>
<td>7669</td>
<td>8002</td>
<td>8031</td>
<td>8295</td>
<td>3865</td>
<td>3949</td>
<td>3955</td>
<td>4020</td>
</tr>
</tbody>
</table>
References


**About the authors**

**Zhanna N. Zenkova**
Cand. Sci. (Phys.-Math.), Associate Professor, MBA;
Associate Professor, Department of Systems Analysis and Mathematical Modeling,
Institute of Applied Mathematics and Computer Science,
National Research Tomsk State University;
36, Lenin Street, Tomsk 634030, Russia
E-mail: zhanna.zenkova@mail.tsu.ru
ORCID: 0000-0002-3776-9935

**Wilson Musoni**
Doctoral Student, Department of Systems Analysis and Mathematical Modeling,
Institute of Applied Mathematics and Computer Science,
National Research Tomsk State University;
36, Lenin Street, Tomsk 634030, Russia
E-mail: wmusoni@uok.ac.rw
Developing a knowledge-based system for the design of innovative product creation processes for network enterprises

Yury F. Telnov\textsuperscript{a}\ E-mail: Telnov.YUF@rea.ru

Vasiliy A. Kazakov\textsuperscript{a}\ E-mail: Kazakov.VA@rea.ru

Vasiliy M. Trembach\textsuperscript{b}\ E-mail: trembach@yandex.ru

\textsuperscript{a} Plekhanov Russian University of Economics
Address: 36, Stremyanny Lane, Moscow 117997, Russia

\textsuperscript{b} Moscow Aviation Institute (National Research University)
Address: 4, Volokolamskoe Shosse, Moscow 125993, Russia

Abstract

The relevance of developing knowledge-based systems used to support innovative processes for creating products and services is related to the objective need to reduce the life cycle of products under the influence of modern digital technologies in developing network enterprises. Well-known research results in the field of model-oriented design of products, processes, systems and enterprises do not fully provide semantic interoperability in the interaction of stakeholders in the innovation process. The aim of this work is to build a knowledge-based system architecture that implements semantic interoperability of network enterprise participants at various stages of the product lifecycle. The work is based on the use of a model-oriented approach to building a digital thread at all stages of the product lifecycle, an ontological approach to semantic modeling of a distributed knowledge base and a multi-agent approach to organizing interaction between interested participants in the innovation process. The paper proposes a functional architecture of a knowledge-based system that includes modules for planning the innovation process, forming product value characteristics and functional requirements, construction and value chain design. A multi-level system of ontologies of the innovation process is also developed and its application in the work of functional modules that provide access to
Introduction

Innovative processes for creating new products are characterized by a high degree of uncertainty and dynamic changes in the requirements for the products, especially at the initial stages of the life cycle associated with the definition of the product concept and subsequent design. At the same time, the trend of generally reducing the product life cycle in the conditions of widespread use of digital technologies, in particular, the technology of “digital twins” and virtualization of interaction between participants in the innovation process of a network enterprise, determines the need to develop new methods for modeling the information space at all stages of the life cycle.

The main problem of existing methods of modeling the information space in the form of project repositories [1] is the weak semantic connectivity of the information resources used, reflecting various aspects of the innovative process of creating products. As a result of emerging semantic problems, it becomes very difficult to obtain consistent design solutions formed by heterogeneous network enterprise stakeholders. This makes it relevant to develop a knowledge-based system that would ensure an integrated, consistent and dynamic use of knowledge throughout the innovation process — from the origination of the product idea to its actual implementation in an industrially manufactured product.

The solution of the problem of integrated representation of knowledge about a product and related business processes is currently based on a model-oriented approach to building systems and enterprises [2, 3], the implementation of the concept of digital twins [4], as well as the concept of dynamic maintenance of product models at all stages of the lifecycle — the concept of a digital thread [5, 6]. Semantic integrity and consistency of information sources used (semantic interoperability) is studied in works related to the construction of ontologies of enterprises [7, 8]. At the same time, the complex linking of these concepts within the framework of creating a unified system based on knowledge is far from a theoretical solution and practical implementation, especially in terms of creating mechanisms for the transition between models at different stages of the product lifecycle and interaction of stakeholders in the innovation process.

In connection with the above, the aim of this paper is to define the architecture of a knowledge-based system based on a multi-level system of ontologies of the innovation process which would ensure semantic consistency of product and process models at various stages of the life cycle, ensuring effective
dynamic interaction of real and potential participants of a network enterprise using multi-agent technologies.

1. Methodological frameworks of knowledge-based systems creating the design of innovative processes

The design of innovative processes for creating products of a networked enterprise, as a rule, has a nonlinear iterative nature, involving a long time of searching and testing conceptual ideas, resources for their implementation, coordination with prospective participants of the network enterprise plans to manufacture and market products and services. This dynamic nature of product design processes necessitates the use of various methods for modeling products and services, as well as related processes at all stages of the life cycle.

Theoretically, the process of creating products based on system modeling is reflected in the OMG Model Driven System Design (MDSD) approach [2]. According to this approach, any product can have a complex structure of subordinate components, which, on the one hand, must have clearly defined requirements at each time, and on the other hand, these components must be represented by a set of models in order to develop the most appropriate solutions that meet the formulated requirements. Making design solutions depends on many factors and, above all, on the availability of resources, capabilities and risks of their implementation. The assessment of capabilities and risks can be characterized by incomplete and unreliable information, as well as variability in the state of resources over time under the influence of various external conditions. These circumstances make it necessary to save various design solutions in the project repository and continuously update information about the status of work on the project.

In this regard, we should pay attention to the organization of information storage in intelligent project repositories that integrate a variety of information sources about the project using ontological tools [1]. This creates a unified information space for stakeholders in the process of creating products that belong to different categories. However, the text form that underlies the documentary organization of project information has a number of limitations related to the isolation of documents from each other. In this case, the model representation of project information provides greater connectivity of the analyzed objects and their parameters, as well as the ability to use a variety of tools for quantitative and qualitative assessment of design solutions.

The development of the concept of dynamic maintenance of information about the state of project decisions regarding the product being developed at various stages of the product life cycle is a concept that is characterized by “the use of integrated information stores, which include, in addition to collecting and analyzing primary data, tools for object representation of the domain in the form of collections of digital twins (DTC) with a description of the structure, composition of objects, their properties, as well as known relationships between them. Promising technologies for organizing integrated information storage should also include methods for their systematic description (construction) and logical and analytical processing” [4].

As a rule, a digital twin is a software and hardware complex that displays the state of components of a real product or equipment in time. The digital twin receives information from its physical counterpart, for example, using the Internet of Things (IoT) and performs, if necessary, control actions to bring the physical object to the required state. At the same time, the digital twin can perform analytical and predictive functions, analyzing large amounts of data and functioning,
in fact, as a software agent endowed with the intellectual ability to model situations. In principle, digital twins can be combined into multi-agent systems to simulate more complex technological processes.

Digital twins are mainly used at the product operation stage, accumulating large amounts of data about the product structure or production process. This data can also be used to improve or redesign the product structure. The result of product design can be immediately reflected in the information structure of the digital twin. At the same time, the concept of digital twins is not suitable for implementing the actual product design processes. This is due to the fact that design requires not only information about the functioning of existing products, but also external information about scientific and technical developments, the behavior of competitors, suppliers, manufacturers and consumers of products.

In this regard, the development of the digital thread concept is of great interest. It considers the creation, implementation and operation of products at all stages of the life cycle [5], modeling, in fact, not individual products and processes, but interrelated processes of the enterprise and their interaction with external stakeholders in production activities. As a result, Model-Based Enterprises are formed [3], which best meets the requirements for creating flexible and dynamically formed network enterprises [9].

The Digital Thread for Smart Manufacturing Systems project includes methods and protocols that extend and complement the end-to-end flow of information that passes through the processes of product design, production, and support, enabling the integration of intelligent manufacturing systems. As a result, the life cycle of bringing the product to the production stage is shortened, while saving costs [6]. The project focuses on establishing the relationship between the product model and the production process model, which makes it possible to make adjustments to design solutions through feedback from the implementation of the production process to the product design. At the same time, the basic principles of system engineering are used in terms of requirements management, verification and validation in a clearly formalized form. However, the problems of transmitting unambiguous semantics of requirements and feedback on their implementation have not been fully resolved. In this regard, the development of the Quality Information Framework (QIF) to formalize the requirements for information exchange in terms of quality measurement is of particular interest.

To attract external stakeholders to the product design process and take into account the competitive environment, it is necessary to expand the product model and production processes to cover all the business processes of the enterprise, in fact, implementing the enterprise engineering approach. In this regard, the application of the Model-Based Enterprise (MBE) concept involves the introduction of advances in standards, testing methods and metrology which allow manufacturers to integrate models of systems, services, products, processes and logistics throughout the production enterprise [10]. At the same time, great importance is attached to ensuring that consumer qualities match the capabilities of their process implementation, as well as the best interoperability of decentralized distributed subsystems in order to form various product configuration versions for specific requirements specifications.

The MBE concept assumes the implementation of the following project principles [10]:

---

1 Quality information framework: https://qifstandards.org/
integration of various types of static and dynamic models of intelligent production systems and analytical systems based on the interdisciplinary exchange of system and project information between them;

transformation of enterprises into service-oriented production systems that allow flexible building of business processes on a dynamic basis using modern digital technologies – IoT, AI, Big Data, etc.;

specification of a complex product structure using models, methods and tools that allow one to implement the concept of fully integrated production systems that are adaptive in real time to changing requirements and conditions within the enterprise and in the external environment;

visualization of product lifecycle data necessary for visual representation and perception of product information by participants of innovation and production processes;

identification of production capabilities that allow for evaluating the ability to meet the stated requirements for the product based on modeling the use of the provided resources at all stages of the life cycle;

synthesis and contextualization of information from traditionally incompatible sources to justify decisions in the operational management of product creation processes at different stages of the life cycle;

extracting and applying knowledge to perform operations based on hybrid intelligent systems in order to analyze unstructured information to ensure decision-making and continuous improvement of production systems.

The implementation of these principles is largely determined by the ability of hybrid intelligent knowledge-based systems to perform semantic mapping of the conceptual model of products and related processes in a multi-level ontology of the innovation process. In this regard, the work [11] on integrating models of created products within the framework of implementing a digital thread based on an ontological approach is known. The work is mainly devoted to the technological aspects of presenting various models in a single format that provides their semantic analysis. To do this, the SysML models of a certain product are converted to JSON documents and then converted to an RDF graph using the SysML ontology. In this case, the model dictionary can be automatically generated and compared with existing domain ontologies. If necessary, during sequential processing of RDF triplets, the identified new terms are added to the existing domain ontology. Then, as a result of a number of transformations and mapping to domain ontologies aligned using the BFO reference ontology, a refined RDF graph is formed, which is then used in the process of making innovative decisions. However, aspects related to multi-agent implementation of interaction between participants in the innovation process based on a multi-level system of ontologies and the corresponding cognitive interpretation of the interaction results are not disclosed in the article [11].

The proposed research focuses on the development of a functional architecture of a knowledge-based system that supports iterative dynamic decision-making on the configuration of new products and related processes with the participation of all stakeholders within the network enterprise. The main focus is on the initial stages of product creation, including the formation of requirements and design of the product and value chain, as a system of related production and logistics processes. At the same time, a multi-level system of ontologies of the innovation process is proposed as a system-forming component which provides semantic interoperability of knowledge about products and related processes.
2. Statement of the problem of developing a knowledge-based system for the design of innovative processes for creating products

In accordance with the classical provisions of system engineering the process of designing products and services includes the following stages [2] (Figure 1):

Forming a product concept involves determining the consumer’s value characteristics (business requirements) and evaluating them from the point of view of marketing attractiveness. In the future, the product concept is specified in terms of the possibility of technical implementation in the form of a set of functional and non-functional requirements.

Product construction design is the definition of a product structure based on an analysis of existing advances in the manufacturing of products and services, as well as promising technologies that can be used to create a product. The result of this stage is the component structure of the product and the definition of a list of technologies for the production of its components.

Detailed design of product creation processes involves the selection of network enterprise participants who will supply or produce certain product components in accordance with the specified requirements and technical design. Thus, at this stage, a value chain is formed that determines the distribution of work among the participants of the network enterprise. At this stage, a large amount of work is associated with coordinating the capabilities of suppliers and manufacturers to create certain components of products in accordance with design solutions and requirements.

Due to possible inconsistencies in the resource and other capabilities of the selected enterprises with the design solutions and requirements, iterative reversals are allowed during the product design process to review the project decisions and even changes in the product concept. This circumstance makes it necessary to dynamically maintain different versions of the project in order to choose the most effective solution version, taking into account all possible limitations (simultaneously maintaining a set of digital threads).

---

**Fig. 1. Design stages of the innovative product creation process**
To justify the design solutions in existing works, it is proposed to use a modeling tool, which is selected depending on the nature of the problem being solved. For example, to solve marketing problems of justifying the choice of consumer qualities of products, these can be statistical models of market assessment, and to assess the performance of the designed product — simulation models. To integrate various types of models, it is proposed to use a system of project repositories united by a single object identification system that links all product-related artifacts at all stages of the lifecycle. The work [12] proposes the architecture for such a product lifecycle management system (System Lifecycle Handler, SLH), which is shown in Figure 2.

The main element of this system is a handle system that allows global identification for any artifact in the digital thread associated with the product project, as well as searching for related artifacts using a system of metadata descriptors such as (a) the artifact name, (b) the artifact type and schema, (c) the artifact author, (d) the artifact location, and (e) other attributes. The digital thread dashboard allows you to view, query, and visualize a digital thread or any of its subgraphs at any stage of the lifecycle. The service system provides the following features:

- connect to corporate data stores (repositories) such as PLM systems, alarm systems, requirements management systems, databases, and environment modeling tools;
- search for individual artifacts;
- linking artifacts from the various repositories and the implementation of the conversion from one form to another;
- synchronization of changes in different repositories;
- graphical representation of the state of artifacts using various visualization tools.

Thus, the presented architecture of the SLH system basically implements the functions of an integrated knowledge management system, which allows maintaining consistent information about the process of creating and subsequent operation of the product. In contrast to the described product lifecycle management system, it is proposed to develop a knowledge-based system (KBS), which would allow, along with the specified components, to implement:

![Conceptual architecture of the product lifecycle management system](image-url)
setting up mechanisms for solving the problem of building innovative processes on the specifics of the domain;
integration of knowledge sources about products and related processes based on a single conceptual model of the domain;
assessment of the prospects for the application of design properties and requirements for products, components and technologies used;
design of value chains with the selection of participants in the innovation process with the best opportunities;
agreement of parameters of manufactured products with all interested participants of the innovation process.

In accordance with these requirements, a KBS architecture for designing innovative product creation processes is proposed (Figure 3). A central component is the ontology of the innovation process, which more deeply reflects the semantic conceptual model of the domain on the basis of which all the above requirements are implemented.

The main functional components correspond to the stages of the innovative product creation process. Each of them contains modules for selecting relevant artifacts from the corresponding data bases, ranking them by importance and evaluating them from the point of view of resource implementation capabilities. Project solutions are saved in the repository for different versions of the project from the point of view of the possibility of dynamically building and evaluating various versions for innovative processes and selecting the best version at each time.

Next, we will consider the features of implementing functional modules for designing innovative product creation processes from the point of view of forming design solutions based on the use of ontologies and databases of artifacts. Mechanisms for ranking design solutions based on the application of QFD and FMEA methods were considered earlier in [9], and the assessment of compliance of resource capabilities with value characteris-
tics and functional requirements based on the application of fuzzy logic methods was considered in [13].

3. Structure of the knowledge base for designing innovative processes for product creation

In the architecture of KBS, a special place is occupied by the system of ontologies [14], which describes the structure of the innovative process of creating products in a network enterprise and the requirements for its organization, characteristic of various domains. This system provides integration of network enterprise databases into a single distributed knowledge base.

The following structure of ontologies is proposed, which will ensure the joint use of KBS by all participants in the innovation process of a network enterprise (Figure 4).

In particular, the system of ontologies includes the following elements:

- **metaontology** or the description language used to describe all other ontologies;
- **upper level reference ontology** that contains key concepts, such as Object, Event, Space, Individual quality, and Time frame, and sets the approach to describing ontologies of domains. This ontology is used to align and share various lower level ontologies;
- **mid-level reference ontology** related to the area of this research-organization of activities and production of network enterprises;
- **domain ontologies** shared by participants in the value chain related to the production of a certain type of product that can be used by enterprises in the relevant industries and included in their knowledge base (refer to the reference ontology and contain product classifications, components, standard requirements and processes, etc.);
- **applied ontology of the network enterprise**, expanded from the upper and mid-level levels of ontologies, which is part of the knowledge base of the network enterprise and includes models of products and services produced, production chains and interaction patterns of the organization;

![Fig. 4. Multilevel system of ontologies of KBS](image-url)
applied ontologies of participants in the value chain, extended from the upper and middle levels of ontologies, which are part of the knowledge bases of the relevant organizations and are used to describe the methods and practices used, information about the activities, products and contracts of the organization.

Heterogeneous data submitted by organizations can be integrated and converted into a standardized description format (in the form of an RDF graph) and aligned according to the ontologies used. Alignment is a prerequisite for making queries (in SPARQL or any other standardized language) to the unified knowledge base to control the integrity and implement the functions of the main modules.

The process of ontological alignment involves establishing correspondences between the concepts used in applied ontologies, and from the point of view of computer science is due to the need to integrate heterogeneous databases developed independently and using their own dictionaries. To facilitate this process, reference ontologies (top-level, middle-level, and subject-level) are used.

The language used for the formalization of ontologies determines metaontology. In the context of the Semantic Web project, or to ensure semantic interoperability of complex systems interacting over the Internet, OWL, a language developed by the World Wide Web consortium, is often used as such a language [15]. Languages based on first-order logic are also used for knowledge exchange in computer systems [16, 17].

When using a knowledge base represented as an RDF graph, OWL metaontology can be described based on the basic RDF schema [18] and include built-in classes and properties that together form the basis of the RDF/XML OWL 2 syntax. Decentralization of the network enterprise and the need to share knowledge bases of different organizations implies alignment and integration of applied ontologies. For these purposes, reference ontologies of the upper levels and reference ontologies of subject areas are used [19].

Among the well-known top-level reference ontologies, we can distinguish the BFO ontology, which is used by over 100 projects [20], GFO projects [21], DOLCE [22], and SUMO². Among the reference subject ontologies of interest for the organization of the innovation process of a network enterprise are product ontologies, such as eClassOWL, GoodRelations and derivatives [23], Google taxonomy³, which are used to describe more than 30,000 types of products, nodes and components, as well as taxonomies of processes and activities [24].

For the organization of interaction and selection of participants of a network enterprise in a single information space, it is proposed to use the reference ontology of the innovation process, the key concepts of which are presented in the article [9].

An example of a system of related ontologies for designing innovation processes is shown in Figure 5. The resulting set of related data and ontologies is a distributed knowledge base of KBS. It can be used to ensure the interaction of organizations that are part of the network enterprise structure at various levels – both for describing quality characteristics, and when designing a production system, organizing the value chain, and the innovation process.

---

² Suggested Upper Merged Ontology (SUMO): http://www.ontologyportal.org/
³ Google Product Taxonomy: https://www.google.com/basepages/producttype/taxonomy.en-US.txt
Fig. 5. Example of a related ontologies system for designing innovation processes.
4. Organization of interaction between participants in the value chain of a network enterprise

The use of a distributed knowledge base in the process of design a value chain and the structure of a network enterprise is carried out within the framework of five main modules:

- planning module (coordinator);
- quality attribute forming module;
- requirements forming module;
- construction design module;
- value chain design module.

The planning module receives information about the type of Product being designed, determines the order and calls the corresponding modules, analyzes the results obtained, and initiates clarification of the missing information. To justify the choice of quality characteristics and product requirements, as well as product construction versions and value chains, an approach based on a combination of QFD and fuzzy logic methods is used [9, 25].

The quality attribute forming module is used at the stage of forming a Product concept to identify and formalize Quality attributes (Figure 6) through which the value of the Product for the consumer is determined.

The reference ontology defines the general concepts of Product and Attribute, as well as the types of attributes in accordance with the Kano method [26]. In shared ontologies, Product types that are specific to a particular domain are added. The organization’s knowledge base stores descriptions of different types of typical attributes for different types of products. This knowledge is accumulated in the course of research on consumer preferences and can be used to determine the attributes of a new product, so that:

"Fig. 6. A fragment of the reference ontology: an example of describing the quality attributes of a product"
selects the Product type corresponding to the specified one;

- a set of Product attributes of different types is generated from the knowledge base that is specific to this type of Product;

- all Required attributes will be associated with the new Product;

- the Expected and Attractive attributes are evaluated and included in the set;

- minimizes the presence of Non-Influencing, Contradictory and Distasteful attributes.

The received attributes are returned to the planning module, selected, and passed on to form a set of requirements.

For each mandatory or expected characteristic, at least one related requirement must be formulated. The formalization of requirements is made using the requirements forming module. Within this module, qualitative Attributes are specified in terms of functional and non-functional Requirements for the Product that ensure the implementation of these Attributes [27] (Figure 7).

The top-level ontology includes a description of the concept of a Requirement and its relationships with other concepts (for example, an Attribute, Function, or Product element). Organizations’ knowledge bases can store Standard requirements specific to certain Product categories.

When creating new Product requirements, the Original requirements are generated. Based on the Standard requirements, an Original requirement is usually created that refers to the Standard requirements (for example, “take into account the requirements of GOST R ISO 12003”). Some Implied requirements may be highlighted based on previous experience from the organization’s knowledge base.

![Fig. 7. A fragment of the reference ontology. An example of the description of the product requirements](image-url)
To conduct further analysis, it is necessary to clarify the wording of requirements as much as possible, including by separating complex Original requirements and separating Derived requirements from their composition.

The resulting set of requirements is returned to the planning module. Fuzzy estimation is used to determine the characteristics of requirements. The priority of a characteristic depends on its type: for required, expected and attractive characteristics it is equal to one, for adverse ones – is equal to zero, for non-binding and contradictory characteristics – a fuzzy scale of the certainty coefficient [0; 1] is applied.

As a result of evaluating the resulting set of requirements, the planning module calls the next module or refers to the previous one for clarification.

Within the framework of the construction design module, the component composition is formalized, the structural Elements of the product and its Functions are highlighted.

A requirement may be related to some Element or Function of a product, either explicitly or indirectly, through the choice of production method, technology, sources of information/knowledge used, regulations and tools. In this way, the design is defined through the Requirements (Figure 8).

Constructive Elements can be included in the Product structure to provide certain functions, mandatory interface requirements, or as a result of inheriting from a typical Product construction that has a product breakdown structure (PBS) that references product ontologies available in the knowledge base.

Based on typical solutions, a product variant (version) or Product project is formed, which combines many types of structural Elements and Functions performed by the product.

Each Project is reflected in structural Models (the composition of structural elements, components or parts, structural and full-scale) and functional Models (product scenarios, simulation models), which confirm compliance with the initial Requirements.

The Product project (project version) is selected by the planning module using a combination of QFD and fuzzy logic methods. Evaluating the total importance of Elements included in the Product construction allows you to rank versions. But the determining factor is whether the selected types of structural elements can be purchased and/or whether they can be produced in accordance with existing Models.

To evaluate this possibility, the planning module calls the value chain design module, which is used to analyze the market and select participants in the innovation Process that has several Versions. In order to choose the Version of Product process, it is necessary to organize interaction between different Enterprises based on their competencies, formalized and published in open sources (Figure 9). The process of matching the requirements and capabilities of enterprises is carried out using multi-agent technology [28].

Providers of structural elements or raw materials that are represented in some version of the Product construction are searched in the register of Enterprises. For each Provider, the ability to deliver a certain category of Elements is evaluated. As initial requirements, the Provider receives a description of the Element, formalized through links to the product ontology and existing structural and functional Models that describe the Element and its purpose, as well as design Requirements.

The absence of constructive Element from Providers associated with mandatory Requirements can lead to one of the following organizational scenarios initiated by the planning module:

- completion of work due to the lack of product Elements necessary for creating the Product;
- revision of the Product concept in order to clarify the characteristics and requirements;
- selecting a different Product construction version according to the rank;
Fig. 8. A fragment of the reference ontology: an example of a product project description
Fig. 9. A fragment of the top-level reference ontology: an example of a description of participants in the value chain.

- select the version of producing the Element to order using the necessary technologies and tools.

In order to identify a Manufacturer that can produce custom Elements according to the Product construction, Process versions are formed. These enterprises could perform work on the production of Elements, as well as work on the design, assembly, processing, adjustment, maintenance and delivery of products. The composition and relationships of Activities are contained in the domain ontology. For each of these types of Activities, possible Tools and Technologies are identified that ensure the production of constructive Elements with specified Design constraints and that meet the established Requirements.

The Process versions are evaluated by the planning module in the same way that Product pro-
pects are evaluated using QFD and fuzzy logic methods. Evaluating the total importance of all Activities using the QFD method allows you to rank versions in the process chain and select Manufacturers based on their competence requirements (ability to perform Activities, knowledge of Tools and Technologies). The algorithm for evaluating deviations of Requirements from the capabilities of the Enterprise based on fuzzy logic is considered in the article [13], while determining the possibility of performing Activities on their own or outsourcing it.

The search for Manufacturers is carried out by comparing the description of the type of Activity and the corresponding Tools and Technologies with the facts presented in the databases of enterprises. If there are no Enterprises that fully meet the specified conditions, then the process is switched to other versions according to the previously defined rank.

A process version with a list of Activities that must be performed to form the final Product construction is saved by the planning module in the knowledge base of the network enterprise and used at the production stage for coordinating work, as well as for developing new types of products.

**Conclusion**

As a result of the study of methods for designing innovative processes for creating products of a network enterprise and their implementation in knowledge-based systems, the following conclusions can be drawn:

* within the framework of the innovation process, the most preferable approach is to build enterprises based on models (Model Based Enterprise, MBE), which best implements the concept of “digital thread”;

* to organize the “digital thread” within the framework of the proposed knowledge-based system, it is necessary to use a distributed knowledge base and a multi-level ontology system for designing innovative processes of product creation which displays both the typical structures of products and value chains, as well as the accumulated production experience and capabilities of enterprises;

* to organize the functioning of KBS in the structure of the system there should be allocated modules for the formation of quality characteristics of the product, requirements for the product, the product construction and value chains (the process of its production), as well as the planning module;

* within the framework of designing the structure of the innovation process of a network enterprise, it is advisable to use a combination of various methods (QFD and fuzzy logic) that ensure the best choice of product construction, value chains and the composition of enterprises participating in the value chain.

The novelty of the proposed knowledge-based system architecture for the design of an innovation process lies in the possible iteration of this process, which allows us to find the best design solutions for both the product construction and the value chain. In addition, the implementation of this architecture will improve the semantic interoperability of participants in the innovation process, and, consequently, the quality of innovative products created.

As part of the further development of the proposed methods and approaches to the development of KBS for the organization of the innovative process of product creation, it is advisable to develop algorithms for individual modules of the system, methods and tools for integrating ontologies included in the distributed knowledge base, methods of multi-agent interaction of network enterprise participants that allow one to dynamically adapt the structure of the innovation process of a network enterprise to requirements that change at all stages of the product lifecycle.

**Acknowledgments**

This work was supported by the RFBR, grants no 19-07-01137a, 18-07-01053a, 18-07-00918a.
References


About the authors

Yury F. Telnov
Dr. Sci. (Economics), Professor;
Head of the Department of Applied Information Technologies and Information Security,
Plekhanov Russian University of Economics,
36, Stremyanny Lane, Moscow 117997, Russia;
E-mail: Telnov.YUF@rea.ru
ORCID: 0000-0002-2983-8232

Vasiliy A. Kazakov
Cand. Sci. (Economics);
Leading Researcher, Research Institute “Strategic Information Technologies”,
Plekhanov Russian University of Economics,
36, Stremynnny Lane, Moscow 117997, Russia;
E-mail: Kazakov.VA@rea.ru
ORCID: 0000-0001-8939-2087

Vasiliy M. Trembach
Cand. Sci. (Tech.), Associate Professor;
Associate Professor, Department 304, Moscow Aviation Institute (National Research University),
4, Volokolamskoe Shosse, Moscow 125993, Russia;
E-mail: trembach@yandex.ru
ORCID: 0000-0001-7499-4368
An approach to identifying bots in social networks based on the special association of classifiers

Vladimir N. Kuzmin
E-mail: kuzmindvn@ya.ru

Artem B. Menisov
E-mail: men.arty@yandex.ru

Ivan A. Shastun
E-mail: shastunivan1982@gmail.com

Space Military Academy named after A.F. Mozhaysky
Address: 13, Zhdanovskaya Street, Saint Petersburg 197198, Russia

Abstract
Currently the use of bots, i.e. auto-accounts in social networks which are managed with special programs but disguised as ordinary users, has serious consequences. For example, bots have been used to influence political elections, distort information on the Internet and manipulate prices on the stock exchange. Many research teams concerned with the detection of such accounts have made use of machine learning methods. However, the practical results of detecting social network bots indicate significant limitations because the methodological tools used have language limitation and ineffective criteria for detection. This article presents improved countermeasures in a methodological approach to develop a universal social network account classifier for minimizing the average risk of errors in bot detection. The application of an assembly of classifiers united by a data adaptation criterion and results from the variance of each model found the formation of a universal classifier for social network accounts. The main results obtained by the authors consist of the criteria system and the categorical (nominal) features transformation approach for the formation of the special ensemble of classifiers. In practice, use of the ensemble of classifiers allows us to increase the effectiveness of bot detection compared to other approaches.

Key words: bot detection; social networks; machine learning; ensemble of models; association of classifiers.

Introduction

The detection of bots on social networks has been a subject of research for over a decade [1] due to their active use for political propaganda purposes. However, to date, no unambiguous interpretation of the term “social network bot” has been formed [1]. In this research, a bot will be understood as a special page (account) of a social network disguised as an ordinary user, which automatically and/or according to a schedule performs actions to publish, promote and comment on materials aimed at achieving certain propaganda or political (economic) goals. Depending on the scope of application, several characteristic goals can be distinguished, as presented in Table 1.

Table 1.

<table>
<thead>
<tr>
<th>No</th>
<th>Applications</th>
<th>Informational purposes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>political</td>
<td>promoting ideology; the imposition of political views; agitation; propaganda; to draw the attraction of the electorate</td>
</tr>
<tr>
<td>2.</td>
<td>economic</td>
<td>advertising of goods and services; increasing brand awareness</td>
</tr>
<tr>
<td>3.</td>
<td>social</td>
<td>increasing personality recognition; black, gray and white PR</td>
</tr>
<tr>
<td>4.</td>
<td>moral</td>
<td>propaganda and change of ideological stereotypes</td>
</tr>
</tbody>
</table>

In modern conditions, the use of bots of social networks has become a threat aimed at discrediting the legitimate government and worsening the controllability of social processes and organizations [2]. Currently, according to the number of bots and the consequences of their actions, the social network Twitter stands out: it has more than a million such accounts [3], and individual botnets contain up to half a million accounts [4].

Many organizations are interested in the development and improvement of means of countering the use of bots [5, 6], for identifying bots, evaluating the results of their actions, and neutralizing the consequences. Analysis of the published research results on this topic [7–22] shows that machine learning and neural networks are widely used to detect bots of social networks. There are two main approaches to detect bots of social networks: based on the processing of materials published by users (Table 2) and based on the processing of the quantitative and qualitative characteristics of the accounts themselves (Table 3).

Despite the achievement of good results of the approaches developed [1, 7–16], the following disadvantages can be identified:

- lack of a sufficiently complete set of data to check the quality of detection of social network bots;
- linguistic limitations of the methods used.

In connection with the above disadvantages, as well as to ensure universality, it is advisable to detect bots based on the results of the analysis of quantitative and qualitative characteristics of accounts, which are interpreted by some authors as “meta-features” [3, 6–8, 11].

As the algorithms that control social networking bots become more sophisticated, so too have algorithms for detecting bots. In studies [17–22], bot detection models range from the simplest [17–19], designed to analyze a single piece of meta-features, to models that use ensemble approaches to analyze large data sets, including a combination of meta-features, actions of social network accounts and content [20–22]. Model ensembles can detect new bot behavior that cannot be detected by individual models since the latter can only detect bots that are sufficiently similar to the data used for training [17].
Results of analysis of approaches and effectiveness of bots detection in social networks (based on the processing of materials published by users)

<table>
<thead>
<tr>
<th>No</th>
<th>Authors</th>
<th>Approach</th>
<th>Effectiveness</th>
<th>Language</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A. Bacciu, M. La Morgia, A. Mei, E. Nerio Nemmi, V. Neri, J. Stefa [7]</td>
<td>latent semantic analysis of tweets</td>
<td>accuracy greater than 0.9</td>
<td>English, Spanish</td>
<td>The research was conducted on data provided at the conference PAN 2019</td>
</tr>
<tr>
<td>2</td>
<td>P. Gamallo, S. Almatarneh [8]</td>
<td>Bayesian classifier</td>
<td>accuracy for English – 0.81, for Spanish – 0.88</td>
<td>English, Spanish</td>
<td>PAN 2019</td>
</tr>
<tr>
<td>3</td>
<td>I. Vogel, P. Jiang [9]</td>
<td>principal component analysis, N-gram</td>
<td>accuracy for English – 0.92, for Spanish – 0.91</td>
<td>English, Spanish</td>
<td>PAN 2019</td>
</tr>
<tr>
<td>4</td>
<td>A. Mahmood, P. Srivivasan [10]</td>
<td>TF-IDF</td>
<td>accuracy – 0.91</td>
<td>English</td>
<td></td>
</tr>
</tbody>
</table>

Results of analysis of approaches and effectiveness of bot detection in social networks (based on processing the quantitative and qualitative characteristics of the accounts)

<table>
<thead>
<tr>
<th>No</th>
<th>Authors</th>
<th>Approach</th>
<th>Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>J. Lundberg, J. Nordqvist, M. Laitinen [12]</td>
<td>random forest</td>
<td>accuracy greater than 0.9</td>
</tr>
<tr>
<td>2</td>
<td>S.R. Sahoo, B.B. Gupta [13]</td>
<td>Petri nets</td>
<td>accuracy – 0.9916</td>
</tr>
<tr>
<td>3</td>
<td>J. Novotny [14]</td>
<td>random forest</td>
<td>accuracy – 0.9</td>
</tr>
<tr>
<td>4</td>
<td>A. Davoudi, A. Z Klein, A. Sarker, G. Gonzalez-Hernandez [15]</td>
<td>Botometer</td>
<td>F-score – 0.7</td>
</tr>
<tr>
<td>5</td>
<td>M.Mazza, S. Cresci, M. Avvenuti, W. Quattrociochi, M. Tesconi [16]</td>
<td>neural network (LSTM)</td>
<td>F-score – 0.87</td>
</tr>
</tbody>
</table>

1. Statement of the research problem

The sets of accounts of the social network $A$ and their states $Y = \{0, 1\}$ are given, and there is a statistical function $y^*: A \rightarrow Y$, the values of which $y_i = y^*(a_i)$ are known only on a finite subset of objects $\{a_1, a_2, \ldots, a_n\} \subseteq A$, and for $y_i = 1$ the account is a bot, and for $y_i = 0$ it is an ordinary user. The “object-state” pairs are use cases. The set of pairs $(a_i, y_i)$ of precedents $A^t = (a_i, y_i)_{i=1}^t$ is a training set for restoring the dependence $y^*$. 
The problem of social networks bot detection is to construct a decision function \( z : A \rightarrow Y \), as close as possible to \( y^*(a) \), and not only on the objects of the training set but also on the entire set \( A \). In other words, it is necessary to determine the state of an arbitrary social network account \( a \in A \). In this case, the probability of correct classification and the probability of errors define the average risk of detecting bot error:

\[
[E] = p_0 E_0 + p_1 E_1 + \ldots + p_i E_i,
\]

where \( H \) — the risk of detecting bot error; \( \mathbb{E} [E] \) — the mathematical expectation of detection errors; \( E \) — the detection errors set; \( \langle E_1, \ldots, E_i \rangle \) — the detection errors; \( i \) — the number of states classes of social networks accounts; \( p_0 \) — the probability of a correct decision; \( \langle p_1, \ldots, p_i \rangle \) — the probability of errors.

Thus, the task of detecting social network bots is to form a decision on the state of a social networks account at the observed moment in time. In turn, the requirements for the quality of the decision are determined in the form of requirements to minimize the risk associated with making the wrong decision:

\[
H = \mathbb{E} [E] = p_1 E_1 + p_2 E_2 \rightarrow \min,
\]

where \( H \) — the risk of detecting bot errors; \( \mathbb{E} [E] \) — the mathematical expectation of detection errors; \( p_1 \) — the probability of type 1 errors, i.e. when a user account is mistakenly classified as a bot; \( p_2 \) — the probability of type 2 errors, i.e. when the bot skips.

2. Methods

2.1. Social networks bot detection criteria

It should be noted that it is not recommended to detect social network bots only by one indicator (for example, only by the number of publications or the number of subscribers) [22]. A combination of such features as the thematic relationship of accounts, activity, anonymity, and, in some cases, data inconsistency is important.

**Thematic relationship of accounts.** The presence of links, subscriptions or other actions by multiple accounts on a specific thematic cluster of accounts (or on one account) is a sign of using bots since one of the main tasks of bots is to “amplify the signal” of other users, not only by commenting and quoting them. In social networks, a ranking system is used that increases the degree of distribution of account materials, depending on the number of not only subscribers but also those who passively view the material or simply navigate to the pages.

**Activity.** The most obvious sign of a bot is its activity. This feature can be determined by open data (for example, the number and frequency of posts and subscriptions since the creation of the account).

**Anonymity.** The third important feature is the degree of account anonymity. In general, the less personal information an account has, the more likely it is a bot. Also, a bot feature is the configured privacy for the page.

**The inconsistency of the data** may lie in the inconsistency of the language of the materials and the place of the account foundation, the place of the account foundation and the time zone of the account.

2.2. Formation of an initial data set

The use of social network bots attracted particular public attention in South Africa when Bell Pottinger used social media to spread negative content [23]. Among the subscribers to the accounts of South African politicians there were bots that followed and moderated their tweets on their pages [24].
Analysis of data from two accounts of the South African politicians (Paul Mashatile, chairman of the African National Congress (ANC) in Gauteng province [25], and Ayanda Dlodlo, a member of the ANC [26]), collected in September 2018, showed that out of 12,000 active subscribers 863 subscribers are shared (Figure 1). Of this category of users, 121 accounts were clearly used to increase the rating of propaganda material. They were selected for the following indicators: the account distributed more than 100 messages per day and was also distinguished by high activity, anonymity, and data inconsistency.

2.3. Description of the methodology for using a special assembly of classifiers

The methodological sequence for identifying social network bots includes the following stages: searching and saving data from social network accounts, preprocessing data, selecting and training individual machine learning models, combining them (models), determining the state of social network accounts.

Stage 1. Search and save data from social network accounts. This phase is aimed at collecting all available information about social network accounts using the API (application program interface) methods of the social network Twitter [27]. The data description is presented in Table 4.

Stage 2. Preprocessing data. This stage includes filling in empty (incorrect) rows, as well as transforming categorical (nominal) features [28].

Feature $f$ of the object $a \in A$ (where $A$ is the set of social network accounts) is the result of evaluating some characteristic of the object [28]. Formally, a feature is a function $f: A \rightarrow D_f$, where $D_f$ is the set of valid feature values. If $D_f$ is a finite set, then $f$ is a nominal feature, and $f_1(a), ..., f_n(a)$ is a feature description of the object $a \in A$. For the research, we will assume that $A = D_{f_1} \times \ldots \times D_{f_n}$.
To transform categorical (nominal) features, we use the hot coding technique [28], which allows all categories to be represented as discrete values. Let us assume that the categorical variable \( d_i \in D_{f_{\text{cat}}} \), \( i = 1, \ldots, k \) for a given similarity of categories \( \text{sim}(d_i, d_j) : D_{f_{\text{cat}}} \times D_{f_{\text{cat}}} \rightarrow [0, 1] \) after that the set of values of the feature \( f_{\text{sim}} \in R \) is determined in the form:

\[
    f_{\text{sim}} = [\text{sim}(d_1, d_1), \ldots, \text{sim}(d_k, d_k)],
\]

where \( d_k \in D_{f_{\text{cat}}} \) — the set of all categories.

The categorical feature transformation approach avoids performing one of the laborious stages of training machine learning models — normalizing records in databases [29].

**Stage 3. Selection and training of models.**

The approach developed is focused on the use of an ensemble of classifiers. Ensembles are well known for their effect of improving the accuracy and generalization of a solution, as well as providing parallelism. They have been successfully used in various problems of binary classification [30]. Since the activity of social network bots includes categorical features, adaptation to the initial data is necessary to use ensembles of models (Stage 2). A particularity of the special association of classifiers is that the classifier union is a wrapper for many different models that operate in parallel (Figure 2) to take advantage of the different strengths of each model [31].

Creation of the special association of classifiers consists in performing the following actions:

1. Development of \( N \) separate models, each of which has its own values of detection accuracy. Since each model is trained with multiple partitions of the data sample for sliding control, the value of the number of models \( (N) \) depends on the statistical robustness of the results and the improvement of the solution.

### Accessible Twitter account information

<table>
<thead>
<tr>
<th>No</th>
<th>Name</th>
<th>Definition</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>identification number</td>
<td>string</td>
<td>4452841</td>
</tr>
<tr>
<td>2</td>
<td>Screen name</td>
<td>display name</td>
<td>string</td>
<td>sToneBirD</td>
</tr>
<tr>
<td>3</td>
<td>Date of creation</td>
<td>account creation date</td>
<td>date</td>
<td>2007-04-13 04:33:54</td>
</tr>
<tr>
<td>4</td>
<td>Favorites</td>
<td>pages that are followed</td>
<td>int</td>
<td>43</td>
</tr>
<tr>
<td>5</td>
<td>Followers</td>
<td>followers</td>
<td>int</td>
<td>386</td>
</tr>
<tr>
<td>6</td>
<td>Friends</td>
<td>friends</td>
<td>int</td>
<td>798</td>
</tr>
<tr>
<td>7</td>
<td>GEO</td>
<td>geographic settings</td>
<td>category</td>
<td>True</td>
</tr>
<tr>
<td>8</td>
<td>Lang</td>
<td>language</td>
<td>category</td>
<td>En</td>
</tr>
<tr>
<td>9</td>
<td>List</td>
<td>lists</td>
<td>int</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>Location</td>
<td>location</td>
<td>category</td>
<td>Doha, Qatar</td>
</tr>
<tr>
<td>11</td>
<td>Protected</td>
<td>privacy</td>
<td>category</td>
<td>False</td>
</tr>
<tr>
<td>12</td>
<td>Status count</td>
<td>number of posts</td>
<td>int</td>
<td>3770</td>
</tr>
<tr>
<td>13</td>
<td>Time zone</td>
<td>Time zone</td>
<td>category</td>
<td>Seoul</td>
</tr>
<tr>
<td>14</td>
<td>URL</td>
<td>address</td>
<td>string</td>
<td><a href="http://pbs.twimg.com/">http://pbs.twimg.com/</a>...</td>
</tr>
<tr>
<td>15</td>
<td>Verified</td>
<td>verification</td>
<td>category</td>
<td>False</td>
</tr>
</tbody>
</table>
2. Train each model separately. The learning from the precedents of each model is reduced to the selection of the best value of the model’s hyperparameters (controllers, external) [32]. For example, in a polynomial regression model, an attempt to optimize the degree of the polynomial over the training sample will result in the selection of the maximum possible degree and overfitting.

3. Combining models and improving the values in the final classifier by the following method: the vector of the error probability over the entire set of precedents for each predicted class (for all classifiers) is summed up and averaged. The value of the class is revealed while minimizing the average risk of detection error (expression (2)):

$$y^*(a) \approx z(a), \text{while}$$

$$\underset{\bar{p_1}, \bar{p_2}}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \left[ (H(p_1, p_2))_i \right],$$

where $z(a)$ is the class value;

$N$ is the number of models in the final classifier;

$H(p_1, p_2)$ is the risk associated with errors of the first and second type.

Stage 4. Determining the state of social network accounts. The class value (bot or valid account) is chosen by opposing the models to each other, taking into account the weights. For example, suppose that there are three models in the ensemble, the output of which is the following values:

- $z_1(a) = 0$, while $H = (0,1; 0,1)$,
- $z_2(a) = 0$, while $H = (0,5; 0,5)$,
- $z_3(a) = 1$, while $H = (0,9; 0,9)$.

In the output of classification presented, the model doing the union of the solutions can interpret this result as $y^*(a) = 0$, but assigning the weights to the models {0,1, 0,1, 0,8} will predict $y^*(a) = 1$. Note that the possibility of such a choice is a significant difference between the methods for combining solutions based on multi-tiered generalization from other approaches, for example, from methods in which the final solution is always chosen from the set of solutions proposed by the basic classifiers [33, 34].

The advantage of the proposed approach is that the ensemble of models can be more easily trained on small input datasets and will improve the bot detection performance compared to any single model.

3. Results

In scientific research on machine learning [35, 36], it is customary to present the results of testing the proposed new method in comparison with other methods on a representative set of problems. The comparison should be carried out under equal conditions using the same methodology (especially if it is a sliding control). Table 5 shows the result of comparing the developed approaches to machine learning models suitable for binary classification and input data into the final classifier.
Such an increase in the classification accuracy can be explained by the fact that each model has a weight characterizing the importance of the contribution to the overall solution, which is calculated by the formula (3). The contribution of each classifier can be interpreted as an assessment of his competence used to scale the outputs (work results) of the classifiers, thereby increasing or decreasing the contribution of each classifier to the overall solution.

Figure 3 shows the variance of the classification results (accuracy) of the developed approach and individual models included in the final ensemble obtained with sliding control.

4. Discussion

To assess the quality of the approach we developed, we define the validation part as equal to 0.3 of the total data set (clause 1.2), which includes 47 records for bots and 242 records for real accounts.

To assess the quality of the output data, let us construct the bot detection confusion matrix shown in Figure 4a.

The confusion matrix displays the number of correct and incorrect detections compared to the actual data:

- \( (0,0) \) — correctly identified real social networks accounts;
- \( (1,1) \) — correctly identified bots;
- \( (0,1) \) — for ordinary accounts, it was decided that they are bots;
- \( (1,0) \) — the decision was made for bots that they are real accounts.

These probabilities of the first and second type can be calculated as the probability of the random variable \( z \) falling into the range.
### Table 1: Comparison of Classifier Performance

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our model</td>
<td>1.00</td>
</tr>
<tr>
<td>Log.reg.</td>
<td>0.95</td>
</tr>
<tr>
<td>SVC</td>
<td>0.90</td>
</tr>
<tr>
<td>LDA</td>
<td>0.85</td>
</tr>
<tr>
<td>QDA</td>
<td>0.80</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.75</td>
</tr>
<tr>
<td>NB</td>
<td></td>
</tr>
<tr>
<td>K-Neighbors</td>
<td></td>
</tr>
</tbody>
</table>

### Figures

**Fig. 3.** The contribution of classifiers to the overall solution

**Fig. 4.** The confusion matrix of bot detection among subscribers of the South African politicians accounts:

- a) using the proposed approach; b) using the CatBoost

**Legend for Fig. 4:**
- **Account:** 240, 2
- **Bot:** 1, 46
- True state
- Predicted state

**Legend for True state:**
- Account: 241, 1
- Bot: 11, 36
of acceptable values of the classes of social networks accounts, that is, \( p_1 = P(0,1) \) and \( p_2 = P(1,0) \). Substituting these values from the confusion matrix into formula (1), we get:

\[
H = \mathbb{E}[E] = \frac{2}{241} \cdot 2 + \frac{1}{47} \cdot 1 \approx 0.038,
\]

where \( H \) — the risk of detecting bots error;
\( \mathbb{E}[E] \) — the mathematical expectation of detection errors.

Let us compare the average risk obtained with the results of the CatBoost model [37] developed by the Russian company Yandex (Figure 4b). It is based on gradient boosting with the implementation of the categorical (nominal) feature transformation approach:

\[
H_{\text{CatBoost}} = \mathbb{E}[E_{\text{CatBoost}}] = \frac{1}{253} \cdot 1 + \frac{11}{47} \cdot 11 \approx 2.57,
\]

where \( H_{\text{CatBoost}} \) — the risk of detecting bots error of the CatBoost model;
\( \mathbb{E}[E_{\text{CatBoost}}] \) — the mathematical expectation of detection errors of the CatBoost model.

We also compare this result with the result obtained on the basis of the average risk of a random choice:

\[
H_{\text{random}} = \mathbb{E}[E_{\text{random}}] = \frac{126}{253} \cdot 126 + \frac{24}{47} \cdot 24 \approx 25.195,
\]

where \( H_{\text{random}} \) — the risk of random detecting bots error;
\( \mathbb{E}[E_{\text{random}}] \) — the mathematical expectation of random detection errors.

Thus, the proposed approach showed the best result, which characterizes an increase in the quality of detecting social network bots.

### Conclusion

The development of new approaches to improve the security of government organizations and users of information web-systems is a constant and urgent task.

An element of the scientific novelty of the approach we developed for identifying bots in social networks is the recommended combination of a number of features: thematic relationship of accounts, activity, anonymity and data inconsistency. A particularity of this approach is taking into account the growing trend of using one set of bots to achieve different information goals.

The practical significance of the study lies in the possibility of applying the proposed approach in the substantiation and development of technical solutions for information security.

The approach we developed to identifying bots in Twitter based on a special combination of classifiers has an advantage in terms of efficiency compared to modern machine learning algorithms and reduces errors in detecting bots. Since the activity of bots in social networks includes categorical features, adaptation to the original data is necessary to use the ensemble of models.

However, despite the advantages of machine learning, one of the main disadvantages of the developed approach may be its impracticality if there are too many unique records, for example, if the string representations of categorical features display typos or combinations of several data in the same records.

As directions for the further development of this study, the following can be distinguished:

- research into the collection of additional data on social network accounts;
- analysis of the impact of data imbalance on training models;
- research into the possibilities of improving the performance of detecting social network bots;
- development of technical solutions to improve services for detecting bots of different types.
References


25. Twitter account Paul Mashatile. Available at: https://twitter.com/PaulMashatile (accessed 4 September 2019).


About the authors

Vladimir N. Kuzmin
Dr. Sci. (Mil.), Professor;
Leading Researcher, Military Institute (Science and Researching),
Space Military Academy named after A.F. Mozhaysky,
13, Zhdanovskaya Street, Saint Petersburg 197198, Russia;
E-mail: kuzmindvn@ya.ru
ORCID: 0000-0002-6411-4336

Artem B. Menisov
Cand. Sci. (Tech.);
Researcher, Military Institute (Science and Researching),
Space Military Academy named after A.F. Mozhaysky,
13, Zhdanovskaya Street, Saint Petersburg 197198, Russia;
E-mail: men.arty@yandex.ru
ORCID: 0000-0002-9955-2694

Ivan A. Shastun
Cand. Sci. (Tech.);
Lecturer, Military Institute (Science and Researching),
Space Military Academy named after A.F. Mozhaysky,
13, Zhdanovskaya Street, Saint Petersburg 197198, Russia;
E-mail: shastunivan1982@gmail.com
ORCID: 0000-0002-1086-5345
Fuzzy regression model of the impact of technology on living standards

Yury A. Zelenkov  
E-mail: yzelenkov@hse.ru  

Elizaveta V. Lashkevich  
E-mail: evlashkevich@edu.hse.ru

National Research University Higher School of Economics  
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

Abstract  
This paper proposes a model of the impact of technology on the standard of living based on fuzzy linear regression. The Human Development Index (HDI) was chosen as a dependent variable as an indicator of the health and well-being of the population. The explanatory variables are the Network Readiness Index (NRI), which measures the impact of information and communication technologies (ICT) on society and the development of the nation, and the Global Innovation Index (GII), which measures the driving forces of economic growth. The analysis is based on data for 2019 for four groups of countries with different levels of GDP per capita. For developed countries, the positive and balanced impact of innovation and ICT on living standards has been confirmed. For two groups of developing countries (upper and lower middle income), the GII coefficient was found to be negative. A more in-depth analysis showed that this is due to the state of political and social institutions. This fact means that without a simultaneous increase in the maturity of institutions, stimulation of other areas of innovative development (education, knowledge and technology, infrastructure) leads to a decrease in the quality of life.

Key words: living standards; technical development; fuzzy regression; quantitative models.

DOI: 10.17323/2587-814X.2020.3.67.81
Introduction

Most countries declare that the highest value is the human being, his or her rights, and freedom (see, for example, Article 2 of the Constitution of the Russian Federation). Therefore, many modern politicians and researchers believe that the main goal of the state is the maximum well-being of citizens, indicators of which are the quality of life, equal access to services, the possibility of self-realization, participation in decision-making, etc. [1, 2]. One of the catalysts promoting well-being, along with political and social institutions, is technological progress [3]. Therefore, it is essential to understand how the use of technology relates to living standards and human development.

The goal of this work is to build quantitative models of the impact of technology on the well-being of citizens of different countries. We pay special attention to modern digital technologies, the use of which, according to many practitioners and researchers, will lead to another industrial revolution, which can radically change both state and social structures and individual behavior models [4]. Therefore, we chose the Human Development Index (HDI1) as the indicator under study, and the Network Readiness Index (NRI2) and the Global Innovation Index (GII3) as explanatory variables. The study was conducted on data from 2019.

We also believe that the impact of technology on well-being is different for developed and developing countries [5, 6]. To identify groups of countries, we used the World Bank approach based on GDP per capita [7] and divided the countries under study into four groups.

The results obtained confirm the positive and balanced impact of innovation and information and communication technologies (ICT) on the standard of living for developed countries. For two groups of developing countries (upper and lower middle income), the GII coefficient was found to be negative. A more in-depth analysis showed that this is due to the state of political and social institutions. This fact means that without a simultaneous increase in the maturity of institutions, stimulation of other areas of innovative development (education, knowledge and technology, infrastructure) leads to a fall in the living standards in developing countries.

1. Review of the literature

Most researchers studying the impact of ICT at the country level have focused on increasing productivity as measured by GDP. Many empirical works have proven that investment in ICT leads to GDP growth in developed countries [5], and the effect of ICT is faster than the effect of other technologies [8]. However, no such positive relationship between ICT and GDP has been found in developing countries. The authors of [5] explain this by the lack of appropriate complementary resources and technologies that must complement ICT to ensure sustainable growth.

Later studies [6, 9] showed that the impact of ICT is not limited to productivity growth (i.e. GDP). On the contrary, the use of ICTs also improves the well-being of the country by helping citizens develop their social capital and achieve social equity, and also ensures equal access to health services and education for all groups of the population. Also, the report [10] argues that a networked environment that makes it easier to develop and sell services can create 140 million new jobs and lift 160 million people out of poverty.

However, not all countries with a similar ICT structure (ratio of a fixed to mobile phones,
access to the Internet, etc.) improve their level of well-being in the same way [6]. This can be explained by the influence of factors such as political and social characteristics, the level of trust, the development of the legislation and regulations.

Thus, more research is needed on the impact of ICT on living standards. Therefore, in this paper, we propose a model of the effects of innovation (measured through the global innovation index, GII), including ICT (Networked Readiness Index, NRI), on the development of the nation (Human Development Index, HDI).

The Human Development Index (HDI) is a statistical tool that integrates three key dimensions: health and life expectancy, educational attainment, and living standards. HDI is calculated as the geometric mean of the normalized indices over these three dimensions [11]. HDI has been published by the United Nations Development Program since 1990 [12]. It is a standard tool for comparing the living standards of different countries. Depending on the HDI value, countries are usually classified according to their level of development: very high, high, medium, and low.

HDI can also be considered as an indirect indicator of the maturity of social institutions since they contribute to human development [13]. Along with an important instrumental role in creating and strengthening certain abilities, social institutions shape individual preferences and behavior, since people cannot be considered completely autonomous. Furthermore, the relationship between people and institutions determines whether a society is peaceful, cohesive, and inclusive.

The Networked Readiness Index (NRI) is a holistic framework for assessing the multifaceted impact of information and communication technologies on society and the development of a nation [14]. The NRI identifies three key dimensions for ICT: individuals/society, enterprises, and governments. It contains elements describing the application of ICTs in national economies, including, for example, the quality of regulations.

NRI is published by the World Economic Forum in partnership with the World Bank, INSEAD Business School and Cornell University. In 2019, the index model was revised, and today it includes four main dimensions (technology, people, government, and influence), each of which is a combination of three sub-indices [14].

However, when studying the impact of technology on the well-being of nations, it would be illogical to limit the research only to the consideration of ICT, since significant advances are being observed in other industries – medicine, agriculture, space research, etc. Therefore, we included the Global Innovation Index (GII) in our model, which is based on an extensive database of detailed indicators for the economies of different countries [15]. GII is published by Cornell University, INSEAD and the World Intellectual Property Organization (a specialized agency of the UN).

Recognizing the pivotal role of innovation in driving economic growth and prosperity, the GII includes indicators that go beyond standard innovation measures, such as the level of research and development. The GII is based on two sub-indices, the Innovation Input Sub-Index and the Innovation Output Sub-Index, each of which is based on seven key elements. Five elements describe the characteristics of the national economy that drive innovation: (1) human capital and research, (3) infrastructure, (4) market sophistication, and (5) business sophistication. Two elements reflect the actual results of innovation: (6) knowledge and technology outputs and (7) creative outputs. To calculate these indices, 80 different indicators are used.
2. Model, data and method

Our study is based on the following hypothesis: the national innovation system (as measured by GII) and the development of ICT in the country (as measured by the NRI) positively affect the level of well-being of citizens (as measured by the HDI), but this effect is different for groups of countries that differ in GDP per capita.

Figure 1 presents a research model showing the indices, their dimensions and relationships between the indices.

The required data was downloaded from the sources provided in the Introduction. For the study, we used the latest actual values of the indices published in 2019; the total sample contains data for 112 countries. In addition to the values of the listed indices, the sample also included the GDP per capita obtained from the World Bank website. Table 1 presents the statistical characteristics of the indices, and Figure 2 – the study of their mutual correlation. In the diagonal elements of the matrix in Figure 2, the distributions of the corresponding indices are shown, in the cells above the

<table>
<thead>
<tr>
<th>Index</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Median</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDI</td>
<td>0.43</td>
<td>0.95</td>
<td>0.76</td>
<td>0.13</td>
<td>0.79</td>
<td>−0.65</td>
<td>−0.46</td>
</tr>
<tr>
<td>GII</td>
<td>20.40</td>
<td>67.20</td>
<td>37.56</td>
<td>11.70</td>
<td>34.45</td>
<td>0.68</td>
<td>−0.65</td>
</tr>
<tr>
<td>NRI</td>
<td>22.07</td>
<td>82.65</td>
<td>52.47</td>
<td>16.54</td>
<td>51.17</td>
<td>0.10</td>
<td>−0.91</td>
</tr>
</tbody>
</table>
diagonal – the pair correlation coefficients, in the cells below the diagonal - scatter diagrams.

As follows from Figure 2, all the variables are strongly correlated, $R^2$ is not less than 0.7, and in some cases, even close to one. The scatter plots show that the GII and NRI indices are significantly correlated with HDI, with an $R^2$ value greater than 0.8. There is also a strong pairwise correlation between GII and NRI, potentially leading to multicollinearity.

All 112 countries were divided into four groups on base of value GDP per capita (Table 2) in accordance with the classification of the World Bank [7], which is used to compile the GII index. Figure 3 shows the distribution of

### Table 2.

**Groups of countries according to the value of GDP per capita**

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcategory</th>
<th>Minimum value of GDP per capita (USD)</th>
<th>Maximum value of GDP per capita (USD)</th>
<th>Number of countries in the group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developed</td>
<td>Developed</td>
<td>12 235</td>
<td>–</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Developing</td>
<td>Upper middle income</td>
<td>3 956</td>
<td>12 235</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Lower middle income</td>
<td>1 006</td>
<td>3 956</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Low income</td>
<td>–</td>
<td>1 006</td>
<td>8</td>
</tr>
</tbody>
</table>
HDI for the selected groups of countries: it can be seen that they differ, and the statistical significance of these differences is confirmed by the results of the ANOVA analysis.

The most widely used statistical method to model the dependencies between variables is linear regression analysis. However, in our case, there are problems that prevent its application. As already noted, this is multicollinearity, as well as a small sample size for each group of countries studied. Therefore, we use fuzzy regression.

When the uncertainty of the system is due to a small amount of data, measurement errors, biased estimates, and other human factors, we are dealing with a fuzzy structure of the systems under consideration. In a statistical regression model, deviations between observed values and calculated values are considered random and associated with measurement errors. In a fuzzy regression model, these deviations are assumed depending on the uncertainty of the structure of the system. Thus, fuzzy regression is applicable in cases where the data structure resists statistical analysis [16, 17].

Fuzzy number $A$ is defined by the membership function $\mu_A : R \rightarrow [0;1]$, where $a$ is a modal value, for which $\mu_A : R \rightarrow [0;1]$. Triangular fuzzy numbers are given by a triplet $A = (a_l, a_c, a_r)$, where $a_l \leq a_c \leq a_r$, interval $[a_l; a_r]$ is a reference set ($a_l$ is a lower bound, and $a_r$ is an upper bound), $a_c$ is a modal value [18]. The membership function of triangular fuzzy number is

$$
\mu_A(x) = \begin{cases} 
0, & \text{if } x \notin [a_l, a_r]; \\
\frac{x - a_l}{a_c - a_l}, & \text{if } a_l \leq x \leq a_c; \\
\frac{a_r - x}{a_r - a_c}, & \text{if } a_c \leq x \leq a_r.
\end{cases}
$$

A crisp number can be considered a special case of a fuzzy one, for which $a_l = a_c = a_r$, corresponding membership function is

$$
\mu_A(x) = \begin{cases} 
0, & \text{if } x \neq a_c; \\
1, & \text{if } x = a_c.
\end{cases}
$$

Let us have $m$ observations $y_1, ..., y_m$ of variable $Y$ and corresponding values of $n$ exogenous variables $x_i$, $i = 0, ..., m, j = 0, ..., n$. A fuzzy linear regression model can be presented as

$$
\hat{Y} = A_0 + A_1 x_1 + ... + A_n x_n.
$$
The problem of a fuzzy regression, formulated in [19], consists in the selection of fuzzy coefficients $A_j$, $j = 0, ..., n$, which must satisfy two conditions [20]:

- Fuzzy numbers $\hat{y}_i \in \hat{Y}$ should contain crisp observations $y_i$ with confidence not less than some given number $h$, i.e. the inequality $\mu_i (\hat{y}_i) \geq h$ must be satisfied;
- Total fuzziness of $A_j$ should be minimal.

In our study, we set $h = 0$, i.e. we replace the condition by the requirement that $y_i$ belongs to a reference set of $\hat{y}_i$.

In the conclusion of this section, it is necessary to make a few comments regarding the interpretation of the fuzzy model. As a result of constructing fuzzy linear regression, we get the set $\hat{y}_i = (a_L; a_C; a_R)$ for each observed value of $y_i$, but it is not an analog of the confidence interval obtained from the statistical linear regression model. A statistical confidence interval indicates confidence (e.g., 95%) that the relationship between the variables is as the model describes. The value of the membership function shows the certainty that the fuzzy value $A$ is equal to $e$. A high level of certainty does not mean a high probability of an event; however, if an event is uncertain, then it is impossible. In this sense, the fuzzy regression model is rougher about the situation, but that is why it is more robust when there is little information about system structure.

### 3. Results

To design the models, we used the PLRLS (Possibilistic Linear Regression combined with the Least Squares) method proposed in [21], since this allows us to obtain estimates of the target variable in the form of asymmetric triangular fuzzy numbers. All calculations were performed using the FUZZYREG package [22] for the R language.

The coefficients obtained for different groups of countries are presented in Table 3. Column $d^2[F, \hat{Y}]$ presents the quality score of the model. It is calculated as the average of the distances between the actual observations $y_i$, converted to fuzzy triangular numbers $f_i \in \hat{F}$, and predicted triangular numbers $\hat{y}_i \in \hat{F}$. According to [16], the distance between two fuzzy numbers $A = (a_L; a_C; a_R)$ and $B = (b_L; b_C; b_R)$ is

$$d^2(A, B) = (a_L - b_L)^2 + (a_C - b_C)^2 + (a_R - b_R)^2.$$

The value determined in this way measures the proximity of two membership functions; for $d^2(A, B) = 0$ the membership functions $A$ and $B$ are equivalent.

The $d^2[F, \hat{Y}]$ values show that the model’s confidence slightly decreases with a decrease in income, i.e. the lower the GDP per capita, the lower the accuracy of describing the dependence of human development on ICT and innovation. However, comparing these distances with the actual HDI values, it can be concluded that the reduction of accuracy is not significant.

Also, when analyzing Table 3, attention should be drawn to the fact that in all cases, except countries with low GDP per capita, the coefficients at NII and GII are the crisp numbers, and the fuzziness of the model is provided by the intercept term.

As in the case of the statistical model, the intercept can be interpreted as the value of the target variable in the case when the values of all regressors are equal to zero. In this case, this means that the intercept term sets a certain “basic” value of the fuzzy set, which for a particular observed object is adjusted by the variables describing it.

These basic values of fuzzy sets for different groups of countries are shown in Figure 4. Note that, the range increases as we move from developed countries to lower-middle-income countries but turns into a crisp number for low-income countries.
Interestingly, there is an overlap of intercepts for developed and upper-middle-income (UMI) developing countries. However, from the data presented in Figures 3 and 5, there is no significant overlap between these groups. This fact can be explained using the coefficients of the variables GII and NRI. For developed countries, both coefficients have similar values, which means that the states of this group are more or less balanced in developing all innovative technologies that contribute to improving the quality of life.

In the group of developing countries with upper middle income, ICT has a much more significant impact on human development. However, the coefficient of the innovation index has a negative value. This shocking result means that innovative activity in the country is associated with a fall in the quality of life.

For a deeper understanding of this phenomenon, consider a group of countries that are at the intersection of the sets of developed and developing countries with upper middle income (Table 4). This analysis is of particular interest.

Table 3.

<table>
<thead>
<tr>
<th>GDP per capita category</th>
<th>Intercept term</th>
<th>NRI</th>
<th>GII</th>
<th>(d^2 (F, \hat{Y}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developed</td>
<td>((0.5861; 0.6094; 0.6427))</td>
<td>0.0024</td>
<td>0.0023</td>
<td>0</td>
</tr>
<tr>
<td>Above middle income</td>
<td>((0.5139; 0.5786; 0.6189))</td>
<td>0.0061</td>
<td>-0.0033</td>
<td>0.01</td>
</tr>
<tr>
<td>Below middle income</td>
<td>((0.1938; 0.2749; 0.3445))</td>
<td>0.0016</td>
<td>-0.0029</td>
<td>0.01</td>
</tr>
<tr>
<td>Low</td>
<td>0.2649</td>
<td>((0.006; 0.008; 0.0016))</td>
<td>0.0006</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Fig. 4. Intercept terms of fuzzy regression models for different groups of countries.
interest to the authors also because the Russian Federation also belongs to this group.

As follows from the data presented, Bulgaria has an NRI value comparable to Russia, while being ahead of it in terms of the innovation index. However, in the human development ranking, Bulgaria is below Russia. Argentina, losing to Russia in both exogenous variables, is ahead of Russia in HDI precisely due to the lower GII value. Similarly, Malaysia outperforms all of the countries listed in Table 4 in both the NRI and the GII but has the lowest HDI score. China is significantly ahead of Russia in technology indices but lags behind in the human development index.

Table 5 presents an analysis of the strengths and weaknesses of innovative development by GII sub-indices based on the 2019 report [15] for several countries comparable to the Russian Federation. A “+” sign in a cell means that the compilers of the report marked this direction as a strong side, a “−” sign corresponds to a weak side. The “+/−” signs correspond to sub-indices in the assessment of which both strengths and weaknesses were recorded.

There are several features common to countries belonging to different groups:

- Developing countries are characterized by less mature institutions compared to developed ones;
- In developing countries (and especially in Russia) the situation with the creation of new products and services is worse. At the same time, the development of human capital and the processes of acquiring and using knowledge in developing countries are rated higher on average, but the impact of knowledge on performance is low;
The strong point of developed countries is the ecological sustainability; The strength of all considered developing countries as opposed to developed countries is the size of the market. Figure 6 shows the average values of the GII sub-indices for three developed countries (Croatia, Romania, Uruguay) and four developing countries with an upper middle income (Argentina, Russia, Malaysia, China), listed in Table 5. As you can see, the selected developing countries are ahead of the developed across all dimensions, except for institutions and infrastructure.

Also, for all the considered developing countries with an upper middle income, the same features of financing innovation were recorded. Public investment (as a percentage of GDP) is at a level comparable to developed countries [23], but private investment is noticeably lower than in developed countries. Moreover, developing countries lag significantly behind in terms of foreign direct investment [24]. This testifies to the lack of faith of local private and foreign investors in the possibility of obtaining a reliable income, which is primarily associated with an assessment of the stability of the political, economic and social environment.

Thus, we can conclude from the data presented that the state of institutions is the main factor hindering the improvement of the quality of life for developing countries. Without increasing the maturity of institutions, stimulating other areas of innovative development (education, knowledge and technology, infrastructure) leads to a decrease in HDI. For Russia, among the possible manifestations of this effect are:

- High percentage of emigration in the most qualified part of the population;

<table>
<thead>
<tr>
<th>HDI Rank</th>
<th>Country</th>
<th>GII</th>
<th>NRI</th>
<th>HDI</th>
<th>GDP per capita (USD)</th>
<th>Country group</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>Croatia</td>
<td>37.8</td>
<td>56.75</td>
<td>0.831</td>
<td>14 910</td>
<td>Developed</td>
</tr>
<tr>
<td>42</td>
<td>Argentina</td>
<td>31.9</td>
<td>51.27</td>
<td>0.825</td>
<td>11 684</td>
<td>Developing / UMI</td>
</tr>
<tr>
<td>43</td>
<td>Oman</td>
<td>31.0</td>
<td>52.87</td>
<td>0.821</td>
<td>16 415</td>
<td>Developed</td>
</tr>
<tr>
<td>44</td>
<td>Russia</td>
<td>37.6</td>
<td>54.98</td>
<td>0.816</td>
<td>11 289</td>
<td>Developing / UMI</td>
</tr>
<tr>
<td>45</td>
<td>Bulgaria</td>
<td>40.3</td>
<td>54.77</td>
<td>0.813</td>
<td>9 273</td>
<td>Developing / UMI</td>
</tr>
<tr>
<td>46</td>
<td>Romania</td>
<td>36.8</td>
<td>55.47</td>
<td>0.811</td>
<td>12 301</td>
<td>Developed</td>
</tr>
<tr>
<td>47</td>
<td>Belarus</td>
<td>32.1</td>
<td>50.34</td>
<td>0.808</td>
<td>6 290</td>
<td>Developing / UMI</td>
</tr>
<tr>
<td>48</td>
<td>Uruguay</td>
<td>34.3</td>
<td>56.4</td>
<td>0.804</td>
<td>17 278</td>
<td>Developed</td>
</tr>
<tr>
<td>49</td>
<td>Kuwait</td>
<td>34.6</td>
<td>53.39</td>
<td>0.803</td>
<td>33 994</td>
<td>Developed</td>
</tr>
<tr>
<td>50</td>
<td>Malaysia</td>
<td>42.7</td>
<td>63.76</td>
<td>0.802</td>
<td>11 373</td>
<td>Developing / UMI</td>
</tr>
<tr>
<td>70</td>
<td>China</td>
<td>54.8</td>
<td>57.63</td>
<td>0.752</td>
<td>9770</td>
<td>Developing / UMI</td>
</tr>
</tbody>
</table>

UMI – Upper middle income
### Strengths and weaknesses of countries based on GII sub-indices

<table>
<thead>
<tr>
<th>Sub-index GII</th>
<th>Developed</th>
<th>Developing/ UMI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Croatia</td>
<td>Romania</td>
</tr>
<tr>
<td><strong>Institutions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political environment</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Regulatory environment</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Business environment</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Human capital &amp; Research</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Tertiary education</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>Research &amp; Development</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Infrastructure</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICT</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>General infrastructure</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Ecological sustainability</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td><strong>Market sophistication</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Credit</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Investment</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Trade, competition &amp; market scale</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Business sophistication</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge workers</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Innovation linkage</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Knowledge absorption</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td><strong>Knowledge &amp; Technology Outputs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge creation</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Knowledge impact</td>
<td>+/-</td>
<td>+</td>
</tr>
<tr>
<td>Knowledge diffusion</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td><strong>Creative outputs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intangible assets</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Creative goods &amp; services</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Online creativity</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Low return on investment in intellectual assets. For example, public investments in supercomputers was analyzed in [25]. It was shown that in Russia the return in the form of new knowledge (publications in prestigious international journals) is at a much lower level than in other countries, both developed and developing. Among the listed reasons for this are: the weak connection between science and industry, low demand for the results of scientific activities in industry, the availability of foreign science-intensive products.

A high share of the state in the economy (according to various estimates, up to 70% of GDP). However, an empirical study [26] shows that knowledge is not an efficiency factor for organizations where the state is the owner.

The tendency towards self-isolation, including in scientific research. In particular, the work [27] shows that the priorities of Russian researchers in the field of ICT do not coincide with the priorities of the international community, and the publications are inferior in quality to the results of foreign authors.

In addition, according to forecasts [28], the development of technology can lead to the emergence of a society controlled by the elite, not restrained by traditional values. Already, in some countries it is possible to establish almost continuous surveillance of every citizen. The information collected can be used by elites in their own interests, which do not coincide with the interests of citizens. The most prominent example of such a platform is WeChat, used by the Chinese government [29, 30].

For developing countries with lower middle income (LMI), the model (Table 3) looks similar to countries with upper middle income — the positive impact of NRI and negative GII. However, in this case, the intercept term is sharply shifted to the lower values (Figure 4) and the influence of ICT on the standard of living is much less expressed. It can be stated that for these countries, all the same problems are significant as for countries with upper middle income, but they are exacerbated by the even lower quality of institutions and infrastructure. These countries include, for example, Ukraine, Moldova, and Kyrgyzstan. However, it should be noted that they are ahead of many lower-middle-income countries in terms of the human development index (Table 6). For example, Ukraine is comparable to China in terms of HDI but has noticeably lower values of NRI and GII indices and GDP per cap-
This situation is primarily associated with political instability, i.e., with the quality of institutions. Still, in general, the post-Soviet countries have a higher HDI value than other countries comparable to them in terms of technological indices. This can be explained by a higher assessment of the population, which is still largely inherited from the USSR.

For low-income countries, the structure of the model changes: an intercept is a crisp number, and the fuzzy is the coefficient of the NRI. Note \( \text{Table 3, Figure 4} \) that the resulting intercept term \( 0.2649 \) is close to the modal value for developing countries with lower middle income \( 0.2749 \). However, for this group of countries, the influence of ICT is expressed, firstly, to the least extent, and secondly, it is described by a fuzzy number. Obviously, this is due to extremely low income, which does not allow people to purchase mobile phones and other tools of accessing information, and business and government to develop telecommunications and information services. Note also that for this group of countries, the GII index has a positive effect on HDI, but small values of the innovation index neutralize this effect. This group includes mainly African states with an extremely unstable political situation and an undeveloped economy (Uganda, Ethiopia, Mali, etc.).

### Conclusion

Thus, fuzzy regression analysis of the dependence of the human development index on the network readiness index and the global innovation index revealed some important patterns of the influence of knowledge, technology, and infrastructure on the welfare and quality of life in a country.

For developed countries, the positive impact of innovation and ICT on living standards has been confirmed, but for developing countries, the coefficient at the global innovation index is negative. In fact, this means that stimulating innovation without developing political and social institutions leads to a fall in the quality of life.

---

**Table 6.**

<table>
<thead>
<tr>
<th>HDI rank</th>
<th>Country</th>
<th>GII</th>
<th>NRI</th>
<th>HDI</th>
<th>GDP per capita</th>
<th>Country group</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>Armenia</td>
<td>34.0</td>
<td>49.84</td>
<td>0.755</td>
<td>4212</td>
<td>UMI</td>
</tr>
<tr>
<td>70</td>
<td>China</td>
<td>54.8</td>
<td>57.63</td>
<td>0.752</td>
<td>9770</td>
<td>UMI</td>
</tr>
<tr>
<td>72</td>
<td>Ukraine</td>
<td>37.4</td>
<td>48.92</td>
<td>0.751</td>
<td>3095</td>
<td>LMI</td>
</tr>
<tr>
<td>74</td>
<td>Mongolia</td>
<td>36.3</td>
<td>39.91</td>
<td>0.741</td>
<td>4121</td>
<td>UMI</td>
</tr>
<tr>
<td>75</td>
<td>Tunisia</td>
<td>32.8</td>
<td>42.04</td>
<td>0.735</td>
<td>3447</td>
<td>LMI</td>
</tr>
<tr>
<td>76</td>
<td>Jordan</td>
<td>29.6</td>
<td>46.97</td>
<td>0.735</td>
<td>4241</td>
<td>UMI</td>
</tr>
<tr>
<td>79</td>
<td>Paraguay</td>
<td>27.1</td>
<td>40.55</td>
<td>0.702</td>
<td>5821</td>
<td>UMI</td>
</tr>
<tr>
<td>80</td>
<td>Moldova</td>
<td>35.5</td>
<td>48.93</td>
<td>0.700</td>
<td>3227</td>
<td>LMI</td>
</tr>
</tbody>
</table>
A more in-depth study of this phenomenon requires the joint efforts of various sciences, but already at the moment it is possible to formulate three closely related mechanisms that explain the negative effect of innovation on the standard of living:

1. With limited resources, their redistribution in favor of stimulating innovation leads to a reduction in support for health care, education, etc.;

2. Innovation systems, created in the interests of the elites, are aimed at controlling society, and not at improving the living standards of citizens;

3. Governments, even trying to develop innovative technologies in the interests of the whole of society, in the absence of transparency and competition, use the available resources ineffectively.

References


**About the authors**

**Yury A. Zelenkov**
Dr. Sci. (Tech.);
Professor, Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;
E-mail: yzelenkov@hse.ru
ORCID: 0000-0002-2248-1023

**Elizaveta V. Lashkevich**
Student, BSc Program “Business Informatics”, Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;
E-mail: evlashkevich@edu.hse.ru
An approach to assessing the functioning of hierarchical socio-economic systems and decision-making based on the EFRA software package

Roman A. Zhukov
E-mail: pluszh@mail.ru

Financial University under the Government of the Russian Federation, Tula Branch
Address: 1a, Oruzheynaya Street, Tula 300012, Russia

Abstract
Modern models and methods for evaluating complex systems are associated with hierarchical socio-economic systems (HSES) implemented on the basis of software systems (expert systems and decision support systems) and used at the regional and municipal levels of government. As usual, such systems have the functionality of analytics and building scenario variants for the development of research objects. However, they do not give quantified values of the state and impact factors at which the complex system under consideration can come to a given state. At the same time, the question of determining such a set state associated with the construction of standards (expected values) for elements, classes or levels of the HSES is still open. In some cases, to make an informed decision it is sufficient to obtain aggregated quantitative estimates and recommendations concerning the further functioning of the research object. This article presents the author’s approach, which allows us to evaluate the functioning of hierarchical socio-economic systems and provides expert opinions for making management decisions implemented on the basis of the EFRA software package. The algorithm includes stages of analysis and synthesis—stages of the basic method of system analysis. The novelty of the proposed approach is the possibility of taking into account the specific conditions of the status and impact of complex systems that provides an opportunity to build their own standard. Additionally, the procedures of standardization and normalization (reduction to a scale from 0 to 1) make it possible to avoid the influence of different units of measurement of results of operation and economies of scale. On the example of regions of the Central Federal district according to data for 2014–2017, estimates of the use of information and telecommunications technologies by the population were obtained, and the optimization problem was solved for the Tula Region, on the basis of which directions related to increasing the region’s readiness for digitalization were proposed.
**Key words:** hierarchical socio-economic system; modeling; evaluation; decision-making; software package.

**Citation:** Zhukov R.A. (2020) An approach to assessing the functioning of hierarchical socio-economic systems and decision-making based on the EFRA software package. *Business Informatics*, vol. 14, no 3, pp. 82–95. DOI: 10.17323/2587-814X.2020.3.82.95

---

**Introduction**

Currently, management activities are increasingly using software tools that ensure informed decision-making, including at the regional and municipal levels. At the same time, most systems (expert systems and decision support systems, DSS) based on built-in models, in addition to Analytics, allow you to form scenario variants of the development of the research object, but do not offer quantitative recommendations aimed at optimizing the functioning of the research object, including the hierarchical socio-economic system (SES), in accordance with the goal and tasks to be solved.

Currently, there are more than a hundred software platforms for modeling socio-economic processes, as well as separate software complexes developed for a specific purpose, including for regional-level HSES.

Thus, from 1975 to 1990, under the leadership of V.I. Gurman, the “Region” set of models was developed [1]. They are based on mathematical models that describe the region at three levels of hierarchy and consider it as an open system, in the form of a set of social, natural and economic subsystems. The dseemodel 1.0 (Dynamic Socio–Ecological–Economic model) software package is well known as a regional modeling tool [2].

Over the past twenty years, CEMI RAS has been conducting research related to modeling complex systems of macro- [3, 4], meso- [5, 6] and micro-levels [7] based on computer modeling systems, including econometric models and neural networks, computable General equilibrium models (CGE models), dynamic stochastic General equilibrium models (DSGE models) and agent-oriented models. Among the CGE models, we note the following: the model of the Russian economy “RUSEC,” “Russia: Center-Federal districts,” “RUSEC—natural monopolies,” “RUSEC – Gazprom,” and “Social Russia,” a model of the socio-economic system of Russia with built-in neural networks. Agent models are represented by such products as the Russian agent model, the regional Governor model, the Eurasia model, and the Roscosmos model, among others.1

Software complexes based on agent-oriented models of “smart cities” are being developed on the basis of the situation centers of the Russian regions [8].

At the micro-level (enterprise level), information systems (IS) are successfully used to solve optimization problems: in production companies, logistics and design organizations (including DSS on SAP, Oracle and IBM platforms), as well as intelligent information systems (AIS), the classification of which is presented in [9]. A number of IS systems are used at the Federal and regional levels. For example, the Ministry of Economic Development of the Russian Federation and its subordinate organizations have about 25 external and 11 internal systems that support management decisions2. Most of them are related to document management, providing legal reference informa-

---

1 Laboratory of agent-based modeling (http://abm.center/publications/)
tion and providing public services. A number
are used for analytical research and forecasts,
for example: “Monitoring, analysis and fore-
casting of socio-economic development and
the financial condition of the regions of the
Russian Federation,” “PROGNOSis,” “Pass-
port of the region,” “Catharsis” [10]. The first
of the presented ICS allows you to make fore-
casts and conduct simulation with the possibil-
ity of selecting the functional form of models
by the user. Unfortunately, the platform does
not provide the ability to generate generalized
indicators, and also does not allow you to bal-
ance the values of individual indicators based
on these indicators (bring them to a given
value) by searching for optimal significant fac-
tors.

At the level of a region (subject of the Russian
Federation) or a municipality, it is often nec-
essary to make enlarged (to a certain extent)
assessments of the functioning of the man-
agement object and recommendations for its
development and decision-making. Detailed
solutions are provided during the development
of appropriate measures. Thus, approaches
and tools that do not require users to have spe-
cial skills in the field of modeling socio-econo-
ic processes and in-depth knowledge in
the field of econometric modeling and statis-
tical processing of information, and that could
provide reasonable quantitative recommenda-
tions to government bodies at various lev-
els (region, municipality, company), aimed
at optimal, in terms of resources used and
costs for implementing the relevant activities,
become particularly relevant with develop-
ment of the management object.

The aim of this article is to present and imple-
ment an approach to assessing the functioning
of the HSES using a modified software package
[11, 12], using the example of studying the use
of information and communication technolo-
gies (ICT) by the population of the regions of
the Central Federal District, including the Tula
Region.

1. Formalized description
of the approach

The proposed approach is based on the clas-
cical method of system analysis, which includes
the stages of analysis and synthesis of manage-
ment decisions [13–15].

We will consider HSES as open territorial
socio-economic systems of the regional level
(meso-level) with a mixed type of economic
relations, in the context of the institutional
type of social relations, with an emphasis on
their subsystems of the environmental type,
interacting with subsystems of object, pro-
cess and project types within the space-time
classification based on the system paradigm
[16].

By functioning, we mean the activity associ-
ated with the performance of certain works by
the object of research in relatively unchanged
conditions, with some restrictions, aimed at
achieving its goals, including ensuring the life
of both its own and other related objects of
interest to it. By the assessment of the func-
tioning of the HSES, we mean the quantita-
tively expressed results of its life activity (its
individual components), – value (axiological)
and practical (epistemological and axiological)
significance from the point of view of the act-
ning and cognizing object, – allowing us to ana-
lyze the correlation of such results with nor-
mative (expected, planned) values, taking into
account specific conditions (factors of state
and impact).

This study examines one of the levels $L_p$ of the
hierarchical socio-economic system, which
is a partially ordered set of $\langle H, R \rangle$ elements
$k_{p,(p-1),v_p,s_q}$ between the elements of which a
relation of non-strict order $R$ is defined [17–
19]. Here $p$ is the hierarchy level (subset $L_p$);
$p-1$ is number of the level element $p-1$,
that the $(p-1)$-level element is subordi-
rate to $p$-level element; $v_p$ is element number;
$s_q$ is element class number (a class is a sub-
set of elements grouped by one of the pos-
sible bases). Examples of classifications can be a set of elements that carry out activities in accordance with the OKVED (all-Russian classifier of economic activities), sector classification [20], space-time classification [21], the division of the HSES into social, economic and environmental components (subsystems — classes) [22]. Each of the elements is characterized by four types of features (it is allowed to study elements through their feature descriptions). These include:

- performance indicator \( y_{P(t),r,v_p} \) is the actual value of the result of the element’s operation. For example, for a region, this is the volume of gross domestic product (GDP) by regions for the corresponding type of OKVED activity;

- state factors \( x_{P(t),r,v_p} \) (for example, average annual number of persons employed by types of economic activity);

- impact factors \( z_{P(t),r,v_p} \) (for example, investments in fixed capital by kinds of economic activity);

- normative (expected) performance indicator \( \hat{y}_{P(t),r,v_p} \) is the norm.

Here \( t \) is the time period, \( t = 1, ..., T \); \( T \) is the number of periods; \( j = 1, ..., J \); \( J \) is the number of state factors; \( u = 1, ..., U \); \( U \) is the number of impact factors. In this case, there is a function such that:

\[
\begin{align*}
 f_{P(t),r,v_p} \cdot x_{P(t),r,v_p} (t), \\
 z_{P(t),r,v_p} (t) \rightarrow \hat{y}_{P(t),r,v_p} (t).
\end{align*}
\]

Then the values of the partial (for a single \( v_p \)-th element) and integral (for a set of elements with an index \( v_p \), which is belonging to the same class) performance indicators will be determined by the formulas [23]:

\[
\xi_{P(t),r,v_p} (t) = \frac{y_{P(t),r,v_p} (t)}{\hat{y}_{P(t),r,v_p} (t)}, \quad (2)
\]

\[
\hat{y}_{P(t),r,v_p} (t) = \frac{\sqrt{\sum_{k=1}^{I} \sum_{l=1}^{J} r_{P(t),r,v_p} \cdot y_{P(t),r,v_p}^0 (t) \cdot \hat{y}_{P(t),r,v_p}^0 (t)}}{\sqrt{\sum_{k=1}^{I} \sum_{l=1}^{J} \hat{y}_{P(t),r,v_p}^0 (t) \cdot \hat{y}_{P(t),r,v_p}^0 (t)}}, \quad (3)
\]

where \( r_{P(t),r,v_p} \) are corresponding paired correlation coefficients between \( i_1 \)-th \( y_{P(t),r,v_p}^0 \) and \( i_2 \)-th \( \hat{y}_{P(t),r,v_p}^0 \) variables (performance indicators are actual and expected (normative), respectively, the values of the latter are determined using the production function (PF)) \((i_1, i_2 = 1, ..., I)\) is the number of performance indicators); the index “0” shows that the normalization procedure was performed (brining it to a scale from 0 to 1) after switching from absolute values of features to their standardized forms. The functional form and parameters of the PF can be obtained using factor analysis of dependencies [24]. The expression in the denominator (3) will be called the aggregated PF (APF).

If the value of the indicator for the \( v_p \)-th element (subsystem) is greater than one, then the functioning of the object under consideration is satisfactory.

The proposed indicator has a number of useful properties that distinguish it from other generalized evaluation indicators that differ in the way of aggregation (convolution) of particular indicators, such as the average of various types [25], summation using weight coefficients [26], or the formation of latent variables [27]. These properties include:

- dimensionality, which allows you to compare features that are of different nature and relate to different processes (for example, economic and social);

- normalization, which provides a reduction in the impact of economies of scale and visibility of results (brining to a scale from 0 to 1);

- normability, which makes it possible to compare the estimates obtained with the standard (expected value).
The presented integral (generalized) performance indicator takes into account the relationship of particular indicators that characterize the functioning of the subsystem and its elements included in the class, as well as the specific conditions under which the research object operates.

The following indicator is used to assess the balanced functioning of the HSES (the coefficient of harmony):

$$H_{dp} = 1 - \frac{\sigma(\xi_{p,i})}{M(\xi_{p,i})}, \quad H_{dp,t} = 1 - \frac{\sigma(\xi_{p,t,i})}{M(\xi_{p,t,i})},$$

(4)

where $\sigma(\xi_{p,i})$, $\sigma(\xi_{p,t,i})$ are means;

$\sigma(\xi_{p,i})$, $\sigma(\xi_{p,t,i})$ are standard deviations,

$i = 1, \ldots, I, I$ is the number of partial performance indicators that characterize the class $s$ (in the case of constructing the level’s harmony coefficient $i = 1, \ldots, Q, Q$ is the number of integral performance indicators corresponding to the number of classes);

$\xi_{p,i}$, $\xi_{p,t,i}$ are variables whose values are determined according to (2) and (3).

Being a derivative of partial and integral performance indicators, the harmony coefficient retains a number of their properties; its value does not exceed one and characterizes the degree of compliance of all indicators with the normative (expected) values under specific conditions of the object of research, as well as their compliance with each other. This distinguishes the coefficient of harmony from other indicators based, for example, on the assessment of the share (contribution) of each of the subsystems to the overall result of the system [28], the intensity of interaction [16] or the coupling coordinated [29].

A set of ten indicators is used to assess the effectiveness of the HSES, including indicators of the functioning effectiveness (four indicators), impact effectiveness (four indicators), and management effectiveness (two indicators). The General formula is defined as follows [23]:

$$Ef(t) = \frac{\xi(t)/\xi(t_0)}{x(t)/x(t_0)}.$$  

(5)

where $\xi(t)$, $\xi(t_0)$, $x(t)$, $x(t_0)$ are partial (or integral) indicators of the performance of the element (subsystem) and its factor attribute (state factors or impact factors, or a generalized factor determined similarly to the numerator (3)) of the current and basic or previous period correspondently. Depending on the type of indicators and factors that are respectively in the numerator and denominator of the expression (5), ten types of performance indicators are formed.

Values of indicators (5) that are greater than 1 can be interpreted as follows: the change in the result is greater than the “cost” of changing it.

The performance indicators presented, in addition to the traditional ones that characterize the ratio of results and costs for their implementation (for example, profitability, etc.) or technical efficiency indicators [30], are not “distorted” by the measurement units used and retain the property of normability.

After the analysis stage of the HSES, which identifies the results of its functioning (its subsystems or elements) that do not meet the expected (normative) values, the decision synthesis stage is performed. For elements (subsystems) for which the partial (integral) performance indicators are less than one, it is necessary to implement the optimization procedure, that is, to find such factors of state and impact (or their changes), in which the expected (normative) result of the functioning of the considered elements (subsystems) would correspond to the actual results. The optimization procedure should be performed using standardized models [31]. One of the particular problems can be formulated as follows.

Let the normative (expected) result $\hat{y}_{p,(p-1),\gamma_{p},i}^0 (t)$ for a population element $k_{p,(p-1),\gamma_{p},i}$ be presented in a general case as:
\[
\hat{y}^0_{p,(p-1),\gamma_1,\gamma_2} (t) = f(x^*_{p,(p-1),\gamma_1,\gamma_2},(t)) + \\
\Delta x^*_{p,(p-1),\gamma_1,\gamma_2} (t), \\
\Delta z^*_{p,(p-1),\gamma_1,\gamma_2} (t) + \Delta z^*_{p,(p-1),\gamma_1,\gamma_2} (t)),
\]

where \( j, s \) are indexes of state factors 
\[ x^*_{p,(p-1),\gamma_1,\gamma_2} (t) \] and indexes of impact factors 
\[ z^*_{p,(p-1),\gamma_1,\gamma_2} (t) \] for them;

\( t \) is time (period);

The problem is posed: at what possible values of normalized PF with known functional form, parameters, factors for known and unknown factors 
\[ \{z^*_{p,(p-1),\gamma_1,\gamma_2} (t)\} \] and 
\[ \{z^*_{p,(p-1),\gamma_1,\gamma_2} (t)\} \] the value of the performance indicator will be equal to one, which will indicate that under these conditions, the result of the element’s functioning can be considered satisfactory.

If the optimality criteria for the functioning of HSES or its subsystems will be the equality unit of the integral index, the coefficient of harmony, performance indicators or their combinations, the search for relevant factors of the status and impact of HSES can be reduced to a multi-objective optimization problem, whose solution can be accomplished by several methods [31–34]. The values obtained can be recommendations for various levels of government when making informed management decisions and developing appropriate measures.
2. Conceptual scheme and algorithm for implementing the approach

The conceptual scheme for implementing the approach consists of four generalized blocks (Figure 1) that implement the corresponding algorithm.

The first block provides for loading data. An example is statistical data for the regions of the Central Federal District (CFD), grouped by year.

The second block provides for the implementation of the following functions.

1. Correlation analysis. For the significance level set by the user, the values of the Pearson correlation coefficients, \( t \)-statistics, and its critical value between feature descriptions of elements for the specified periods are output. Significant coefficients are highlighted in color.

2. Factor analysis of dependencies.
   2.1. Selection of effective and factor features (state factors and impact factors) for building models. There is a choice of several effective features.
   2.2. Choosing the functional form of models. Provides a choice of linear, logarithmic, exponential, and power multiplicative forms.
   2.3. The choice of evaluation period. Provides a choice of the start and end periods. If only the number of periods is specified, the estimate is made starting from the last specified period.
   2.4. The construction of models. This function is implemented using two methods: the ordinary least squares method (OLS) (with opportunity of backward selection). This generates models for absolute and standardized variables (standardized models).
   2.5. Evaluating the quality of models and their parameters. It includes an assessment of the quality of models using the Fisher criterion (the coefficient of determination is evaluated) and model parameters using the Student criterion, with the corresponding calculation of calculated and critical values and standard errors, as well as the subsequent formation of an expert opinion based on the results of factor analysis of dependencies. The quality of models can be evaluated for both linearized and non-linearized models.
   2.6. Additional testing for multicollinearity of factors (Farrar–Glauber test) and for heteroscedasticity of a number of residues (Spearman rank correlation test).
   2.7. Calculation of additional characteristics: the average relative error of actual and normative values, as well as elasticity coefficients for the corresponding factors.
   2.8. Point and interval estimation of expected (normative) values of performance indicators. The following formula is used:
\[
\hat{y}_{p,(p-1)\nu_p, r_q} = \hat{y}_{p,(p-1)\nu_p, r_q} \pm t_{\alpha,n,\nu_p} \cdot s_j \cdot \left(1 + \frac{[XZ]^T \cdot ([XZ]^T)^{-1} \cdot [XZ]_p}{m} \right)^{1/2},
\]

(10)

where \( s_j \) is standard error;
\( \hat{y}_{p,(p-1)\nu_p, r_q} \) are calculated values using a formula similar to (6), but for absolute values of factor and result attributes;
\( t_{\alpha,n,\nu_p} \) is confidence level (determined from the Student distribution table), \( \alpha \) is significant level, \( n \) is the number of observations, \( J \) is the number of model parameters;
\( [XZ] \) is matrix of state and impact factors;
\( [XZ]_p \) is vector of expected values (this vector can be entered in the appropriate form, or downloaded from a file).

2.9. Passing the built models to the block for generating evaluation indicators. The transfer can be performed in two ways: complete (in the case of simultaneous evaluation of several selected performance indicators) or backward selection (in the case of building and evaluating models based on a single performance indicator). If the resulting indicator has a negative character, then the procedure for inverting the indicator is provided (replacing its values with reverse values with a “minus” sign).

3. The block for generating evaluation indicators is launched after entering the initial data and working out the second block. It provides for the implementation of the following operations.

3.1. Setting the evaluation periods “from and to” and entering the number of elements belonging to only one element of a higher level. In addition, operation 3.1. is used to check whether input data is correct. If the entered number and the evaluation period do not match, a message is displayed indicating the reason why subsequent procedures cannot be started.

3.2. Selection of effective and factor features for the formation of private indicators for evaluating the functioning of the selected elements and their totality, forming a class or a set of classes.

3.3. Calculation of partial (formula (2)) and integral (formula (3)) indicators of performance, harmony (formula (4)) and efficiency of various types (formula (5)).

3.4. Setting the required output results for evaluating the functioning of elements and classes (classes) of the same level of the HSES.

4. The unit of decision-making. Starts after the third block is completed. It includes the following procedures.


4.2. Setting the required optimization parameters, including an additional optimization option for an integral (generalized) indicator, as well as optimization for each of the factor features.

4.3. Optimization. At this stage, an optimization algorithm is implemented, which consists in searching for state and impact factors separately and by a generalized factor for particular integral performance indicators. In the present version, the restriction (9) is not used.

Each of the blocks provides output of results in the form of tables and graphs, as well as expert opinions.

Thus, the conceptual scheme and the algorithm presented allow us to implement the proposed approach to evaluating the functioning of levels, classes and individual elements of the HSES.

3. Results of the assessment of the use of information and communications technologies by the population of the Central Federal District and the Tula Region

The approach presented here was tested on the example of the regions of the Central Federal District (excluding Moscow).

The choice of evaluation indicators was based on available information obtained from open sources and traditionally used in the analysis of the level of information and communi-
cations technologies (ICT) development at the regional level, as well as significant factors that could influence such results.

Due to limited information and the lack of information for a number of indicators up to 2014 (although the methodology allows using data with different lengths (number of points)), the assessment period 2014–2017 was selected for the Central Federal District regions. The information base was compiled according to Rosstat data. As a tool, the EFRA software package was used, which implements the presented approach [35].

Based on the correlation analysis (the corresponding module of the EFRA program was applied with $t$-statistics for the significance level $\alpha = 0.05$), 7 out of 10 effective features and 6 out of 12 factor features (functioning factors) were selected that characterize the peculiarities of ICT use in the regions. These features were used in the construction of partial and integral performance indicators (indicators) (Table 1). All cost indicators were adjusted for the level of inflation and brought to the level of 2007.

The regression analysis module is used for building models. Since the choice of the functional form of models is not justified in the research (this requires more in-depth study) due to the lack of information about them, the linear form of models is accepted as the simplest in terms of evaluating its parameters. We used the least square method (backward selection) to build the models and insignificant factors were excluded for $t$-statistics with the level of significance $\alpha = 0.05$.

Table 1.

Results and factors (conditions)
of functioning of the Central Federal District regions

<table>
<thead>
<tr>
<th>No</th>
<th>Name of the indicator</th>
<th>Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>State and impact factors</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Average annual number of persons employed, thousand pers.</td>
<td>$x_1$</td>
</tr>
<tr>
<td>2</td>
<td>Nominal average monthly wages of employees of organizations, rubles</td>
<td>$x_2$</td>
</tr>
<tr>
<td>3</td>
<td>Percentage of households with a personal computer, %</td>
<td>$x_3$</td>
</tr>
<tr>
<td>4</td>
<td>Consolidated budget expenditures by education, million rubles</td>
<td>$z_1$</td>
</tr>
<tr>
<td>5</td>
<td>Consolidated budget expenditures by social policy, million rubles</td>
<td>$z_2$</td>
</tr>
<tr>
<td>6</td>
<td>The costs of ICT (total), million rubles</td>
<td>$z_3$</td>
</tr>
<tr>
<td></td>
<td>Results</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Percentage of households with Internet access, %</td>
<td>$y_1$</td>
</tr>
<tr>
<td>8</td>
<td>Percentage of households with broadband Internet access, %</td>
<td>$y_2$</td>
</tr>
<tr>
<td>9</td>
<td>Users of the Internet, %</td>
<td>$y_3$</td>
</tr>
<tr>
<td>10</td>
<td>Users who access the Internet every day, %</td>
<td>$y_4$</td>
</tr>
<tr>
<td>11</td>
<td>Number of connected mobile subscriber devices per 1000 population, units</td>
<td>$y_5$</td>
</tr>
<tr>
<td>12</td>
<td>The number of active subscribers of broadband access to the Internet, individuals, thousand pers.</td>
<td>$y_6$</td>
</tr>
<tr>
<td>13</td>
<td>Number of active mobile subscribers using Internet access services, thousand pers.</td>
<td>$y_7$</td>
</tr>
</tbody>
</table>
Factors related to consolidated budget expenditures were not included in the models, since the coefficients describing their impact on the performance indicator were statistically insignificant. From a qualitative point of view, this may be due to the lack of information about the targeted use of expenditures in terms of training the population in computer literacy and social support, for example, for the purchase of office equipment. Therefore, when developing proposals, we should not exclude from consideration the organization of educational events and social support measures aimed at increasing the level of use of ICT in the life of the population.

The results $\hat{y}_i^*$ and $\hat{y}_i^*$ are more related to the average annual number of persons employed than to the level of wages. At the same time, $\hat{y}_i^*$ depends on the nominal average monthly wages of employees of organizations. The results $\hat{y}_i^*$, $\hat{y}_i^*$, and $\hat{y}_i^*$ also significantly depend on the ICT costs that occur in the region.

The models presented can be used both for developing forecasts and calculating expected values of regional performance results, and for evaluating them, as well as for making management decisions.

Using the equations (1)–(3), performance indicators were calculated ($i = 1, \ldots, 7$ corresponds to the number of the performance indicator in Table 1) for the regions of the Central Federal District, for which it is possible to assess the level of the population in terms of ICT use. The results are shown in Figure 2 in 2017.

The results show that, although in absolute values the Tula Region ranks second among all the regions of the Central Federal District in most indicators, when taking into account the existing conditions (factors) identified as significant, the functioning of the Tula Region in a number of indicators does not reach the standard (expected) value. This indicates insufficient use of the region’s opportunities (potential) to increase the level of digitalization.

For the Tula Region, using the decision-making module of the EFRA complex, in the simplest case (the task of optimizing each indicator separately without taking into account restrictions), the state and impact factors were found, at which the values of indicators would reach the standard.

The solution of such a problem can be interpreted qualitatively as follows.

Interpretation 1: what are the values $\Delta x_{k,i,j}^{*} (t)$, $\Delta z_{k,i,q}^{*} (t)$ of overspending (underutilization) of state and impact factors in the $k$-th region.

Interpretation 2: what values $\Delta x_{k,i,j}^{*} (t)$, $\Delta z_{k,i,q}^{*} (t)$ should be used to intensify the use of $x_{k,i,j} (t)$ and rationalize $z_{k,i,q} (t)$ in order to achieve the norm in the $k$-th region.

The calculation results for the Tula Region are shown in Table 2. This table indicates:

$\xi_i$ corresponds to the assessment of the $i$-th result of functioning $y_i$;

the sign indicates by how many percent there is an overspend (“−”) or excess (“+”) of the factor attribute, so the norm is not reached;

0 indicates that optimization is not required for this subject;

>100 indicates that the norm cannot be reached if only one factor changes,

the dash indicates that this factor is not used in the model.

The results obtained can serve as a basis for a more in-depth analysis of the causes of underutilization (overspending) of existing conditions (factors) and further management decision-making, as well as the development of appropriate measures.

It can be concluded that the areas to focus on when developing measures to increase the level of ICT use by the population of the Tula Region include:

- the organization of educational services, including with the support of state and local governments;
Fig. 2. Values of performance indicators for the Central Federal District regions in 2017:

a) the integral (generalized) indicator $\zeta_i$; 
b) $\xi_1$;  
c) $\xi_2$;  
d) $\xi_3$;  
e) $\xi_4$;  
f) $\xi_5$;  
g) $\xi_6$;  
h) $\xi_7$. 

INFORMATION SYSTEMS AND TECHNOLOGIES IN BUSINESS
In terms of further work in the field of increasing the level of ICT use by the population using the approach presented, the following steps can be proposed:

1. In-depth assessment of conditions (state factors and impact factors) on the level of digital literacy of the population by expanding the information base of the study, both in terms of sectional and cross-sectional data in the context of municipalities and regions, with the construction of regional development models with point and interval estimates in the short and medium term;

2. The assessment of the effectiveness of digitalization management in the region;

3. The assessment of the harmony (balance) of the region’s functioning in terms of digitalization using the author’s methodology;

4. The detailed elaboration of measures, including in quantitative terms on the basis of several classes of optimization problems to be solved, aimed at ensuring balanced growth and increasing digital literacy of the region’s population.

**Conclusion**

This article presents an approach to assessing the functioning of complex systems (HSES) and decision-making, implemented in the EFRA software package. The difference between the approach used and similar ones is the ability to take into account the specific conditions for the functioning of the HSES and the formation of their own standards for them. Designed indicators allow you to compare different objects that operate in different conditions, and the procedures used to build such indicators eliminate the influence of units of measurement and economies of scale.

The integral indicator takes into account the mutual influence of particular results of the HSES functioning, which is rarely used in the formation of integral estimates. The harmony coefficient and performance indicators in combination with partial and integral performance indicators form a system of indicators for evaluating the functioning of the HSES, and the optimization method used allows you to find the values of state factors and impacts to achieve the normative values of the estimated indicators of the object of research.
Based on the example of the Central Federal District regions, the level of ICT use by the population of the subjects of the Russian Federation was analyzed for 2014-2017, and the necessary changes in the state and impact factors were found for the Tula Region, in which the evaluation indicators considered would reach the normative (expected) value. On the basis of the results obtained using the EFRA software package, a number of guidelines for the development of the region in terms of ICT and improved digital literacy of the population of the Tula Region aimed, ultimately, at increasing the readiness of the region for digitization.

References


**About the author**

**Roman A. Zhukov**

Cand. Sci. (Phys.-Math.), Associate Professor;
Researcher, Associate Professor, Department of Mathematics and Informatics, Financial University under the Government of the Russian Federation, Tula Branch, 1a, Oruzheynaya Street, Tula 300012, Russia;
E-mail: pluszh@mail.ru
ORCID: 0000-0002-2280-307X