ISSN 2587-814X (print), ISSN 2587-8158 (online)
Russian version: ISSN 1998–0663 (print), ISSN 2587–8166 (online)

Vol. 16 No. 2 – 2022

BUSINESS INFORMATICS

HSE SCIENTIFIC JOURNAL

CONTENTS

Publisher:
National Research University
Higher School of Economics

The journal is published quarterly

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

Acting Editor-in-Chief
E. Zaramenskikh

Computer Making-up:
O. Bogdanovich

Website Administration:
I. Khrustaleva

Address:
26-28, Shabolovka Street,
Moscow 119049, Russia

Tel./fax: +7 (495) 772-9590 *28509
http://bijournal.hse.ru
E-mail: bijournal@hse.ru

Circulation:
English version – 100 copies,
Russian version – 100 copies,
online versions in English and Russian – open access

Printed in HSE Printing House
44, build.2, Izmaylovskoye Shosse,
Moscow, Russia

© National Research University
Higher School of Economics

ABOUT THE JOURNAL

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: modeling of social and economic systems, digital transformation of business, innovation management, information systems and technologies in business, data analysis and business intelligence systems, mathematical methods and algorithms of business informatics, business processes modeling and analysis, decision support in management.

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 Scopus, Web of Science Emerging Sources Citation Index (WoS ESCI), Russian Science Citation Index on the Web of Science platform (RSCI), EBSCO.

International Standard Serial Number (ISSN): 2587-814X (in English), 1998-0663 (in Russian).

EDITORIAL BOARD

ACTING EDITOR-IN-CHIEF

Evgeny P. Zaramenskikh
National Research University Higher School of Economics, Moscow, Russia

EDITORIAL BOARD

Habib Abdulrab
National Institute of Applied Sciences, Rouen, France

Sergey M. Avdoshin
National Research University Higher School of Economics, Moscow, Russia

Andranik S. Akopov
National Research University Higher School of Economics, Moscow, Russia

Fuad T. Aleskerov
National Research University Higher School of Economics, Moscow, Russia

Alexander P. Afanasyev
Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia

Anton A. Afanasyev
Central Economics and Mathematics Institute, Russian Academy of Sciences, Moscow, Russia

Eduard A. Babkin
National Research University Higher School of Economics, Nizhny Novgorod, Russia

Sergey I. Balandin
Finnish-Russian University Cooperation in Telecommunications (FRUCT), Helsinki, Finland

Vladimir B. Barakhnin
Federal Research Center of Information and Computational Technologies, Novosibirsk, Russia

Alexander P. Baranov
Federal Tax Service, Moscow, Russia

Jorg Becker
University of Munster, Munster, Germany

Vladimir V. Belov
Ryazan State Radio Engineering University, Ryazan, Russia

Alexander G. Chkhartishvili
V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia

Vladimir A. Efimushkin
Central Research Institute of Communications, Moscow, Russia

Tatiana A. Gavrilova
Saint-Petersburg University, St. Petersburg, Russia

Hervé Glotin
University of Toulon, La Garde, France

Alexey O. Golosov
FORS Development Center, Moscow, Russia

Andrey Yu. Gribov
CyberPlat Company, Moscow, Russia

Alexander I. Gromoff
National Research University Higher School of Economics, Moscow, Russia

Vladimir A. Gurvich
Rutgers, The State University of New Jersey, Rutgers, USA

Laurence Jacobs
University of Zurich, Zurich, Switzerland

Liliya A. Demidova
Ryazan State Radio Engineering University, Ryazan, Russia

Iosif E. Diskin
Russian Public Opinion Research Center, Moscow, Russia

Nikolay I. Ilyin
Federal Security Guard of the Russian Federation, Moscow, Russia

Dmitry V. Isaev
National Research University Higher School of Economics, Moscow, Russia

Alexander D. Ivannikov
Institute for Design Problems in Microelectronics, Russian Academy of Sciences, Moscow, Russia

Valery A. Kalyagin
National Research University Higher School of Economics, Nizhny Novgorod, Russia

Tatiana K. Kravchenko
National Research University Higher School of Economics, Moscow, Russia

Sergei O. Kuznetsov
National Research University Higher School of Economics, Moscow, Russia

Kwei-Jay Lin
Nagoya Institute of Technology, Nagoya, Japan

Mikhail I. Lugachev
Lomonosov Moscow State University, Moscow, Russia

Svetlana V. Maltseva
National Research University Higher School of Economics, Moscow, Russia

Peter Major
UN Commission on Science and Technology for Development, Geneva, Switzerland

Boris G. Mirkin
National Research University Higher School of Economics, Moscow, Russia

Dmitry M. Nazarov
Ural State University of Economics, Ekaterinburg, Russia

Dmitry E. Palchunov
Novosibirsk State University, Novosibirsk, Russia

Panagote (Panos) M. Pardalos
University of Florida, Gainesville, USA

Óscar Pastor
Polytechnic University of Valencia, Valencia, Spain

Joachim Posegga
University of Passau, Passau, Germany

Konstantin E. Samouylov
Peoples' Friendship University, Moscow, Russia

Kurt Sandkuhl
University of Rostock, Rostock, Germany

Christine Strauss
University of Vienna, Vienna, Austria

Ali R. Sunyaev
Karlsruhe Institute of Technology, Karlsruhe, Germany

Victor V. Taratukhin
University of Munster, Munster, Germany

José M. Tribolet
Universidade de Lisboa, Lisbon, Portugal

Olga A. Tsukanova
Saint-Petersburg National Research University of Information Technologies, Mechanics and Optics, St. Petersburg, Russia

Mikhail V. Ulyanov
V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia

Raissa K. Uskenbayeva
International Information Technology University, Almaty, Kazakhstan

Markus Westner
Regensburg University of Applied Sciences, Regensburg, Germany

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.

ABOUT THE GRADUATE SCHOOL OF BUSINESS

HSE Graduate School of Business was created on September 1, 2020. The School will become a priority partner for leading Russian companies in the development of their personnel and management technologies.

The world-leading model of a 'university business school' has been chosen for the Graduate School of Business. This foresees an integrated portfolio of programmes, ranging from Bachelor's to EMBA programmes, communities of experts and a vast network of research centres and laboratories for advanced management studies. Furthermore, HSE University's integrative approach will allow the Graduate School of Business to develop as an interdisciplinary institution. The advancement of the Graduate School of Business through synergies with other faculties and institutes will serve as a key source of its competitive advantage. Moreover, the evolution and development of the Business School's faculty involves the active engagement of three professional tracks at our University: research, practice-oriented and methodological.

What sets the Graduate School of Business apart is its focus on educating and developing globally competitive and socially responsible business leaders for Russia's emerging digital economy.

The School's educational model will focus on a project approach and other dynamic methods for skills training, integration of online and other digital technologies, as well as systematic internationalization of educational processes.

At its start, the Graduate School of Business will offer 22 Bachelor programmes (three of which will be fully taught in English) and over 200 retraining and continuing professional development programmes, serving over 9,000 students. In future, the integrated portfolio of academic and professional programmes will continue to expand with a particular emphasis on graduate programmes, which is in line with the principles guiding top business schools around the world. In addition, the School's top quality and all-encompassing Bachelor degrees will continue to make valuable contributions to the achievement of the Business School's goals and the development t of its business model.

The School's plans include the establishment of a National Resource Center, which will offer case studies based on the experience of Russian companies. In addition, the Business School will assist in the provision of up-to-date management training at other Russian universities. Furthermore, the Graduate School of Business will become one of the leaders in promoting Russian education.

The Graduate School of Business's unique ecosystem will be created through partnerships with leading global business schools, as well as in-depth cooperation with firms and companies during the entire life cycle of the school's programmes. The success criteria for the Business School include professional recognition thanks to the stellar careers of its graduates, its international programmes and institutional accreditations, as well as its presence on global business school rankings.

DOI: 10.17323/2587-814X.2022.2.7.20

Reduction of dimension in the problem of optimal management of a freight cars fleet using unmanned locomotives

Fedor A. Belousov
E-mail: sky_tt@list.ru

Nerses K. Khachatryan
E-mail: nerses-khachatryan@yandex.ru

Ivan V. Nevolin
E-mail: i.nevolin@cemi.rssi.ru

Central Economics and Mathematics Institute, Russian Academy of Science
Address: 47, Nakhimovsky Prospect, Moscow 117418, Russia

Abstract

This paper considers the problem of optimal management of a fleet of freight cars by a transport railway operator. The solution to this problem is an optimal plan, which is a timetable for the movement of freight and empty railway cars, following which the transport operator will receive the maximum profit for the estimated period of time. This problem is reduced to the problem of linear programming of large dimension. Unlike the works of other authors on this topic, which mainly deal with methods of numerical solution of the corresponding linear programming problems, this article focuses on an algorithm that allows one to reduce their dimensionality. This can be achieved by excluding from the calculation those routes that obviously cannot be involved in the solution, or whose probability of participation in the final solution is estimated as extremely low. The effectiveness of the proposed modified algorithm was confirmed both on a model example (several stations, a short planning horizon) and on a real example (more than 1 000 stations, a long planning horizon). In the first case, there was a decrease in the dimension of the problem by 44%, while in the second – by 30 times.

Keywords: railway freight transportation, optimal plan, optimal management of the fleet of cars, linear programming, theory of schedules, operations research, unmanned locomotives

Citation: Belousov F.A., Khachatryan N.K., Nevolin I.V. (2022) Reduction of dimension in the problem of optimal management of a freight cars fleet using unmanned locomotives. *Business Informatics*, vol. 16, no. 2, pp. 7–20. DOI: 10.17323/2587-814X.2022.2.7.20

Introduction

One of the most popular modes of transport for cargo in the Russian Federation is rail. Publications devoted to railway logistics can be divided into the following main groups according to the type of tasks studied:

- railway network infrastructure design tasks;
- railway planning tasks;
- tasks of managing the fleet of locomotives and wagons.

In the first group, works [1–4] can be distinguished. The second group, in particular, is represented by the tasks of forming the timetable of freight trains, as well as the tasks of forming freight flows [5, 6]. One of the approaches to the formation of cargo traffic is presented in the works of Khachatryan and Beklaryan [7–16]. These articles present macroscopic dynamic models in which the process of organizing railway freight transportation is the formation of freight traffic based on the interaction of neighboring stations. These models make it possible to predict dynamics of station congestion and flows arising on the railway network by using a given procedure for organizing cargo traffic. Several configurations of sections of the railway network are considered. The first one is an extended section of the railway line which is characterized by an infinite number of stations in both directions, and also characterized by the absence of hub stations. This configuration of the transport network is suitable for describing transnational transportation (for example, transportation along the Trans-Siberian railway with a length of more than 9.000 km). The second configuration defines the movement of cargo traffic through a closed chain of stations. The third is characterized by a finite number of stations and determines the movement of cargo traffic between two hub stations.

The presented work is devoted to the problem of optimal management of a fleet of railway freight cars. Railway transport operators are faced with the task of optimal management of a fleet of freight cars to maximize profit. Such management is carried out, on the one hand, on the basis of wagons' dislocation, on the other – on the basis of requests for the transportation of goods. Requests are submitted by customers. Each request specifies stations of departure and destination, the volume of cargo transported, expressed in wagons and the rate that the customer is going to pay for each wagon of transported cargo. In addition to the rate that the customers pay to the transport operator for the provision of wagons, they also pay to Russian Railways for the transportation of loaded wagons. The costs of transporting empty wagons are covered by the transport operator. From the entire list of requests, the transport operator selects those that are most profitable for it to execute. Any selected requests can be completed either in full or in part. In accordance to dislocation of wagons and the available list of requests, creation of a wagon management plan implies preparation of a timetable for the movement of loaded and empty wagons, taking into account known time standards. Thus, the task is to find the optimal plan to manage the fleet of freight cars for a certain period of time (as a rule, the plan is drawn up for a month). In [17] a multi-commodity flow model defined on a space-time graph is presented, and an algorithm that allows reducing it to a linear programming problem is proposed.

The problems with a close formulation were previously considered in [18, 19]. In these articles, the authors focus on the methods of numerical solution of the resulting linear programming problem. In particular, they talk about the use of the column generation method [20, 21] and modification of this method [22, 23], which is based on the Danzig–Wulff decomposition [24]. A similar problem was considered in an earlier article [25]. The model presented in this paper was developed at the request of one of the largest transport operators in Latin America. Its peculiarity is that

the freight transportation plan is drawn up taking into account the previously known schedule of locomotives, whereas in this article and in the works [18, 19], the time of movement of cars on each of the routes is determined solely based on the standards of the Russian Railways (i.e., the schedule of locomotives in our case is unknown and it is not important for us; this is an internal matter of Russian Railways).

In contrast to these works, the authors of this article do not focus their attention on methods of numerical solution of linear programming problems, but offer methods for constructing space-time graphs that serve as the basis for setting both the objective function and the constraints of a linear programming problem having a smaller dimension. At the same time, the type of the objective function and constraints does not change. Thus, in this paper, a modification of the algorithm described in [17] is proposed which makes it possible to significantly reduce the dimension of the model. Before proceeding to its description, let us present the statement of the linear programming problem itself. To do this, the following notations are going to be introduced:

N – number of stations involved in planning;

T – planning horizon, measured in days; for simplicity one month is taken as the length of the planning horizon in this work (i.e. $T = 30$ or 31);

t – the discrete parameter responsible for time is measured in days and takes values $t = 1, 2, ..., T$;

$C = \left\{C_{ij}\right\}_{i,j=1}^{N}$ – $(N \times N)$-matrix, which elements characterize the tariff set by Russian Railways for an empty run of one wagon from station i to station j;

$\Theta 1 = \left\{\Theta 1_{ij}\right\}_{i,j=1}^{N}$ – $(N \times N)$-matrix, which elements characterize the time (in days) of movement of loaded wagons from station i to station j in accordance with Russian Railways standards (time is rounded to a larger integer);

$\Theta 2 = \left\{\Theta 2_{ij}\right\}_{i,j=1}^{N}$ – $(N \times N)$-matrix, which elements characterize the time (in days) of movement of empty wagons from station i to station j in accordance with Russian Railways standards (time is rounded to a larger integer);

$P = \left\{P_{ij}\right\}_{i,j=1}^{N}$ – $(N \times N)$-matrix, which elements characterize the rate specified by the customer in the request for transportation of one loaded wagon from station i to station j;

$\bar{Q} = \left\{\bar{Q}_{ij}\right\}_{i,j=1}^{N}$ – $(N \times N)$-matrix, which elements characterize the number of loaded wagons specified in the corresponding request for cargo transportation from station i to station j. All elements of the matrix take non-negative;

$\bar{S}^0(t) = \left\{\bar{S}_i^0(t)\right\}_{i=1}^{N}$ – vector of dimension N, that characterizes the initial location of wagons on day t, the i-th element of this vector equals to the number of wagons that arrived at station i at time $t \in \{1, ..., T\}$. All values of this vector take non-negative integer values.

Then the above linear programming problem takes the form

$$PC^T \cdot K \to \max_{K \geq 0}, \tag{1}$$

subject to

$$(A_{out} - A_{in}) \cdot K = S_0, \tag{2}$$

$$A_Q \cdot K \leq Q, \tag{3}$$

where

K – is a vector the first part of which is responsible for freight routes, the second part corresponds to empty routes, in fact this vector is a transportation plan;

PC – is a vector, the first part of which is responsible for freight rates, the second part corresponds to costs for empty rates, the product of $PC^T \cdot K$ gives a profit that transport operator gets for planning horizon;

A_{out} – is a matrix that takes into account the outgoing routes from each station;

A_{in} – is a matrix that takes into account the incoming routes to each station;

S_0 – is a vector of the initial distribution of wagons by stations and by time;

A_Q – is a matrix such that the product $A_Q \cdot K$ shows the volume in wagons that must be executed for each of the requests in accordance with the solution K;

Q – is a vector of the volume of orders (in wagons) specified in the requests.

Constraint (2) is a balance constraint, i.e. its implementation guarantees that in each period the number of cars entering the station will coincide with the number of cars leaving. The fulfillment of constraint (3) guarantees that the volume of executed cargo routes will not exceed the volumes specified in the requests.

The algorithm presented in [17] gives the dimension of the problem (1)–(3) equals to TN^2, i.e. the number of elements in the vectors K and PC is $2TN^2$. The dimension of the matrices A_{out} and A_{in} turns out to be $TN \times 2TN$, the dimension of the matrix A_Q is $N^2 \times 2TN^2$, the dimension of the vector S_0 is TN, the dimension of Q is N^2.

1. Algorithms for the generation of matrices and vectors for the problem (1)–(3)

The most noticeable reduction in the dimension of the problem (1)–(3) can be achieved by taking into account only freight routes in vector , the use of which will lead to partial or full execution of orders. In other words, reducing the dimension of the problem can be organized by removing from the vector those cargo routes that in any case will not be involved. In addition to vector , the corresponding transformations must be performed in all other vectors and matrices of the problem (1)–(3) so that they are all consistent.

In this section, an algorithm for generating all matrices and vectors of reduced dimension for the problem (1)–(3) is described. In addition to removing unnecessary cargo routes from consideration, the algorithm also provides for the possibility of excluding empty routes selected for any reason from the calculation.

PC, Q, S_0 vectors and dynamic lists

Let us introduce new variables, *routes_from_station_cargo*, *routes_to_station_cargo* and *routes_from_station_empty*, *routes_to_station_empty*, which are dynamic lists with elements taking integer values. The elements of these variables are responsible for the numbers of those routes that are going to be taken into account in the calculation. The *routes_from_station_cargo* list contains numbers of outgoing stations for each of the considered cargo routes; *routes_to_station_cargo* contains the corresponding numbers of incoming stations for the same cargo routes. Similarly, *routes_from_station_empty* contains outgoing station numbers for empty routes; *routes_to_station_empty* contains incoming station numbers for corresponding empty routes. If we take into account all possible routes, as is done in [17], then the number of elements contained in the new variables is going to be equal to N^2. However, these variables were introduced so that fewer routes could be taken into account, thereby reducing the dimension of the problem.

Let us fill in the variables *routes_from_station_cargo* and *routes_to_station_cargo*. We will take into account only those routes that are in the requests; therefore, if there is a request from station i to station j, that is, $P_{ij} > 0$ then add the element i to the variable *routes_from_station_cargo* on the right, and add the element j to the variable *routes_from_station_cargo* on the right. Simultaneously with each addition of elements to the variables *routes_from_station_cargo* and *routes_to_station_cargo*, we will compose a vector p by sequentially adding elements P_{ij}

from below, and we will also compose a vector Q by sequentially adding elements $\overline{Q}_{ij}$ from below characterizing the volumes of the corresponding applications. Thus, at each iteration, the dimension of the vectors p and Q, as well as the variables *routes_from_station_cargo* and *routes_to_station_cargo* increases by one.

It can be seen that the resulting vector Q takes into account only cargo routes, and not all possible routes, as was in [17]. Due to this, the dimension of the vector Q is reduced from N^2 to N_{cargo}.

It is assumed that for each pair of stations i and j, there can be no more than one request from station i to station j. If there are two requests for a pair of stations i and j from i to j, then in this case a duplicate of station м is created, let us call it $\hat{i}$, and in the variables *routes_from_station_cargo* and *routes_to_station_cargo* not only i and j are added, but also $\hat{i}$ and j, respectively. In the case of two bids, the vector p is filled with the corresponding bids in the same sequence as the variables *routes_from_station_cargo* and *routes_to_station_cargo*. Communication between station i and its duplicate $\hat{i}$ is instantaneous and free of charge. At the same time, there is no back route from station i to station $\hat{i}$. All incoming routes are directed to station i, one can get to station $\hat{i}$ only through station i. This is done so that cyclic flows from station i to $\hat{i}$ and back do not appear in solutions. The case when there may be more than two requests from station i to station j is not considered in this article – this is a separate topic, the disclosure of which we will leave for subsequent works in this direction.

Similarly, the *outes_from_station_empty* and *routes_to_station_empty* variables are filled with station numbers only of those routes that were selected according to some criteria. Simultaneously with filling in the variables *routes_from_station_empty* and *routes_to_station_empty*, by analogy with the vector p, a new vector $\tilde{C}$ is filled by adding elements C_{ij} from below (the cost of an empty run from station i to station j). The order of adding elements to vector $\tilde{C}$ corresponds to the order of adding elements to the variables *routes_from_station_empty* and *routes_to_station_empty*. Thus, at each iteration, the dimension of the vector $\tilde{C}$, as well as the variables *routes_from_station_empty* and *routes_to_station_empty* increases by one.

One way to reduce the dimension due to empty routes is to remove from consideration those empty routes, the arrival stations in which are not departure stations for any of the requests for loaded routes. The idea is that there is no need to come to such stations by empty routes, since cars can only leave there by other empty routes, which is unlikely to be optimal. The exception is routes from these stations to themselves, it is better to leave such routes in numerical calculation, since either wagons from the previous period arrive at these stations, or they are the final destination for some loaded routes. Here it is necessary to make a reservation that this method of removing empty routes from consideration is justified, provided that the time and financial costs of empty runs from an arbitrary station A to an arbitrary station B are always no more than when carrying out two consecutive empty runs from A to some station C and from C to B. In practice, this condition is fulfilled, so the exclusion of such double empty routes from consideration does not lead to a deterioration in the target indicators of the solutions obtained.

Using N_{cargo} we denote the dimension of vector p, which coincides with the number of elements in the variables *routes_from_station_cargo* and *routes_to_station_cargo*; using N_{empty} we denote the dimension of the vector $\tilde{C}$, which also coincides with the number of elements in the variables *routes_from_station_empty* and *routes_to_station_empty*. In other words, N_{cargo} characterizes the number of all cargo routes corresponding to the list of requests; N_{cargo} characterizes the number of all possible empty routes that are taken into account when searching for the optimal plan.

Let us construct a vector PC of a smaller dimension compared to a similar vector in [17]. To do this, we perform $T-1$ consecutive concatenation of the vector p so that we get a vector of dimension $T \cdot N_{cargo}$. Next, we also add the vector $\tilde{C}$ to the resulting vector by sequential concatenation $T-1$ times. Assign the value of the resulting vector to the vector PC; its dimension of this vector is $T \cdot (N_{cargo} + N_{empty})$. The new dimension of the problem (1)–(3) is also equal to $T \cdot (N_{cargo} + N_{empty})$. Obviously, the elements of vector K correspond to the same routes that correspond to both the rates and costs of vector PC.

The algorithm for creating vector S_0 described in [17] will remain unchanged. Namely, the system of vectors $\overline{S}^0(t)$, $t \in \{1, ..., T\}$ is transformed into a vector S_0 by sequential concatenation of vectors corresponding to each moment of time. With this concatenation, the first N elements of the new vector correspond to the vector $\overline{S}^0(1)$, the next N elements correspond to the vector $\overline{S}^0(2)$, etc. Thus, the dimension of the vector S_0 equals to TN.

A_{in} and A_{out} matrices

Each of the A_{in} and A_{out} matrices consists of two parts. The first part is responsible for cargo routes, the second for empty ones. These matrices are sparse matrices; any nonzero element in them takes a single value. Let us construct A_{out} matrix, which is responsible for outgoing routes originating from each station. At the zero iteration, the A_{out} matrix is a zero matrix of size $T \cdot N \times T \cdot (N_{cargo} + N_{empty})$.

Let us denote by $Index\_cargo\_out[1]$ a dynamic list of indexes $k \in \{1, ..., N_{cargo}\}$, for which $routes\_from\_station\_cargo[k] = 1$ In other words, the variable $Index\_cargo\_out[1]$ contains the numbers of those freight routes among the first N_{cargo} elements of vector K, the starting point for which is station 1. For an arbitrary station $i \in \{1, ..., N\}$, the interpretation of the variable $Index\_cargo\_out[i]$ is similar.

To account for outgoing routes from station 1 in the first time period, it is necessary for all $k \in Index\_cargo\_out[1]$ elements $A_{out}[1, k]$ to be assigned the value 1. To account for outgoing routes from station 1 in the second time period, it is necessary to assign value 1 to the elements $A_{out}[1 + N, k + N_{cargo}]$, $k \in Index\_cargo\_out[1]$. N is added to the first component of the coordinates of the A_{out} matrix, since the period in the vector S_0 is equal to N. In other words, the first N elements in this vector are responsible for N stations in the first time period, and the next N elements are responsible for the same N stations in the second time period, etc. N_{cargo} elements are added to the second component of coordinates of the A_{out} matrix, since the period in the first part of the PC vector responsible for the rates of loaded runs is equal to N_{cargo}. In other words, the first N_{cargo} elements in the first part of the PC vector are responsible for routes starting in the first time period; the next N_{cargo} elements are responsible for the same routes, but starting in the second time period, and so on. In other words, the first N_{cargo} elements in the first part of the PC vector are responsible for routes starting in the first time period; the next N_{cargo} elements are responsible for the same routes, but starting in the second time period, and so on. Continuing this logic further, it is clear that all elements of matrix A_{out} with coordinates $[1 + (t-1) \cdot N, k + (t-1) \cdot N_{cargo}]$, $k \in Index\_cargo\_out[1]$ have to be assigned to value 1. Thus, to obtain the A_{out} matrix, it is necessary for each station $i \in \{1, ..., N\}$ to create a dynamic list $Index\_cargo\_out[i]$ with such numbers $k \in \{1, ..., N_{cargo}\}$ for which $routes\_from\_station\_cargo[k] = i$. Further, for all $i \in \{1, ..., N\}$ for which $Index\_cargo\_out[i] \neq 0$, elements of A_{out} matrix with coordinates $[i + (t-1) \cdot N, k + (t-1) \cdot N_{cargo}]$, $k \in Index\_cargo\_out[i]$, $t \in \{1, ..., T\}$ have to be assigned to value 1. The first part of the A_{out} matrix responsible for cargo routes has been formed. Similarly, the second part of this matrix is formed, which is responsible for empty routes. To do this, for each station

$i \in \{1, \ldots, N\}$, other *Index_empty_out*[i] lists are formed with the following numbers $k \in \{1, \ldots, N_{empty}\}$, for which *routes_from_station_empty*[k] = i. In other words, the variable *Index_cargo_out*[i] contains the numbers of those empty routes among the N_{empty} elements of the vector K following $T \cdot N_{cargo}$ elements, the starting point for which is station i.

Further, for all $i \in \{1, \ldots, N\}$ for which *Index_empty_out*[i] $\neq 0$, elements of A_{out} matrix with coordinates $[i + (t-1) \cdot N, k + (t-1) \cdot N_{empty} + T \cdot N_{cargo}]$, $k \in$ *Index_empty_out*[i], $t \in \{1, \ldots, T\}$ have to be assigned to value 1. Since the first $T \cdot N_{cargo}$ elements in vector K are responsible for cargo routes, the rest are responsible for empty ones. Then in the case of considering empty routes in the second component of the A_{out} matrix, there is an additional term $T \cdot N_{cargo}$. After performing all the described operations, the construction of the A_{out} matrix is completed.

At the next stage, the A_{in} matrix is formed. This matrix is responsible for the incoming routes to each station. At the zero iteration A_{in} matrix is a zero matrix of size $T \cdot N \times T \cdot (N_{cargo} + N_{empty})$. To form this matrix, we additionally need information about the travel time for each of the routes, that is, the values of the matrices $\Theta 1$ and $\Theta 2$. Denote by *Index_cargo_in*[1] a dynamic list of those indexes $k \in \{1, \ldots, N_{cargo}\}$ of the variable *routes_to_station_cargo*[k] for which *routes_to_station_cargo*[k] = 1. For an arbitrary station $i \in \{1, \ldots, N\}$, the interpretation of the variable *Index_cargo_in*[1] is similar. Then, in order to account for incoming routes to station 1, departures on which are carried out in the first time period, elements of A_{in} matrix with coordinates $[1 + \Theta 1[routes\_from\_station\_cargo[k], 1] \cdot N, k]$, $k \in$ *Index_cargo_in*[1] have to be assigned to value 1. Similarly, to account for incoming routes to station 1, departures on which are carried out in the time period $t \in \{1, \ldots, T\}$, elements of A_{in} matrix with coordinates $[1 + (\Theta 1[routes\_from\_station\_cargo[k], 1] + t - 1) \cdot N, k + (t-1) \cdot N_{cargo}]$, $k \in$ *Index_cargo_in*[1] have to be assigned to value 1.

For an arbitrary station $i \in \{1, \ldots, N\}$ dynamic variables *Index_cargo_in*[i] are compiled from those indexes $k \in \{1, \ldots, N_{cargo}\}$ of the variable *routes_to_station_cargo*[k] for which *routes_to_station_cargo*[k] = i. For all $i \in \{1, \ldots, N\}$ for which *Index_cargo_in*[i] $\neq 0$, elements of A_{in} matrix with coordinates $[i + (\Theta 2[routes\_from\_station\_cargo[k], i] + t - 1) \cdot N, k + (t-1) \cdot N_{cargo}]$, $k \in$ *Index_cargo_in*[i], $t \in \{1, \ldots, T\}$ have to be assigned to value 1.

The first part of the A_{in} matrix relating to loaded routes is constructed. It remains to construct the second part of this matrix relating to empty routes. Denote by *Index_empty_in*[i] a dynamic list of those indexes $k \in \{1, \ldots, N_{cargo}\}$ of the variable *routes_to_station_empty*[k] for which *routes_to_station_empty*[k] = i. For all stations $i \in \{1, \ldots, N\}$ for which *Index_empty_in*[i] $\neq 0$, elements of A_{in} matrix with coordinates $[i + (\Theta 2[routes\_from\_station\_cargo[k], i] + t - 1) \cdot N, k + (t-1) \cdot N_{empty} + T \cdot N_{cargo}]$, $k \in$ *Index_empty_out*[i], $t \in \{1, \ldots, T\}$ have to be assigned to value 1.

A_Q matrix

The A_Q matrix is needed to calculate the volume of completed requests, so only loaded routes are taken into account when calculating this indicator. This means that in the matrix A_Q, which has dimension $N_{cargo} \times T \cdot (N_{cargo} + N_{empty})$, the last $T \cdot N_{empty}$ columns consist exclusively of zero elements, non-zero elements are found only in the first $T \cdot N_{cargo}$ columns. At the zero iteration, we take the zero matrix as the A_Q matrix.

Since the first $T \cdot N_{cargo}$ elements of vector K are ordered with the period N_{cargo}, i.e. the first N_{cargo} elements are responsible for loaded routes outgoing in the first time period, the next N_{cargo} elements are responsible for the same routes outgoing in the second time period, etc. Then in the first row of A_Q matrix T units are written, the first of which is put on the first position, the

next to the position N_{cargo} + 1, the next to the position $2N_{cargo}$ + 1, etc. In other words, in the first row of A_Q matrix, T elements are assigned to the unit starting from the first element and then with the period N_{cargo} elements. In the next row of the A_Q matrix, the unit is also assigned T elements with the period N_{cargo}, but starting from the second element of the second row. In the third row of the A_Q matrix, the algorithm is repeated, but the unit is assigned elements starting from the third element of the third row. This continues until the last line of N_{cargo}. As a result, if we consider the first N_{cargo} rows and the first N_{cargo} columns, we get a unit matrix, if we consider the next N_{cargo} columns, we also get a unit matrix, etc. If we consider the first $T \cdot N_{cargo}$ columns of the matrix A_Q, we will see T sequentially composed unit matrices of dimension $N_{cargo} \times N_{cargo}$, the remaining columns of the matrix are zero.

This section provides algorithms for the construction of all components of the problem (1)–(3) – objective function and constraints. It is shown that the dimension of both the vector of variables of the objective function and the constraint matrices is noticeably reduced. We will demonstrate this both on a model example (several stations, a short planning horizon) given in [17] and on a real example (more than 1,000 stations, long planning horizon).

2. Reducing the dimensionality of the problem on model and real examples

Here is a statement of the model example from [17]. The number of stations is 4 ($N = 4$), the planning horizon is 3 days ($T = 3$). The list of received requests consists of five items, which are shown in *Table 1*.

Based on the list of requests, it is necessary to make two matrices – a matrix of rates P, the elements of which are written in conventional units, and a matrix of request volumes $\bar{Q}$:

$$P = \begin{pmatrix} 0 & 0 & 2.9 & 0 \\ 1.1 & 0 & 2.3 & 0 \\ 0 & 1.9 & 0 & 2.1 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \ \bar{Q} = \begin{pmatrix} 0 & 0 & 3 & 0 \\ 5 & 0 & 4 & 0 \\ 0 & 7 & 0 & 6 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Next, we will give the travel time of both loaded and empty routes in the form of the values of the matrices $\Theta 1$ and $\Theta 2$:

$$\Theta 1 = \begin{pmatrix} 0 & 2 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 \\ 1 & 2 & 1 & 0 \end{pmatrix}; \ \Theta 2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 \\ 1 & 2 & 1 & 1 \end{pmatrix}.$$

Recall that the diagonal elements of the matrix $\Theta 2$ are equal to one. This is due to the fact that if the cars need to be left at the station until the next day, then this is equivalent to the

Table 1.

List of requests for cargo transportation in the model example

No.	Departure station	Destination station	Volume of requests (in wagons)	Rate (in conditional units)
1	1	3	3	2.9
2	2	1	5	1.1
3	2	3	4	2.3
4	3	2	7	1.9
5	3	4	6	2.1

fact that they are, as it were, sent from this station to itself on a one-day run.

The values of the Russian Railways tariffs for empty runs, as well as the rates expressed in conventional units, are characterized by the values of the elements of the matrix C:

$$C = \begin{pmatrix} 0 & 1.9 & 1.3 & 1.9 \\ 1.2 & 0 & 1.8 & 0\ 9 \\ 1.1 & 1.2 & 0 & 1.6 \\ 1.3 & 1.5 & 1.2 & 0 \end{pmatrix}.$$

As part of this task, it is assumed that cars can stay at stations until the next day for free, so the diagonal elements of the matrix C are zero.

The initial distribution of wagons is characterized by the following vectors:

$$\overline{S}^0(1) = \begin{pmatrix} 0 \\ 2 \\ 1 \\ 3 \end{pmatrix}; \quad \overline{S}^0(2) = \begin{pmatrix} 5 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

During the period $t = 3$, the wagons do not arrive, which is equivalent to the zero vector $\overline{S}^0(3)$.

Dimension of the problem

The dimension of the problem presented in [17] is $2TN^2 = 96$. We calculate dimension when solving the same problem using the algorithm presented in this paper. The proposed algorithm gives the dimension $T \cdot (N_{cargo} + N_{empty})$. Therefore, in order to calculate it, it is necessary to know the values of the parameters N_{cargo} and N_{empty}. To determine N_{empty}, it is necessary to understand which empty routes are planned to be included in the calculation, which are not. We exclude from consideration those empty routes, the arrival stations in which are not departure stations for any loaded runs from clients' requests. There is one such station and this is station four. We will remove from consideration all empty routes in which the destination is station 4. We will leave only the route from 4 to 4 (the car remains at the station until the next period). It turns out that empty routes from 1 to 4, from 2 to 4 and from 3 to 4 are removed from consideration. We get that $N_{empty} = N^2 - 3 = 13$. As for N_{cargo}, its value is equal to the number of requests, i.e. in our case $N_{cargo} = 5$. Thus, the dimension when using the new algorithm turns out to be equal to $T \cdot (N_{cargo} + N_{empty}) = 54$. It turns out that specifically for this example the dimension of the problem has decreased by about 44%. Separately, we note that in practice, empty routes can be excluded from the calculation for other reasons, for example, empty routes can be ignored, the tariff for which is higher than a certain threshold value. Therefore, in real problems, it is possible to achieve an even greater reduction in dimension compared to $2TN^2$.

As an example, we can consider the problem of finding the optimal plan which was solved in practice for $N = 1126$ stations, with a planning period of $T = 30$ days and the number of requests for cargo transportation equal to 1616. The dimension of the problem when solving it by the algorithm from [17] is $TN^2 = 76072560$. To determine the dimension of the problem, which is obtained using the algorithm presented in this paper, one needs to calculate N_{cargo} and N_{empty}. Obviously, $N_{cargo} = 1616$, which corresponds to the number of requests. To calculate N_{empty}, it is necessary to remove from consideration all empty routes in the direction to stations that do not appear in requests as departure stations. In addition, empty routes with tariffs exceeding 50 000 rubles are removed from consideration. As a result, the number of empty routes that should be taken into account in the calculation equals to $N_{empty} = 82058$, i.e. approximately 6.5% of all possible empty routes take part in calculation; the number of all empty routes is equal to $N^2 = 1267877$. As a result, the dimension of the problem is equal to $T \cdot (N_{cargo} + N_{empty}) = 2510220$, i.e. it decreases by about 30 times.

Linear programming

One can write out the linear programming problem (1)–(3) for the model example. To do this, we determine the values of the matrices A_{in}, A_{out} and A_Q, as well as the vectors PC, S_0, Q. After that, we solve this problem and compare the resulting solution with the solution from [17].

Dynamic lists *routes_from_station_cargo*, *routes_to_station_cargo* and *routes_from_station_empty*, *routes_to_station_empty* are the following:

routes_from_station_cargo = {1, 2, 2, 3, 3},

routes_to_station_cargo = {3, 1, 3, 2, 4},

routes_from_station_empty = {1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4},

routes_to_station_empty = {1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 4}.

One can make up the vectors p and c:

$p = (2.9 \quad 1.1 \quad 2.3 \quad 1.9 \quad 2.1)^T$

$c = (0 \quad 1.9 \quad 1.3 \quad 1.2 \quad 0 \quad 1.8 \quad 1.1 \quad 1.2 \quad 0 \quad 1.3 \quad 1.5 \quad 1.2 \quad 0)^T$.

The PC vector is obtained by sequential concatenation of the resulting vectors:

$PC = (p^T, p^T, p^T, c^T, c^T, c^T)^T$.

Vector Q, which is responsible for the volume of requests, takes the following form:

$Q = (3 \quad 5 \quad 4 \quad 7 \quad 6)^T$.

The vector S_0, which characterizes the initial distribution of wagons by time and by stations, takes the form:

$S_0 = (0 \quad 2 \quad 1 \quad 3 \quad 5 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0)^T$

We get the matrices A_{in}, A_{out} and A_Q. Since these matrices are sparse matrices, and nonzero elements can only take single values, therefore, as in [17], one can write these matrices in a sparse format, specifying the coordinates of the elements taking value 1. Let us write out the coordinates of unit elements of A_Q matrix, the dimension of which is equal to $N_{cargo} \times T \cdot (N_{cargo} + N_{empty}) = 5 \times 54$ (hereafter the numbering of rows and columns begins with one):

(1, 1), (2, 2), (3, 3), (3, 4), (4, 5), (1, 6), (2, 7), (3, 8), (4, 9), (5, 10), (1, 11), (2, 12), (3, 13), (4, 14), (5, 15).

Coordinates of unit elements of A_{in} matrix are the following:

(7, 1), (5, 2), (11, 3), (10, 4), (12, 5), (11, 6), (9, 7), (5, 16), (6, 17), (7, 18), (5, 19), (6, 20), (7, 21), (5, 22), (6, 23), (7, 24), (5, 25), (10, 26), (7, 27), (8, 28), (9, 29), (10, 30), (11, 31), (9, 32), (10, 33), (11, 34), (9, 35), (10, 36), (10, 37), (9, 38), (11, 40), (12, 41).

The list of coordinates of the unit elements of A_{ou} matrix:

(1, 1), (2, 2), (2, 3), (3, 4), (3, 5), (5, 6), (6, 7), (6, 8), (7, 9), (7, 10), (9, 11), (10, 12), (10, 13), (11, 14), (11, 15), (1, 16), (1, 17), (1, 18), (2, 19), (2, 20), (2, 21), (3, 22), (3, 23), (3, 24), (4, 25), (4, 26), (4, 27), (4, 28), (5, 29), (5, 30), (5, 31), (6, 32), (6, 33), (6, 34), (7, 35), (7, 36), (7, 37), (8, 38), (8, 39), (8, 40), (8, 41), (9, 42), (9, 43), (9, 44), (10, 45), (10, 46), (10, 47), (11, 48), (11, 50), (12, 51), (12, 52), (12, 53), (12, 54).

The dimension of the matrices A_{in} and A_{ou} is $T \cdot N \times T \cdot (N_{cargo} + N_{empty}) = 12 \times 54$.

Let us write out the solution that was obtained by using MatLab. Since vector K, consisting of $T \cdot (N_{cargo} + N_{empty}) = 54$ elements, also mainly consists of zero elements, we write out values of only non-zero elements:

$K_3 = 2$; $K_4 = 1$; $K_6 = 3$; $K_{13} = 2$; $K_{14} = 4$; $K_{15} = 6$; $K_{26} = 1$; $K_{28} = 2$; $K_{31} = 2$; $K_{40} = 3$.

One can write out the same solution in a more understandable format of matrices $K1(t)$ and $K2(t)$, which are $(N \times N)$-matrices. Elements of these matrixes characterize the number of loaded ($K1(t)$) and empty ($K2(t)$) wagons sent from station i to station j at the time $t \in \{1, \ldots, T\}$.

$$K1(1)=\begin{pmatrix}0&0&0&0\\0&0&2&0\\0&1&0&0\\0&0&0&0\end{pmatrix};\ K1(2)=\begin{pmatrix}0&0&3&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{pmatrix};$$

$$K2(1)=\begin{pmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&1&0&2\end{pmatrix};\ K1(3)=\begin{pmatrix}0&0&0&0\\0&0&2&0\\0&4&0&6\\0&0&0&0\end{pmatrix};$$

$$K2(2)=\begin{pmatrix}0&0&2&0\\0&0&0&0\\0&0&0&0\\0&0&3&0\end{pmatrix};\ K2(3)=\begin{pmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{pmatrix}.$$

Graphically, this solution is shown in *Fig. 1*. The width of each lane in this picture characterizes the number of wagons sent in a given direction.

The target value of profit, calculated according to the rule $PC^T \cdot K$, is equal to 32.3.

If we compare the obtained solution with the solution from [17], it is clear that they differ, but the values of the target functional expressing the final profit are the same. Thus, the comparison of solutions to the same problem shows that the problem has at least two different solutions.

Conclusion

This article is a continuation of the work [17]. It presents a modified algorithm for solving the problem of optimal management of a fleet of freight cars. The essence of the proposed approach is to exclude from the calculation those loaded or empty routes about which it is known in advance that they either will not be involved in the final solution or the probability of these routes appearing in the solution is estimated as very low. The model example presented in the paper shows that the use of

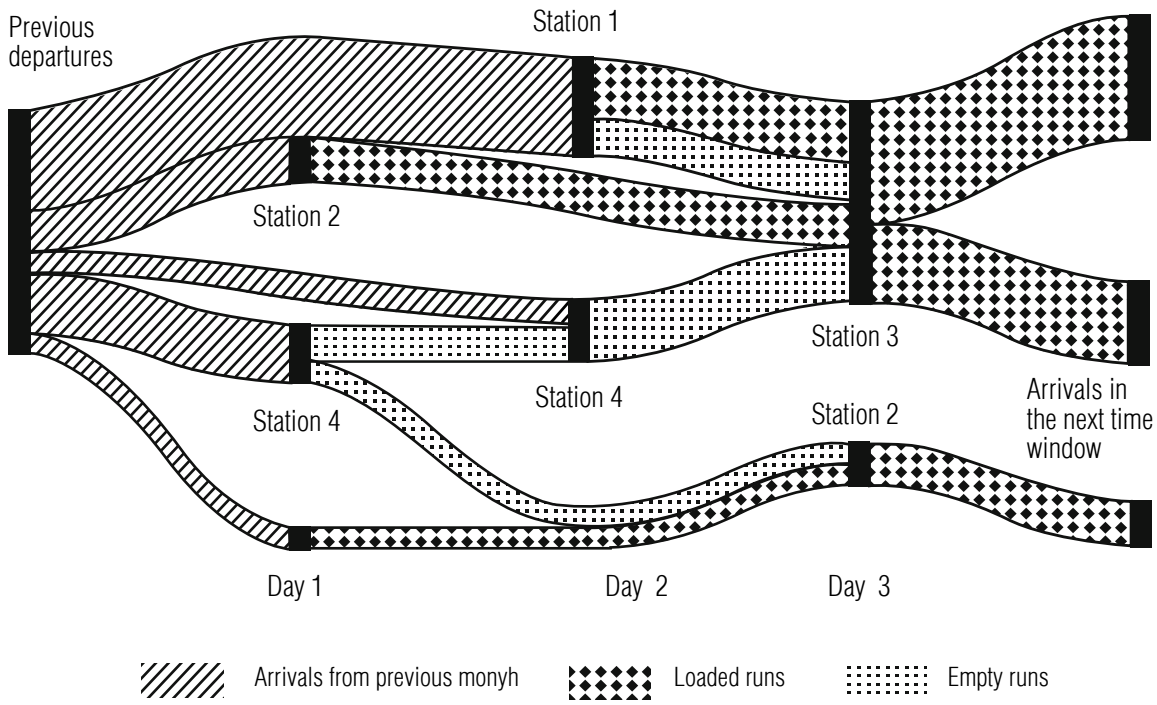

Fig. 1. Schematic representation of the resulting solution.

the modified algorithm leads to a reduction in dimension by about 44%. In practice, as a rule, there is a much more noticeable decrease in dimensionality, in particular because the exclusion of an even larger number of empty routes from the calculation due to additional features (for example, to exclude too expensive, too long empty routes). So, in the problem mentioned in the previous section, which was solved on the basis of real data, the use of an improved algorithm leads to a thirty-fold reduction in dimension compared to the algorithm from [17].

Separately, we note that the potential of methods that allow us to significantly reduce the dimension for transport problems has not been fully exhausted. It can be shown that the space-time graph that is being constructed within the framework of the presented approach can be reduced even more without loss in the quality of final solutions (reducing the space-time graph will obviously lead to a decrease in the dimension of the transport problem). To do this, one can divide all stations into three categories. The first category includes stations to which wagons arrive from the previous period and which do not participate in requests for cargo transportation either as departure stations or as destination stations. The second category includes stations that appear in the requests as destination stations, but not as departure ones, and the third category includes the remaining stations, that is, the stations indicated in the requests as departure ones. For the first category of stations, one can build outgoing empty routes exclusively for those days in which cars arrive from previous period (previous month) only to stations of the third category. In other words, as soon as wagons get to these stations, they are immediately sent by an empty run to the stations from which requests for cargo transportation can be executed. For stations of the second category, incoming empty routes are not built, but only outgoing empty routes are built in stations of the third category. For stations of the third category, a full-fledged space-time graph with incoming and outgoing empty routes is being built. The description of the specified algorithm may become the subject of one of the following articles in this direction. Reducing the space-time graph, and hence the dimension of the transport problem, can be achieved by other more subtle methods. For example, when constructing a space-time graph for stations of the second and third types, it is possible to additionally take into account from what earliest moment in time wagons may begin to appear in these stations and not to build a graph for the corresponding stations until this moment of model time. In this case, in the struggle to reduce the dimension, the only payment is an even greater complication of the algorithms for the formation of matrices and vectors for the problem (1)–(3), which in turn increases the probability of errors when creating such algorithms.

In addition to efforts to further optimize algorithms for the formation of matrices and vectors, another direction for the development of this type of tasks is modernization of the formulation of the optimal management problem of the fleet of freight cars in order to take into account more restrictions. In the current version, the transport problem is of exclusively scientific interest, but not practical in any way. For railway transport operators, who are the main customers of such models, it is important to be able to take into account a sufficiently large number of factors, among which is the possibility to take into account various types of wagons, the prohibition for some types of wagons to enter certain territories, accounting of sediment stations, restrictions on the minimum or maximum number of wagons that must move in the specified directions during the planning horizon etc. The study of the problems described above may be the subject of future research. ■

Acknowledgments

The reported study was funded by RFBR, project number 19-29-06003.

References

1. Higgins A., Ferreira L., Kozan E. (1995) Modeling single-line train operations. Transportation Research Record, vol. 1489, pp. 9–16.
2. Kraay D., Barker P., Chen B. (1991) Optimal pacing of trains in freight railroads: model formulation and solution. Operations Research, vol. 39, no. 1, pp. 82–99. https://doi.org/10.1287/opre.39.1.82
3. Ferreira L., Murray M. (1997) Modelling rail track deterioration and maintenance: current practices and future needs. Transport Reviews, vol. 17, no. 3, pp. 207–221. https://doi.org/10.1080/01441649708716982
4. LeBlanc L. (1976) Global solutions for a nonconvex nonconcave rail network model. Management Science, vol. 23, no. 2, pp. 131–139. https://doi.org/10.1287/mnsc.23.2.131
5. Lazarev A., Musatova E., Kvaratskheliya A., Grafov E. (2012) Schedule theory. Problems of transport system management. Moscow: MSU (in Russian).
6. Lazarev A., Musatova E., Grafov E., Kvaratskheliya A. (2012) Schedule theory. Problems of railway planning. Moscow: ICS RAS (in Russian).
7. Beklaryan L., Khachatryan N. (2006) Traveling wave type solutions in dynamic transport models. Functional Differential Equations, vol. 13, no. 2, pp. 125–155.
8. Beklaryan, L., Khachatryan N. (2013) On one class of dynamic transport models. Computational Mathematics and Mathematical Physics, vol. 53, no. 10, pp. 1649–1667. https://doi.org/10.7868/S0044466913100037
9. Khachatryan N. (2013) Dynamic model of organization of cargo transportation with a limited capacity of the distillation ways. Business Informatics, vol. 26, no. 4, pp. 62–68.
10. Khachatryan N., Akopov A. (2017) Model for organizing cargo transportation with an initial station of departure and a final station of cargo distribution. Business Informatics, vol. 39, no. 1, pp. 25–35. https://doi.org/10.17323/1998-0663.2017.1.25.35
11. Khachatryan N., Akopov A., Belousov F. (2018) About quasi-solutions of traveling wave type in models for organizing cargo transportation. Business Informatics, vol. 43, no.1, pp. 61–70. https://doi.org/10.17323/1998-0663.2018.1.61.70
12. Khachatryan N.K., Beklaryan G.L., Borisova S.V., Belousov F.A. (2019) Research into the dynamics of railway track capacities in a model for organizing cargo transportation between two node stations. Business Informatics, vol. 13, no. 1, pp. 59–70. https://doi.org/10.17323/1998-0663.2019.1.59.70
13. Beklaryan L., Khachatryan N., Akopov A. (2019) Model for organization cargo transportation at resource restrictions. International Journal of Applied Mathematics, vol. 32, no. 4, pp. 627–640. https://doi.org/10.12732/ijam.v32i4.7
14. Khachatryan N., Beklaryan L. (2021) Study of flow dynamics in the model of cargo transportation organization along a circular chain of stations. Economics and Mathematical Methods, vol. 57, no. 1, pp. 83–91. https://doi.org/10.31857/S042473880013024-5
15. Khachatryan N. (2020) Study of flow dynamics in the model of cargo transportation organization between node stations. International Journal of Applied Mathematics, vol. 33, no. 5, pp. 937–949. https://doi.org/10.12732/ijam.v33i5.14
16. Khachatryan N. (2021) Modeling the process of cargo transportation between node stations. International Journal of Applied Mathematics, vol. 34, no. 6, pp. 1223–1235. https://doi.org/10.12732/ijam.v34i6.12
17. Belousov F.A., Nevolin I.V., Khachatryan N.K. (2020) Modeling and optimization of plans for railway freight transport performed by a transport operator. Business Informatics, vol. 14, no. 2, pp. 21–35. https://doi.org/10.17323/2587-814X.2020.2.21.35
18. Sadykov R., Lazarev A., Shiryaev V., Stratonnikov A. (2013) Solving a freight railcar flow problem arising in Russia. Proceedings of the 13th Workshop on Algorithmic Approach for Transportation Modelling, Optimization, and Systems (ATMOS'13), Leibniz, Germany, 5 September 2013, pp. 55–67. https://doi.org/10.4230/OASIcs.ATMOS.2013.55

19. Lazarev A., Sadykov R. (2014) Management problem of railway cars fleet. Proceedings of the XII All-Russian Meeting on Management Issues (VSPU 2014), ICS RAS, Moscow, Russia, 16–19 June 2014, pp. 5083–5093 (in Russian).
20. Desaulniers G., Desrosiers J., Solomon M. (2005) Column generation. New York: Springer. https://doi.org/10.1007/b135457
21. Lübbecke M. (2011) Column generation. Wiley Encyclopedia of Operations Research and Management Science. John Wiley & Sons. https://doi.org/10.1002/9780470400531.eorms0158
22. Frangioni A., Gendron B. (2009) 0–1 reformulations of the multicommodity capacitated network design problem. Discrete Applied Mathematics, vol. 157, no. 6, pp. 1229–1241. https://doi.org/10.1016/j.dam.2008.04.022
23. Sadykov R., Vanderbeck F. (2013) Column generation for extended formulations. EURO Journal on Computational Optimization, vol. 1, nos. 1–2, pp. 81–115. https://doi.org/10.1007/s13675-013-0009-9
24. Dantzig G., Wolfe P. (1960) Decomposition principle for linear programs. Operations Research, vol. 8, no. 1, pp. 101–111. https://doi.org/10.1287/opre.8.1.101
25. Fukasawa R., Aragão M.P., Porto O., Uchoa E. (2002) Solving the freight car flow problem to optimality. Electronic Notes in Theoretical Computer Science, vol. 66, no. 6, pp. 42–52. https://doi.org/10.1016/S1571-0661(04)80528-0

About the authors

Fedor A. Belousov

Cand. Sci. (Econ.);

Senior Researcher, Laboratory of Dynamic Models of Economy and Optimization, Central Economics and Mathematics Institute, Russian Academy of Science, 47, Nakhimovsky Prospect, Moscow 117418, Russia;

E-mail: sky_tt@list.ru

ORCID: 0000-0002-3040-3148

Nerses K. Khachatryan

Cand. Sci. (Phys.-Math.);

Deputy Director of CEMI RAS for Scientific Work, Central Economics and Mathematics Institute, Russian Academy of Science, 47, Nakhimovsky Prospect, Moscow 117418, Russia;

E-mail: nerses-khachatryan@yandex.ru

ORCID: 0000-0003-2495-5736

Ivan V. Nevolin

Cand. Sci. (Econ.);

Leading Researcher, Laboratory of Experimental Economics, Central Economics and Mathematics Institute, Russian Academy of Science, 47, Nakhimovsky Prospect, Moscow 117418, Russia;

E-mail: i.nevolin@cemi.rssi.ru

ORCID: 0000-0002-8462-9011

DOI: 10.17323/2587-814X.2022.2.21.35

The development of a model for a personalized learning path using machine learning methods*

Ekaterina S. Morozevich [a]
E-mail: katyamorozevich@mail.ru

Vladimir S. Korotkikh [b]
E-mail: vskor@bk.ru

Yevgeniya A. Kuznetsova [a]
E-mail: zhenya.kuz-1997@yandex.ru

[a] Reshetnev Siberian State University of Science and Technology
Address: 31, Krasnoyarsky Rabochy Av., Krasnoyarsk 660037, Russia

[b] Municipal budget general education institution secondary school No. 4
Address: 9, st. Naberezhnaya, Divnogorsk 663091, Russia

Abstract

Today, the economy is undergoing a digital transformation. Its key barriers are a lack of qualified personnel, competencies and knowledge, as well as internal resistance in organizations. It can be overcome through quality staff development and training. An urgent problem is to build a personalized learning path. Modern research is aimed at the implementation of recommendation systems in order to select relevant material. However, these recommendations are based on digital traces; the student's full personal profile, as well as organizational values are not considered. This study aims is to create an intelligent guide that would accompany an employee throughout his life in the organization, involving him in the learning process according to a personalized path based on a complex personal profile and reactions to educational material, training soft and hard skills in accordance with the values of the organization and the employee. Methods of system analysis, system engineering, psychodiagnostic research (the DISC model, Rowe's "Decision-making Style" methodology, Honey and Mumford's method of determining activity styles, psychotype test), software design and artificial intelligence (matrix factorization and neural networks) were used in

* The article is published with the support of the HSE University Partnership Programme

this study. The study was conducted on a unique database collected as part of its implementation and consisting of educational tasks for soft skills development, plus data on their implementation by users with different soft skills profiles. An intelligent guide model has been developed and implemented as a software component for an enterprise management system. The basis consists of psychodiagnostic modules, organizational management, training and recommendations. The intelligence of the system we developed allows you to qualitatively form a personalized learning path that will involve an employee not only in the learning and development process, but also in achieving organizational goals. The organization receives T-shaped specialists who have a proactive position and are capable of self-organization by investing in the development of employees. The results of this study can be used by enterprises not only at the organizational level, but also through broadcasting in the education system to form an education ecosystem in accordance with the requirements of innovative development of a given region's economy.

Keywords: training and development, supra-professional competencies, adaptability, recommendation system, machine learning, matrix factorization, neural network

Citation: Morozevich E.S., Korotkikh V.S., Kuznetsova Y.A. (2022) The development of a model for a personalized learning path using machine learning methods. *Business Informatics*, vol. 16, no. 2, pp. 21–35. DOI: 10.17323/2587-814X.2022.2.21.35

Introduction

Today, the economy is undergoing a digital transformation. This is a revolutionary process of transforming an organization's business model, not only by using digital technologies, but also by introducing fundamental organizational changes in technology, culture, operations and the principles of creating new products.

The connecting link between the integration of deep organizational transformations and new technologies are people – employees with their knowledge, skills and values. According to Deloitte Consulting LLC, 53% of organizations already understand that in the current conditions from 50 to 100% of their employees should acquire new skills and abilities [1].

Organizations are created by people with their value system's investment in the structure being created. An employee who comes to an organization at a certain stage of the life cycle does not realize what is happening, because he or she does not know what happened, and does not guess what will happen. In view of this, the urgent task is the development of personnel, taking into account the points of contact between the values of the employee and the organization.

Activities' digitalization generates an increase in data volume that requires further processing. It is data that underlies digital transformation, being the lifeblood of the transition process to the digital economy.

Already in 2020, according to Data Age Report, about 51 zettabytes of information were generated by mankind, and by 2025 this data volume will increase almost 3.5 times and amount to 175 zettabytes [2]. This shows the trend of digital data exponential growth.

The users' activity recorded by various devices represents their digital traces. Today, digital footprints occupy a significant part of

the big data cloud. Their key direction is to extract information about the preferences of potential customers and offer products which they will be interested in, that is, increasing sales [3].

However, the information extracted from digital traces can be used not only in external interactions, but also in the internal activities of the company. Modern companies are increasingly using data to increase employee engagement in achieving and separating goals, that is, to increase the productivity of both employees and organizational processes, thereby creating new sources for their competitive advantage [4].

Thus, recommendation systems are an integral part of the major market players today. The recommendation system is a complex of algorithms, programs and services designed to form relevant recommendations to users regarding the object of information search [5].

Many studies note [6–8] that it is possible to overcome the key barriers of digital transformation in the form of a shortage of qualified personnel, competencies and knowledge, as well as internal resistance in organizations [9–10] by creating a system of training and staff development. This system is based on the formation of a personalized learning path [11] relying on the values of the organization and the employee and relevant educational material, in other words, on those objects, actions, tasks, which the employee will perform with increased interest and, as a result, with maximum efficiency.

In the education field, the use of information systems based on recommendations is designed to solve one of the student's main problems – choosing educational material. Often such material is training programs and educational courses. When selecting educational material in such systems, the student's personal characteristics are usually not taken into account, but his preferences are used [12]. Recently, researchers have begun to actively study using personal characteristics in educational recommendations. For example, the use of information about the learning style and emotional reactions [13]. Moreover, the issue of the recommendation systems used in online learning is increasingly rising in the research [14].

Thus, this study's purpose is to create an intelligent guide which would accompany an employee throughout his life in the organization, involving him in the learning process according to a personalized path based on data on the personal profile and reactions to educational material, training soft and hard skills in accordance with the values of the organization and the employee.

1. An intelligent system of personnel training and development concept

The principles of system engineering provided the basis for the intelligent guide development for the employees' growth and training. The requirements' collection of interested parties to the system was being developed as the breakaway point in the creation of an intelligent guide.

As the result of analyzing the requirements, it was revealed that the modular architecture of the system will not only take into account all the stakeholders' requirements, but also differentiate access to various functions depending on the positions held. The modular architecture proposed for implementation in accordance with the identified requirements is shown in *Fig. 1.*

A database was created using the MySQL database management system to implement the intelligent guide and support the work of the proposed modules.

The diagnostic module's main task is the formation of a digital profile of an employee, which will make it possible to take into account

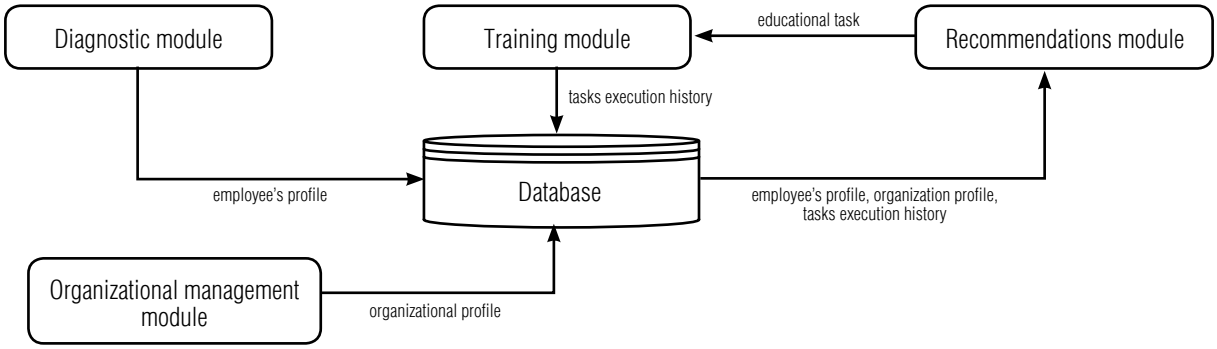

Fig. 1. Intelligent guide model.

all the characteristics of the individual and will become the basis for effective training and development.

Modern trends are such that organizations are increasingly focusing on supra-professional skills or so-called soft skills, which are much more difficult to reproduce with the help of digital transformation tools and which underlie the ongoing changes [15]. Soft skills are the basis for the effective development of professional skills or so-called hard skills.

Thus, it was proposed to form a digital profile of an employee, consisting not only of personal data (gender, age, education, position, work experience), but also revealing the levels of development of employee soft skills in demand in the labor market in the digital economy.

Comprehensive automated diagnostics were developed to determine the level of proficiency in one or another soft skill (*Fig. 2*). The DISC model, Rowe's "Decision-making Style" methodology, Honey and Mumford's method of determining activity styles, and a psychotype test were used as diagnostic material. These methods reveal the psychoemotional characteristics of a person in 16 projections with their respective competencies.

The organizational management module is designed to fix organizational values through formation of job profiles. First of all, an organizational structure is created through this module. After that, the head of the structural unit determines the points of the critical and desired profile of the employee for each position under consideration, in accordance with the labor functionality and labor actions, as well as the supra-professional skills designated on the basis of organizational values.

Thus, having compiled a position profile for each element of the organizational structure, a comprehensive profile of the organization is formed which reflects all the features of its activities.

The educational process organization is the basis of the training module. Users in the framework of soft skills development perform unique educational tasks (*Fig. 3*). For each task, the user evaluates how much he liked the task, its effectiveness and complexity. Thus, a data set is formed which characterizes the educational tasks' performance by users and is used in the future to form relevant recommendations.

The recommendations module is designed to form a personalized learning path in accordance with the values of the employee and the organization and to select the most relevant educational material for the employee.

It is possible to determine the soft skills directions' development of an employee within the interests of the organization if there is an

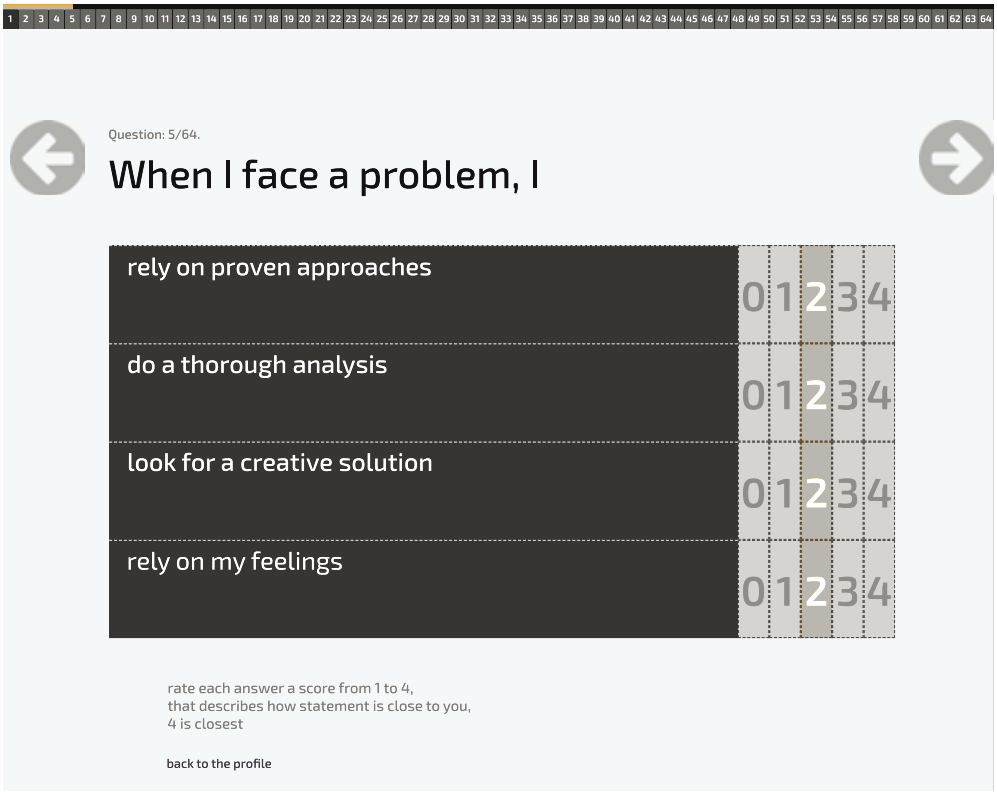

Fig. 2. Diagnostic module.

employee and position profile. However, as mentioned earlier, it is necessary to synchronize the values of the employee and the organization. So it is also necessary to determine the desired development directions for the employee. Within the framework, an employee development needs questionnaire depending on his value orientations has been developed.

After the diagnostics, employee is asked to answer a list of questions to form the desired employee profile and determine the competencies that he would like to develop

The overlapping of the obtained profiles on each other makes it possible to visualize the level of competencies' expression and to identify the directions of an employee's personal development in accordance with the needs of the organization, all of which leads to the formation of their common development path.

The next step is the direct selection of relevant educational material using a recommendation system. To select an algorithm for its operation, a study was conducted on technologies such as machine learning based on matrix factorization and a neural network.

2. Computational experiment

2.1. Problem Statement

There are many users $U = \{u_1, u_2, \ldots, u_n\}$ and many educational tasks $T = \{t_1, t_2, \ldots, t_m\}$.

The $V_{n \times m}$ matrix contains the scores assigned by users to educational tasks. There will be a number in place $v_{ij} (i \in 1, \ldots n, j \in 1 \ldots m)$ if the u_i user has evaluated the task t_j and is empty otherwise.

It is required to find a vector $\widehat{v_i}$ containing the user's already known estimates v_{ij}, as well as the

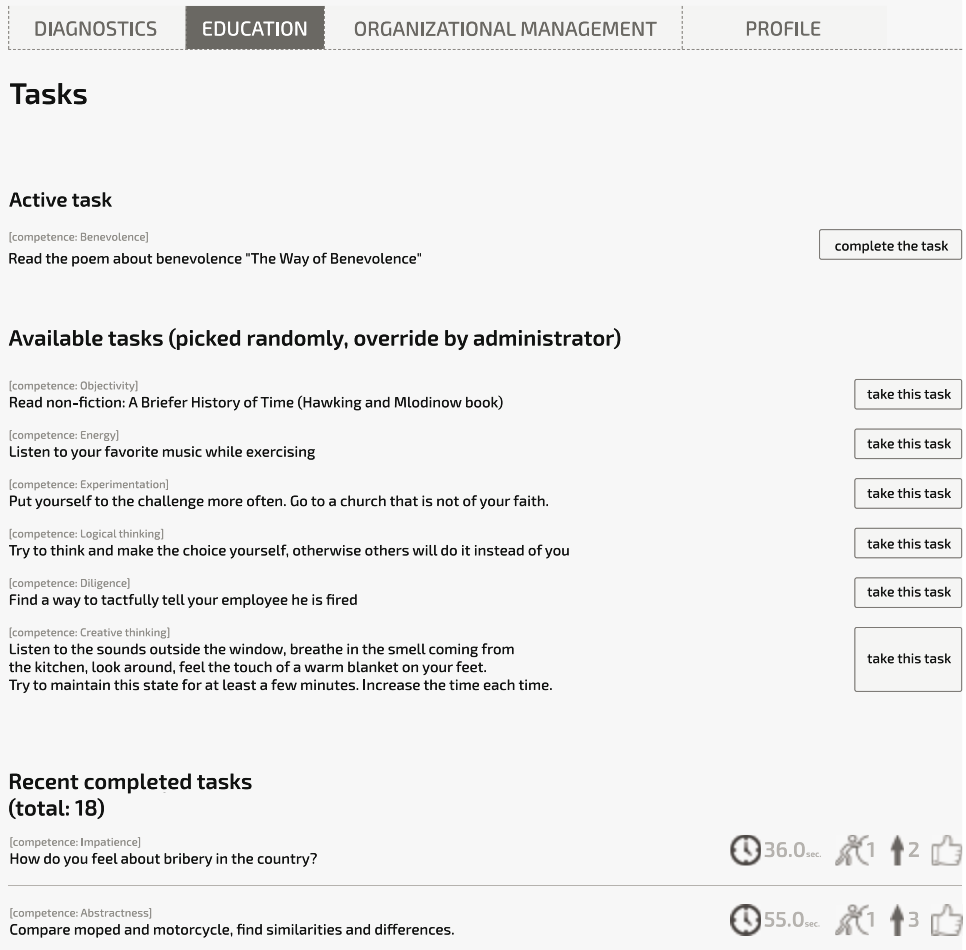

Fig.3. Training module.

expected ones $\widehat{v_{ij}}$. Based on the obtained vector, you need to rank the list of educational tasks T for the u_i user.

2.2. Machine learning based on matrix factorization

Matrix factorization implies decomposition of the original matrix into a product of two matrices of small rank. The user's interaction with the object is modeled as a scalar product of the vectors of user's imagine and the object in the factor space. Factorization models work effectively with highly sparse matrices [16–18].

Let us imagine the evaluation matrix V as a product of two matrices:

- $W_{n\times k}$ matrix, which contains both latent user characteristics and explicit ones (gender, education, work experience and user profile values consisting of 16 parameters);
- $H_{k\times m}$ matrix, which is characterized educational tasks.

Let us fill in the latent characteristics of the W and H matrices with random variables based on the law of uniform distribution on the interval $\left[0;\sqrt{\max\{v_{ij}\}/k}\right]$.

Next, we will solve the minimization problem:

$$\text{argmim}\,||V-\widehat{V}||+\alpha||H||+\beta||W||, \quad (1)$$

where $\widehat{V}$ is the matrix obtained by approximation from W and H;

α, β – algorithm parameters.

Each algorithm's iteration for finding a solution to minimize the error consists of the following steps:

1. Fix the H matrix.

2. Find the error $\delta = |v_{ij} - \widehat{v}_{ij}|, j \in 1, ..., n$.

3. Find new values $W_{ip} = W_{ip} - \epsilon (\delta H_{pj}^T + \lambda W_{ip})$, where $p \in 1, ..., k$;

λ – regularization parameter;

ϵ – learning rate.

4. Fix the W matrix.

5. Find the error $\delta = |v_{ij} - \widehat{v}_{ij}|, j \in 1, ..., m$.

6. Find new values $H_{pj} = H_{pj} - \epsilon(\delta W_{ip}^T + \lambda H_{pj})$, where $p \in 1, ..., k$;

λ – regularization parameter;

ϵ – learning rate.

When solving the problem experimentally, the rank of the matrices W and H $k = 40$ was found. This value is a boundary value for a given case in which the root-mean-square and absolute error of solving the factorization problem are optimal. With a value of $k > 40$, the advantage is not traced as a result, and the calculation time increases. Conversely, with a value of $k < 40$, too rough an approximation is obtained.

Based on the test results (*Table 1*), with an increase in the learning rate from 0.005 to 0.03, there is an improvement in the results (a decrease in error, an increase in prediction accuracy). At the value $\epsilon = 0.04$, the results are usually equivalent to the results at $\epsilon = 0.03$ or worse. With the number of iterations 20 ($\epsilon = 0.03$), high accuracy of factorization is already observed on the given samples. Based on these results, we use the following parameters for the algorithm: $\epsilon = 0.03$; the number of iterations is 20.

With an increase in the amount of data received and a decrease in the sparsity of data, it may be necessary to increase the number of iterations. The control of the number of iterations was automated by testing after the end of factorization and comparing the accuracy of the prediction with the previous testing. As the prediction accuracy decreases, the number of iterations increases. Otherwise the number of iterations remains the same. Another alternative is to set the permissible accuracy of factorization, at which the algorithm should be stopped.

Table 1.

Speed and factorization errors (n = 14000)

Iterations' number	ϵ	0.005	0.01	0.02	0.03	0.04
20	absolute average	0.197	0.119	0.064	0.05	0.048
	the quadratic mean	0.24	0.09	0.013	0.007	0.0064
	time, sec.	6	6	5	5	5
40	absolute average	0.113	0.057	0.028	0.02	0.022
	the quadratic mean	0.082	0.013	0.002	0.001	0.001
	time, sec.	11	10	10	11	11
60	absolute average	0.075	0.03	0.017	0.014	0.014
	the quadratic mean	0.03	0.003	0.0009	0.0006	0.0006
	time, sec.	16	15	16	15	15

One of the main tasks in the construction of recommendation systems is to solve the problem of 'cold' start [19]. It occurs when new elements appear in the system: whether it is a user or preference objects.

In the problem under consideration, two variants of cold start of users can be distinguished: the user did not perform the educational tasks; the user performed the educational tasks, but the factorization task was solved before his registration; that is to say, there are no user factors in the system.

Let us consider the first case. Nothing is known about the user's preferences. We know only about his profile. It is some vector which is formed from the values of gender, seniority, position. If the user has passed the test, then the vector will also have a soft skills profile of 16 components.

The problem of cold start in this case is proposed to be solved using cosine similarity. To do this, there is a cosine between the new user and each existing user. It is ranked in descending order, and the user to whom the maximum cosine corresponds is selected. As a result, recommendations are formed for a new user already based on the similarity of the user found through the cosine:

$$\cos\varphi = \frac{\left(\overline{W_a}, \overline{W_b}\right)}{\|\overline{W_a}\| \times \|\overline{W_b}\|}. \qquad (2)$$

In the second case, the problem is proposed to be decided by solving the factorization problem for a specific user. In this case, the W and V matrices will take the form of row vectors. The problem factors H matrix will be fixed, and the search for values will be performed only for the W vector-string. The search algorithm is reduced to a special case: $i = 1$, and the steps of the matrix factorization algorithm from the 4th to the 6th are excluded.

We also need to note the 'cold' start problem for educational tasks. At the current stage of the study, tasks for which there are no factors are randomly offered to users.

To assess the quality of predictions, the set of V estimates was divided into samples for V_{train} training and V_{test} testing. The model was tested with the following quantitative characteristics:

- Users' number: 303;
- Task's number: 11726;
- Task's number completed by users: 17733;
- Training sample: 75%;
- Test sample: 25%;
- Matrices' rank: 40;
- Factorization steps: 20;
- Range of original matrix' values: 0–10 (normalized data).

The V matrix is very sparse; most of the cells are not filled (more than 95% of the cells). When evaluating the performance of the model, three types of approximation were carried out: for efficiency, complexity and preferences. Four cases were considered:

Using only latent factors (marked as *FREE* on graphs);

Using fixed user factors along with latent (soft skills profile, gender, position, education; designated as *UV*);

Using fixed task factors along with latent ones (the average values of performance indicators, complexity and preferences set by users, the level of complexity, the orientation of tasks to competencies, are designated as *TV*);

Using fixed factors, users and tasks (designated as *UV+TV*).

Fifty factorizations were carried out on random data for each case. Then these data were averaged. The testing also took into account the cold start of users based on cosine similarity. The cold start of tasks was not taken into account in testing.

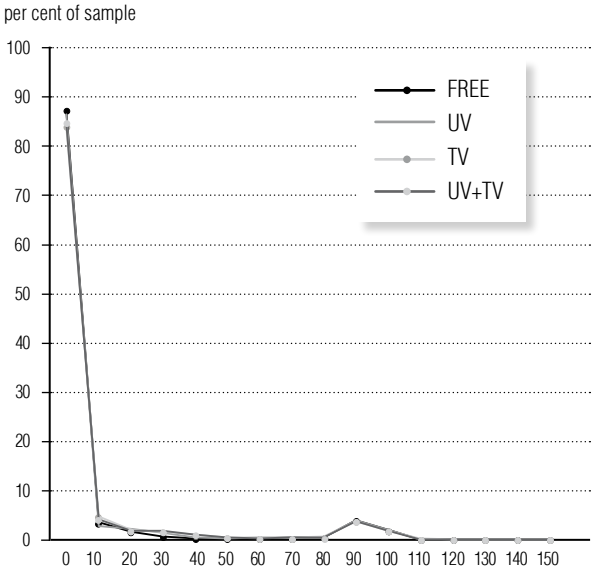

Fig. 4. Error in predicting preference using factorization on percentage intervals.

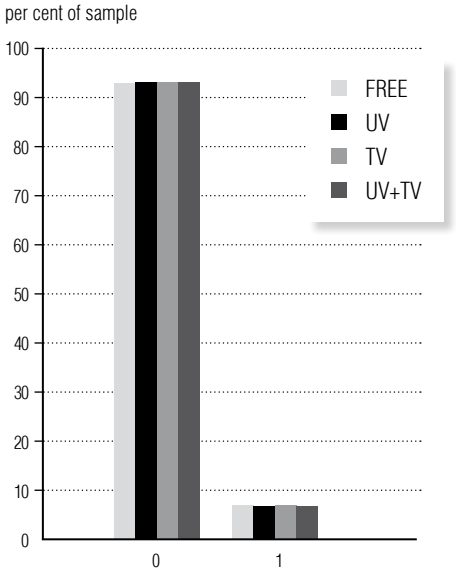

Fig. 5. Error in predicting preference using factorization on discrete values [0; 1].

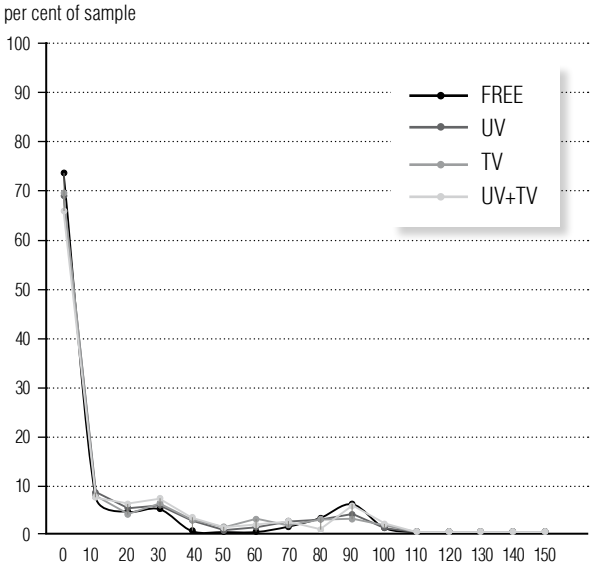

Fig. 6. Error in predicting preference using cosine similarity on percentage intervals.

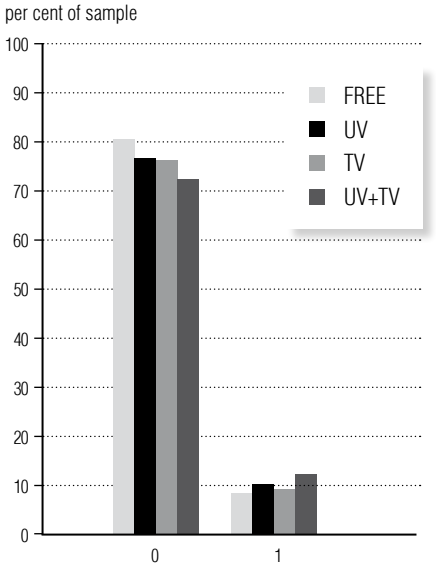

Fig. 7. Error in predicting preference using cosine similarity on discrete quantities [0; 1].

The data for the preference prediction error are presented in *Fig. 4–7*. The data for the efficiency prediction error are presented in *Fig. 8–11*.

The fixed factors' use mainly worsens the result, as can be seen from the graphs (*Fig. 4–11*). This may be due to the small size of the data set.

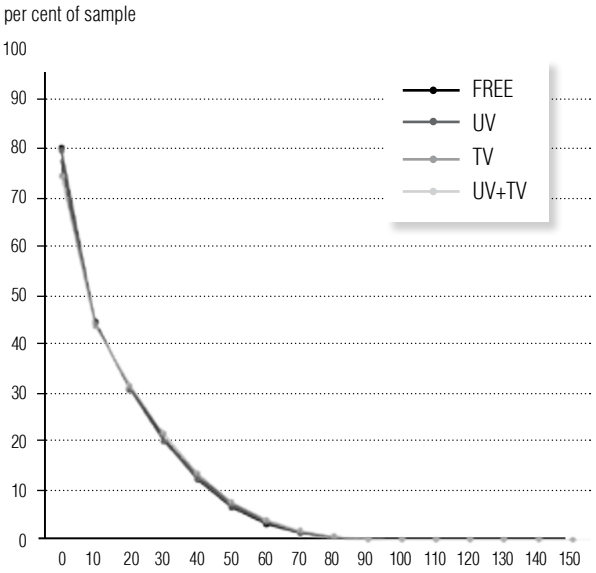

Fig. 8. Error in predicting efficiency using factorization on percentage intervals.

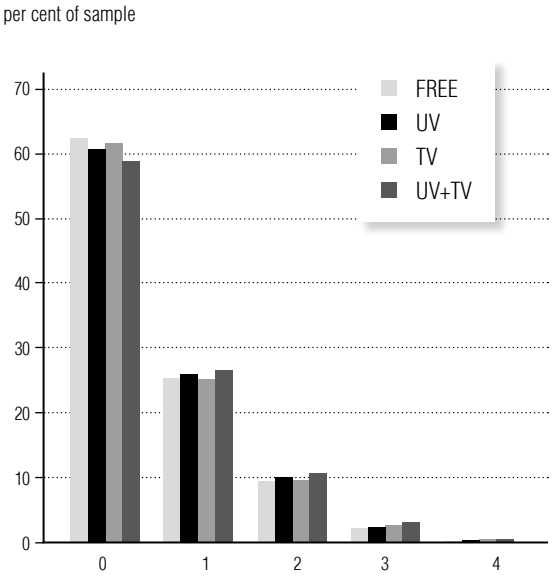

Fig. 9. Error in predicting efficiency using factorization on discrete quantities [1; 5].

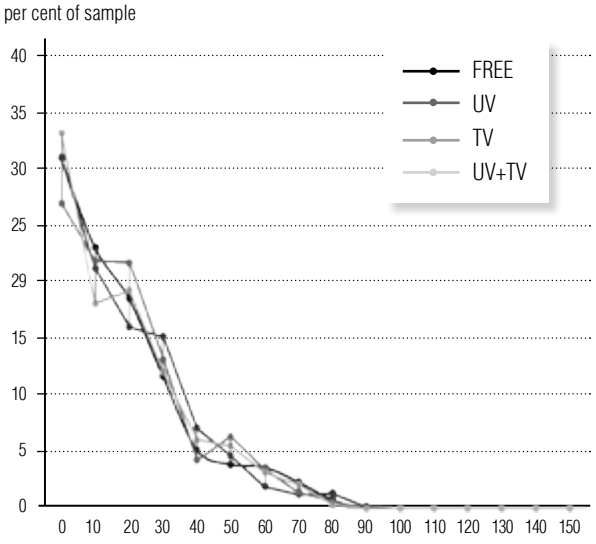

Fig. 10. Error in predicting efficiency using cosine similarity on percentage intervals.

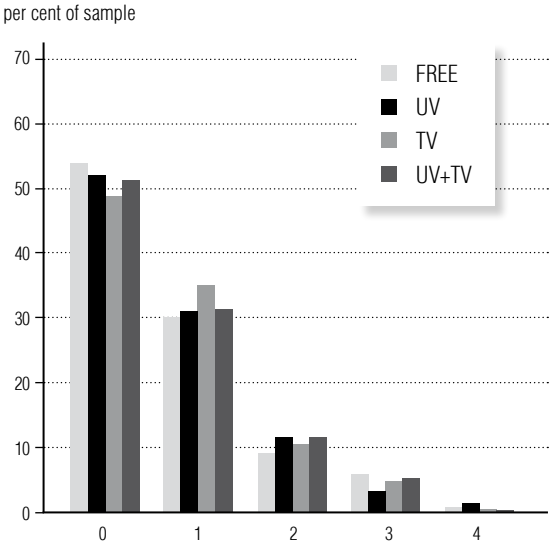

Fig. 11. Error in predicting efficiency using cosine similarity on percentage intervals and on discrete quantities [1; 5].

2.3. Neural network

A neural network is a computational structure assembled into computational elements' network created in the likeness of a biological neuron [20]. Neural networks are capable of solving a wide range of tasks in various fields of activity [21].

To solve this problem, such neural network architectures as a convolutional neural network [22] and a multilayer perceptron accord-

ing to Rumelhart [23] were considered. The latter proved to be the most effective. In addition, a convolutional neural network requires more data and training time.

Two methods were investigated: encoding words with a unique number and embedding words (Word2Vec) to vectorize the educational task [24]. Encoding words with a unique number is faster, but it does not allow you to record any relationships between words, as well as the similarity of words.

Using the Word2Vec model for word embedding allows you to capture subtle relationships between words, but requires a large amount of data for training and is a lengthy process. Due to the absence of the need for constant vectorization of educational tasks, but only when replenishing the database of educational tasks, the Word2Vec model was used.

The input layer of the neural network has one neuron, which receives a vector containing data about the user and tasks.

The activation function 'ReLU' is activated on the hidden layers, the activation function 'Sigmoid' is activated on the output layer, which allows you to amplify weak signals and is not saturated by strong signals [25].

During training and testing the neural network, a data set consisting of 16 862 vectors containing data about users and tasks was used. The data were divided into training and test samples in the proportion of 80/20.

On the studied data set, the neural network with the following topology showed the greatest accuracy: two hidden layers (the first layer – 64 neurons, the second layer – 128 neurons), the size of the batch – 32, the number of epochs – 20.

The neural network results and the prediction accuracy are presented in *Table 2*.

To construct a vector, the network is launched as many times as the values are contained in the database of educational tasks. Thus, after all iterations, a vector $\widehat{v_i}$ is formed on the basis of which it is possible to select the most preferred tasks for the user.

3. A comparison of Artificial Intelligence technologies

Two approaches to the formation of recommendations were considered: neural network and machine learning based on matrix factorization. Both approaches have both advantages and disadvantages.

As opposed to vectors fed to the neural network input, the factor matrix consists of unique user vectors that contain latent features that are not taken into account by the neural network. The neural network, in turn, relies only on a given known user vector consisting of parameters laid down by the researcher. It allows us to talk about a more personalized approach when using matrix factorization.

The neural network also requires the mandatory availability of not only all data about the user, but also their correctness. When using machine learning based on matrix factoriza-

Table 2.

Neural network results

	Preference	Complexity	Effectiveness
Standard deviation	0.056	0.102	0.098
Accuracy, %	93.31	68.79	61.0

tion, you can make recommendations after several evaluations without having any more information.

The matrix factorization's disadvantage is a cold start, when the user has not yet evaluated any tasks, and he needs to give recommendations. However, this problem is solved very quickly using the method described earlier.

Despite the fact that the researchers were faced with the task of ranking, it was necessary to compare the work of these methods according to three criteria: preferences (binary value), complexity (discrete [1; 4]) and efficiency (discrete [1; 5]). The comparison results are presented in *Table 3*.

As can be seen from *Table 3*, on discrete values, the accuracy of neural network predictions decreases compared to matrix factorization. Based on the above, the researchers decided to use machine learning based on matrix factorization as the core of the recommendation module.

Conclusion

The problem of low adaptability of management in the conditions of digital transformation was considered in this article. Within the research framework, a model of an employee's intelligent guide has been developed. This model allows us to level this problem due to the qualitative development of the employee's competence profile, taking into account the value orientations of the employee and the organization. The model's implementation has been carried out in the form of a software component for the enterprise management system.

The new competencies' development will lead to the expansion of the employee's profile and ability to perform new tasks in the area of interest, thereby obtaining a high-quality result. The organization, in turn, will receive new opportunities for development in the market and rapid adaptation to changing conditions. Thus, by investing in employee development, the organization receives T-shaped specialists with a proactive position and capable of self-organization. This directly leads to the realization of the potentials of both the employee and the organization.

The intelligent guide has a modular architecture, which is due to the interdisciplinarity and novelty of the research, as well as the simplicity of implementing the received developments into real management practice. This approach made it possible to conduct qualitative research in two projections: the development of comprehensive automated soft skills diagnostics of an employee's profile and relevant filling of an individual educational path with educational material using artificial intelligence, and then to combine the results into a single system for training and staff development.

In the proposed approach artificial intelligence is expressed by machine learning using matrix factorization. Such intelligence allows you to qualitatively select educational material which will be interesting to the employee from his personal positions and maximally involve

Table 3.

Accuracy of artificial intelligence technologies' predictions

	Preference	Complexity	Effectiveness
Neural network	93.3%	68.8	61%
Matrix factorization	93.1%	76.4	62.4%

him not only in the learning and development process, but also in achieving organizational goals.

The proposed database structure allows you to collect digital traces of users describing biographical characteristics, skills and level of their proficiency at a certain stage of development, the history of choosing and completing educational tasks, and further use of the collected data as the basis for the work of the recommendations module and the formation of personalized proposals.

The interdisciplinarity and novelty of the research also determine the variability of its development. As part of further research, it is planned to integrate the software component into the enterprise's business process management system. This is supposed to predict the time when the labor resource is not loaded and offer educational tasks to fill it. It will be the basis of a fundamentally new approach to considering downtime not as a loss, but as an opportunity for the development and training of employees.

The developed software component can be used by enterprises not only at the organizational level, but also through broadcasting in the education system. On the one hand, enterprises will be able to reduce staff downtime, expand the personal competence profile of employees, and increase the growth of the production potential of both employees and enterprises in the market. On the other hand, educational institutions will be able to train highly qualified personnel within the framework of orientation to the values of the market and the values of students with the formation of individual educational trajectories. Such cooperation will make it possible to form an education ecosystem in accordance with the requirements of innovative development of a region's economy. ■

References

1. Deloitte Insights (2020) *Results of the study "International trends in human resource management – 2020" in Russia*. Available at: https://www2.deloitte.com/content/dam/Deloitte/ru/Documents/human-capital/russian/hc-trends-2020-Russia.pdf (accessed 22 November 2021) (in Russian).
2. TAdviser (2020) *Big Data global market*. Available at: https://tadviser.com/a/e.php?id=129607 (accessed 22 November 2021).
3. Fatyanov A.A. (2018) Big Data in the digital economy: its value and legal challenges. *Economics. Law. Society*, no. 4, pp. 37–40 (in Russian).
4. Reinsel D., Gantz J., Rydning J. (2018) *The digitization of the world – from edge to core.* IDC White Paper. Available at: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (accessed 22 November 2021).
5. Ali N.M., Gadallah A.M., Hefny H.A., Novikov B.A. (2021) Online web navigation assistant. *Vestnik Udmurtskogo Universiteta. Matematika. Mekhanika. Komp'uternye Nauki*, vol. 31, no. 1, pp. 116–131. https://doi.org/10.35634/vm210109
6. Gokhberg L. (ed.) (2019) *What is the digital economy? Trends, competencies, measurement*. Moscow: HSE (in Russian).
7. Kulikova M.Kh., Kulikova M.Kh., Magomadov M.A. (2020) IT in education: today, tomorrow and always. Proceedings of the *I Student Scientific and Practical Conference "Information technologies in business and education", Grozniy, 20 February 2020*, pp. 85–90 (in Russian). https://doi.org/10.36684/21-2020-1-32-35
8. Geissbauer R., Lübben E., Schrauf S., Pillsbury S. (2018) *Digital Champions. How industry leaders build integrated operations ecosystems to deliver end-to-end customer solutions*. PwC Strategy& Global Digital Operations Study. Available at: https://www.strategyand.pwc.com/gx/en/insights/industry4-0/global-digital-operations-study-digital-champions.pdf (accessed 22 November 2021).

9. Harvard Business Review (2017) *High-performance sourcing and procurement driving value through collaboration*. Harvard Business School Publishing. Available at: https://hbr.org/resources/pdfs/comm/scoutrfp/HighPerformanceSourcing.pdf (accessed 22 November 2021).

10. Dolganova O.I., Deeva E.A. (2019) Company readiness for digital transformations: problems and diagnosis. *Business Informatics*, vol. 13, no 2, pp. 59–72. http://doi.org/10.17323/1998-0663.2019.2.59.72

11. Komarov V.A., Sarafanov A.V. (2021) IoT systems in the process of multidisciplinary training of personnel for the digital economy and their design. *Business Informatics,* vol. 15, no 2, pp. 47–59. http://doi.org/10.17323/2587-814X.2021.2.47.59

12. Rivera A.C., Tapia-Leon M., Lujan-Mora S. (2018) Recommendation systems in education: A systematic mapping study. Proceedings of the *International Conference on Information Technology & Systems (ICITS 2018), Libertad City, Ecuador, 10–12 January 2018* (eds. A. Rocha, T. Guarda). Advances in Intelligent Systems and Computing, vol. 721, pp. 937–947. https://doi.org/10.1007/978-3-319-73450-7_89

13. Bustos López M., Alor-Hernández G., Sánchez-Cervantes J., Paredes-Valverde M., Salas-Zárate M.P. (2020) EduRecomSys: An educational resource recommender system based on collaborative filtering and emotion detection. *Interacting with Computers*, vol. 32, no. 4, pp. 407–432. https://doi.org/10.1093/iwc/iwab001

14. Urdaneta-Ponte M.C., Mendez-Zorrilla A., Oleagordia-Ruiz I. (2021) Recommendation systems for education: Systematic review. *Electronics*, vol. 10, no. 14, article ID 1611. https://doi.org/10.3390/electronics10141611

15. Bughin J., Hazan E., Lund S., Dahlström P., Wiesinger A., Subramaniam A. (2018) *Skill shift: automation and the future of the workforce.* McKinsey & Company. Available at: https://www.mckinsey.com/featured-insights/future-of-work/skill-shift-automation-and-the-future-of-the-workforce (accessed 22 November 2021).

16. Rubtsov V.N. (2017) *Matrix factorization based on deep learning for collaborative filtering*. Student Theses. Moscow: HSE. Available at: https://www.hse.ru/en/edu/vkr/206744221 (accessed 22 November 2021) (in Russian).

17. Strömqvist Z. (2018) *Matrix factorization in recommender systems: How sensitive are matrix factorization models to sparsity*. Uppsala University Publications. Available at: https://uu.diva-portal.org/smash/get/diva2:1214390/FULLTEXT01.pdf (accessed 22 November 2021).

18. Mojsyuk-Dran'ko P.A., Revotyuk M.P. (2020) Matrix factorization methods for recommendation systems. Proceedings of the *International Conference Information Technologies and Systems 2020 (ITS 2020)*. Minsk: Belarusian State University of Informatics and Radioelectronics, pp. 193–194. Available at: https://libeldoc.bsuir.by/bitstream/123456789/41339/1/Moysyuk_Dranko_Metody.pdf (accessed 22 November 2021) (in Russian).

19. Kuznetsov I.A. (2019) *Machine learning methods and algorithms for preprocessing and classification of semi-structured text data in scientific recommendation systems*. Moscow: NRNU MEPhI. Available at: https://ds.mephi.ru/documents/90/Кузнецов_И_А_Текст_диссертации.pdf (accessed 22 November 2021) (in Russian).

20. Golovko V.A., Krasnoproshin V.V. (2017) *Neural network data processing technologies*. Minsk: Belarusian State University. Available at: https://elib.bsu.by/bitstream/123456789/193558/1/Golovko.pdf (accessed 22 November 2021) (in Russian).

21. Lisovsky A.L. (2020) Application of neural network technologies for management development of systems. *Strategic decisions and risk management*, vol. 11, no. 4, pp. 378–389 (in Russian). https://doi.org/10.17747/2618-947X-923

22. Bezdan T., Bacanin Džakula N. (2019) Convolutional neural network layers and architectures. Proceedings of the *Sinteza 2019: International Scientific Conference on Information Technology and Data Related Research, Belgrade, 20 April 2019* (ed. Milovan Stanišić), pp. 445–451. https://doi.org/10.15308/Sinteza-2019-445-451

23. Postarnak, D.V. (2012) The critical analysis of models of neural networks. *Vestnik Tyumenskogo gosudarstvennogo universiteta*, no. 4, pp. 162–167. Available at: https://elibrary.ru/item.asp?id=17758787& (accessed 22 November 2021) (in Russian).

24. Young T., Hazarika D., Poria S., Cambria E. (2018) Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75. https://doi.org/10.1109/MCI.2018.2840738

25. Iqbal T., Qureshi S. (2020) The survey: Text generation models in deep learning. *Journal of King Saud University – Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2020.04.001

About the authors

Ekaterina S. Morozevich

Process architect, Research Management Department, Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk 660037, Russia;

E-mail: katyamorozevich@mail.ru

ORCID: 0000-0001-8796-3157

Vladimir S. Korotkikh

System administrator, Municipal budget general education institution secondary school No. 4, 9, st. Naberezhnaya, Divnogorsk 663091, Russia;

E-mail: vskor@bk.ru

ORCID: 0000-0001-9270-3376

Yevgeniya A. Kuznetsova

Engineer, Research Management Department, Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk 660037, Russia;

E-mail: zhenya.kuz-1997@yandex.ru

ORCID: 0000-0002-3559-8394

DOI: 10.17323/2587-814X.2022.2.36.48

Technologies of collective intelligence in the management of business processes of an organization

Boris B. Slavin
E-mail: bbslavin@gmail.com

Financial University under the Government of the Russian Federation
Address: 49, Leningradsky Prospekt, Moscow 125993, Russia

Abstract

With the digitalization of the economy, the creative component of an organization's activities increases. Standard business process management methods stop working due to the rise in uncertainty of the task solution time. Currently, there are no effective technologies for managing intellectual activity processes in organizations. The role of collective intelligence technologies for knowledge management in organizations has long been discussed in the literature, but there are still no concrete proposals on implementation. This work aims to show how collective technologies can solve the problems of managing business processes of intellectual activity. The possibility of collective intelligence technologies for increasing labor productivity is demonstrated. Models for distributing tasks by competencies and synergy from collaboration are proposed for this demonstration. The paper shows that competencies are the primary metric that can be used to measure work with knowledge in an organization. But they should also be considered when organizing group activities. A simple model example shows that the correct distribution of tasks by competencies allows you to increase the speed of solving tasks by a group by several times. In real cases, calculations using computing resources are necessary. A model is also proposed that demonstrates increasing the joint activity of a creative employee and an analyst. It is shown that business process management should be supplemented by mapping the competence model and group work options to the stages of business processes. This will allow you to manage the business processes of intellectual activity.

Keywords: collective intelligence, competencies, knowledge management systems, business processes, intellectual activity, synergy, brainstorming, group work

Citation: Slavin B.B. (2022) Technologies of collective intelligence in the management of business processes of an organization. *Business Informatics*, vol. 16, no. 2, pp. 36–48.
DOI: 10.17323/2587-814X.2022.2.36.48

Introduction

Organizations that have passed or are undergoing a stage of digital transformation are starting to compete in the innovation market. To do this, they need to create new products and services on a much larger scale than before. In the pre-digital era, the departments involved in the introduction of innovations were few, but they coped with their work. Today, many technologists, developers and managers are involved in the creation of innovative products. That is why DevOps and BizOps technologies are becoming popular, involving the creation of a continuous pipeline from the development of new products (Dev – Development) to their commissioning (Ops – Operations) and back, or even starting with ideas generated by business (Biz – Business). In this regard, the share of employees of such organizations engaged in creative intellectual activity increases by a multiple. It is no coincidence that industries at the center of the "digital vortex" are the main consumer of creative personnel today, and they feel "hunger" in them.

However, not only problems with the labor market arise during the transition to a knowledge society. Increasing the share of creative activity requires a radical revision of approaches to business process management. Let us show this with a simple example. *Figure 1* shows a simple business process consisting of four stages. For each stage of the business process, graphs of the probability of its execution in time are given, where the value "1" means execution. The condition that the entire business process will be manageable is that each stage must be completed on time. Usually, even a little more time is laid on the execution so that it is highly likely to be executed. Moreover, the execution time is standardized, and according to such normatives, it is possible to accurately predict when the business process will be executed. For example, in car services, both the labor intensity and the time of order execution are calculated in this way. The entire business process management system in the organization is built on the fulfilment of this rather obvious condition.

However, if the business process concerns creative activity, the end time of the stages will be unpredictable. The stage may end much earlier than the allotted time, or it may end much later. *Figure* 2 shows the probability densities of completing the stages in the case of creative activity.

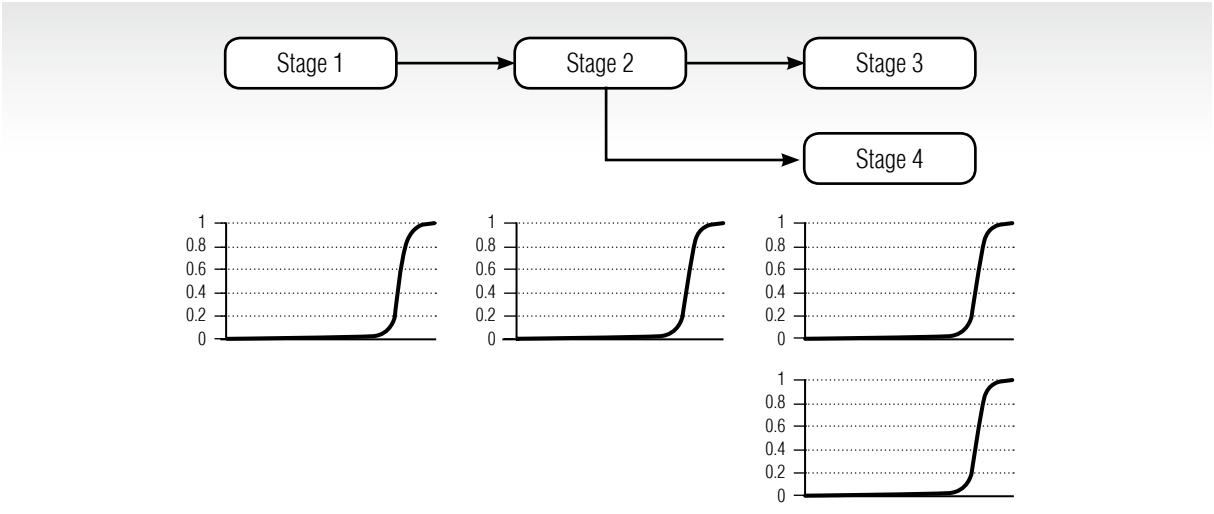

Fig. 1. Diagram of the business process and the probability of completion of stages on time in normal activities.

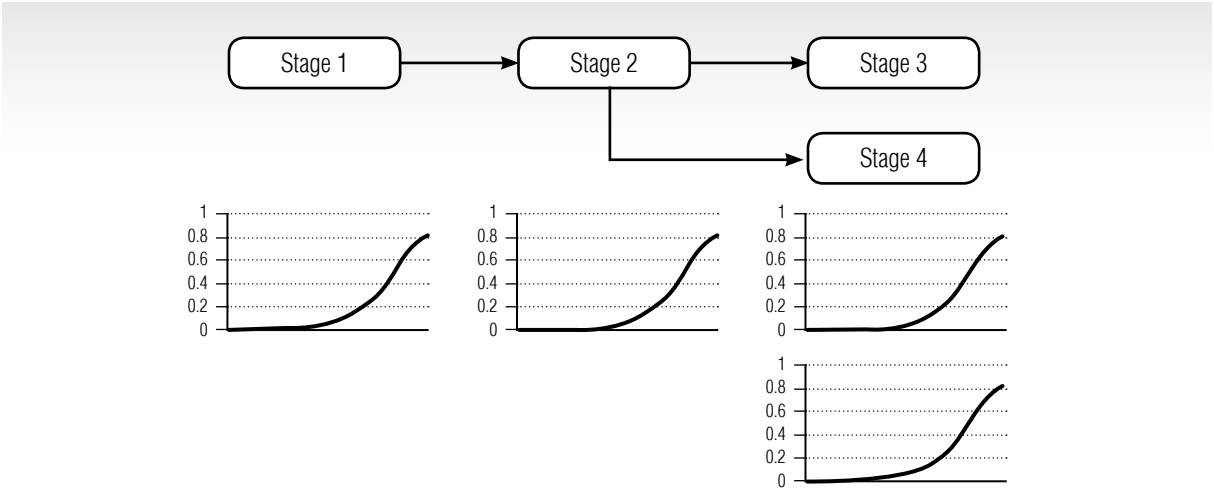

Fig. 2. The scheme of the business process and the probability of execution of stages in creative activity.

In fact, such a business process will be unmanageable, since the probabilities that the stages will not be completed on time will multiply, and the end time of the entire process will become unpredictable. You can, of course, significantly increase the time for each of the stages, but then the efficiency of the business process will be excessively low and employees will be idle most of the time. When innovation departments were small, they were simply taken out of the framework of the business process management system and set goals with indefinite deadlines. In the case when innovation units are integrated into common business processes, as is the case today with companies caught in the center of digital transformation, business process management systems will not work. Collective intelligence technologies and a competency-based approach to process management should help solve this problem.

1. From knowledge management to competence management

In [1] it is shown that the analogue of a commodity in the knowledge economy will not be knowledge itself, but the ability of people to operate with knowledge, i.e., their competence [2]. Unlike knowledge, the cost of competencies is proportional to the cost price, i.e., the costs necessary for human education. If the part of the costs of teachers can be attributed to several students (which means that such costs can be reduced), then the time resources of the student himself are not replicated, and in the future, they will probably determine a significant share of the cost of training. Moreover, investments in competencies are quite market-based and give the same return as investments in the production of goods and services. For example, according to economists, one year of study increases the salary by 10% on average. This, in turn, means that total investment in human competencies will be more profitable as we move to the knowledge economy.

Due to the special role of competencies in the innovation economy, studies of the possibility of building competence management systems, organizing creative activities, measuring and increasing the value of human capital become relevant [3]. Currently, only personnel accounting information systems have become widespread, but the market is already growing demand for such subject-oriented information systems as Talent management, Career Devel-

opment Planning, Competence Management, etc. Competence management as a concept first appeared in the field of education 40 years ago [4], but it became widespread in business later [5]. Darnton [6] formulated the main components of the competence management process, to which he attributed: the relationship of employee competencies with the strategy and objectives of the enterprise, the conditions for the development of competencies, their classification, development planning and competence control. Competence management is part of the general knowledge management system (Knowledge Management System, KMS) [7], if by knowledge we mean explicit and implicit knowledge. For the first time to highlight "tacit" or implicit knowledge (tacit knowledge) back in 1958, Polani [8, p. 23] proposed, referring to them the knowledge that a person has beyond what he can say. In relation to knowledge in an organization, Nonaka and Takeuchi [9] used the term "implicit knowledge" in their book "The company that creates knowledge", calling it more often "unformalized" knowledge: "Unformalized knowledge (or "implicit") is personal and dependent on the situation and therefore difficult to formalize and disseminate" [9, p. 84]. Competencies include implicit knowledge, and, unlike knowledge, are measurable. Since it is possible to talk about managing something only if the object of management is measurable, the concept of competence management makes sense in contrast to knowledge management. However, you can still use the term of knowledge management, meaning by it is the management of people's competencies for the creation and use of knowledge.

2. Technologies of collective intelligence

The competence approach plays a special role in the technologies of collective intelligence. The concept of collective intelligence has a broad interpretation, and in one form or another (wisdom of the crowd, collective intelligence, etc.) it can be found in scientific literature dating back many hundreds of years ago [10]. The term Collective Intelligence itself was apparently first introduced by David Wexler, the creator of the so-called Wexler intelligence assessment scales. Wexler [11, p. 906] argued that collective intelligence arises only when group members use common intellectual resources in their activities. Many works were devoted to the possibilities of collective creative activity at the end of the last century. Let us note as an example the book by the Fischer spouses [12] "Distributed minds: Achieving high productivity through the collective intelligence of working groups", which discusses approaches to the collectivization of knowledge in organizations. Nevertheless, the problems of collective intelligence received the most attention only with the development of the Internet [13]. It was the era of the Internet that was marked by a strong interest in the problems of collective intelligence.

At the end of the last century, Canadian publicist Levy [14] published a book entitled "Collective intelligence: Mankind's emerging world in cyberspace", in which he called for creating a society where cyber technologies have a humanizing influence and contribute to the emergence of "collective intelligence." Heylighen (author of the book "The global superorganism: An evolutionary-cybernetic model of the emerging network society" [15]) wrote that it is very important to learn how to use network communications to increase "collective intelligence" in such a way that group intelligence exceeds the sum of the intelligences of group members [16, p. 92]. Researchers of collective intelligence pay special attention to the Wikipedia project. For example, American scientists from Carnegie Mellon University have identified the relationship between the complexity of Wikipedia content and the competence of the editors of this project [17]. Horost [18, p. 251], who generally views all network resources as a global brain with memory, nodes and synapses, wrote about Wikipe-

dia as a collective knowledge base: "Wikipedia is distinguished by its "intelligence," which it develops through collective consciousness and content editing. And again, we see the total sum of many individual judgments about what is important and what is not... The resulting knowledge differs from PageRank, but both resources complement each other perfectly. In combination, they form, as it were, the incipient frontal lobes, the hippocampus and a kind of long-term memory Network".

Zettsu and Kiyoki [19] wrote the fact that collective intelligence technologies are one of the tools for knowledge management on the Internet. In [20], it was generally proposed to consider all social networks as a knowledge infrastructure (knoware) of collective intelligence. The authors introduce such concepts as a "supernet of knowledge," which includes media networks, user networks and knowledge networks. A group led by Malone [21] conducts many studies on the topic of collective intelligence at the Massachusetts Institute of Technology in the USA. Scientists of this group are studying various ways of applying collective intelligence technologies, both for organizing global network projects and for improving business efficiency. Such works as [22, 23] are devoted to the use of collective intelligence technologies as special information systems of enterprises. In [24], collective intelligence technologies were considered as technologies for improving the efficiency of human activity by analogy with the use of business intelligence tools.

In [25] Malone and his colleagues proposed a classification, which they called the "genome" of collective intelligence. However, in fact, this classification did not reveal the features of collective intelligence technologies, but simply allowed ranking all global network projects. Many researchers, following Malone, also do not distinguish between crowdsourcing technologies and technologies of collective intelligence [26, 27]. However, there is another point of view. So Gruber [28, p. 4], describing crowdsourcing technologies and social networks, writes that they can only claim to be called a "collection of intelligences, but they are not a single collective intelligence, since they do not support group thinking.

There is currently no consensus on what collective intelligence technologies should include. This study supports and develops the point of view that collective intelligence technologies are tools and systems "that unite into groups the necessary number of people who have their own individual goals but organized in such a way that the overall intelligence and effectiveness of the group increases" [29, p. 219]. Within the framework of this approach, it is possible to define collective intelligence technologies as a special form of "information technologies that contribute to the collective solution of intellectual and creative tasks using network communications" [10].

3. The role of competencies in collective intelligence technologies

Let us look at some examples that show the effectiveness of collective intelligence technologies in organizing creative activities. The first thing that such technologies allow – due to the correct consideration of competencies in the distribution of tasks that the group solves – is to significantly speed up their solution. Let us assume that we have a group of four employees with different competencies (let there be 6 of them), which indicate the probability of solving a problem for this competence (this is how the human intelligence index, IQ, is usually measured). For simplicity, let these probabilities be either 0 (there is no competence) or 1 (competence allows you to solve the problem with probability 1). Then the range of competencies of such a group can be described by a rectangular matrix shown in *Fig. 3a*. The first employee uniquely solves the tasks of the first two and fourth competencies, the second – from the third to the fifth, etc.

a)

1	0	1	0
1	0	1	1
0	1	0	0
1	1	0	0
0	1	1	0
0	0	1	1

b)

1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
3	**3**	**4**	**2**

Fig. 3. Competence matrix (a), group distribution matrix (b).

Let this group also solve problems collected in four sets of six competencies each. And let us assume that these tasks are distributed evenly among the participants (*Fig. 3b*). With such assumptions, the tasks will be solved only if the solvers have similar competencies – the results of the work are shown in the row under the matrix in *Fig. 3b*. It can be seen that on average the participants of such a group will solve three tasks.

However, if we distribute the tasks among the participants in a different way (see the group distribution matrix in *Fig. 4b*), keeping the load on each participant – six tasks, we can ensure that each of the participants solves all six tasks. That is to say, the productivity of the group will be twice as high and simply due to the correct selection of competencies. It is the finding of a group (or collaborative) distribution matrix that is necessary when organizing work within the framework of collective intelligence technology.

In the case when a larger number of employees participate in the group, and the probability of their solving problems differs from 0 or 1, numerical calculations must be carried out, while the difference in productivity may be even higher. The algorithm for organizing group work based on the calculation of the collaborative matrix is an analogue of the division of labor, but for intellectual activity. It is clear that it is difficult to accurately measure the probability of solving certain tasks, but it is possible to assess the speed of their solution by one or another specialist. The correct division of the overall task into subtasks and the correct selection of personnel facilitates the effective use of intellectual resources. In the current practice of conducting complex scientific research, managers are still carrying out such a distribution, relying only on intuition.

a)

1	0	1	0
1	0	1	1
0	1	0	0
1	1	0	0
0	1	1	0
0	0	1	1

b)

3	0	1	0
0	0	1	3
0	4	0	0
3	1	0	0
0	1	3	0
0	0	1	3
6	**6**	**6**	**6**

Fig. 4. Competence matrix (a), heterogeneous group distribution matrix (b).

4. Synergy in collective intelligence

The division of people by technical competencies is not the only condition for the effectiveness of collective intellectual activity. It is important to properly organize joint work on one task, or synergy, which considers creative and analytical competencies. The division of experts into analysts and "idea generators" is an important component of the brainstorming method. Altshuller [30, p. 10] in his book "The invention algorithm" describes this method proposed by the American journalist Alex Osborne at the end of the 1930s: "There are people who are good at "generating" ideas, but do not cope well with their analysis. And vice versa: some people are more inclined to critically analyse ideas than to "generate" them. Osborne decided to separate these processes. Let one group, having received a task, only put forward ideas, even the most fantastic ones. Let the other group only analyze the ideas put forward". Even though the "generation" of ideas and their analysis can be considered different competencies, it is especially necessary to take them into account when organizing intellectual activity, since one task cannot be divided into the phase of developing ideas and the phase of their concretization; joint work is necessary.

To understand how the synergy effect is achieved when a creative participant ("generator" of ideas) and an analyst interact, we can use model probability density functions for solving a problem. If we assume that the time for solving the problem is the same for both specialists (and is equal to 10), the probabilities of their solving the problem will look something like as shown in *Fig. 5*, where F_i is the probability of solving the problem by the "generator" of ideas, and F_a is the analyst. An expert analyst is unlikely to solve the problem ahead of time $t = 6$, and almost certainly will solve it by time $t = 14$, while an expert with creative competencies will solve the problem only by time $t = 20$, but it is likely that he can solve the problem even with small values of t.

The distribution function can be interpreted not only as the probability density of solving the problem, but also as the percentage of task completion. Of course, a separate task can either be completely solved, or it will not be solved. But in some cases, a partial solution of the problem makes real sense – for example, when performing some research, when one scientist can conduct only part of the research and another can

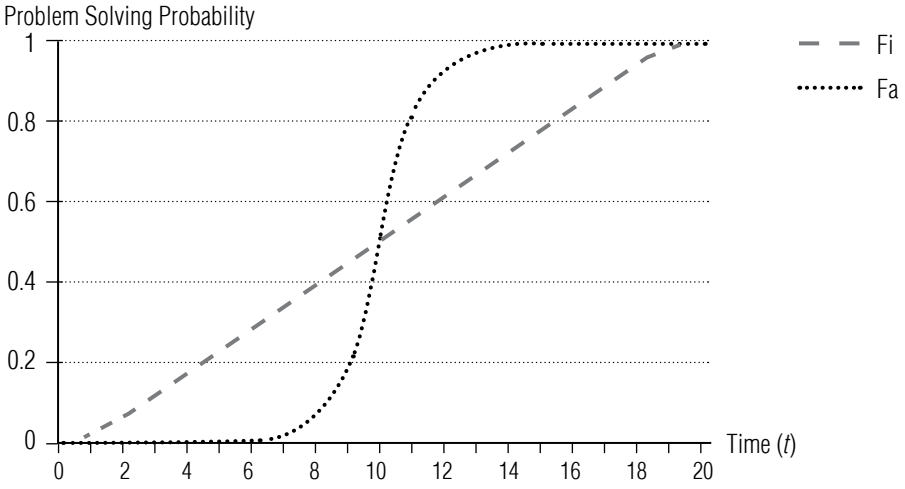

Fig. 5. Probabilities of solving the problem by the "generator" of ideas and the analyst.

finish it. This interpretation of the distribution function allows us to simulate a situation when two specialists are working on a task at once, and one is an analyst, and the second is a "generator" of ideas. When transferring a task to each other, its volume (or probability) should not change. Mathematically, this means that the probability function of the joint solution of the problem (in the case of transferring the problem from one to another) must be continuous.

The continuity of the probability function of the joint solution of the problem is quite obvious, but only this property does not allow us to determine the moment when it is possible to transfer the task to another participant. It is possible to formulate a hypothesis that when transferring a task from one participant to another, equality is necessary not only for the volume of the solved problem, but also for the dynamics of its solution. There is no evidence of this hypothesis yet, but there are empirical facts that partially confirm its validity. So, in [31], students' collaboration was studied in work which was carried out remotely using network tools (blogs, wiki, etc.), and it was shown that students participate in collaboration with great success when the style of problem solving (skills, knowledge, goals and plans) of their partners is closer and clearer to them. This hypothesis means that the collaborative probability function of a joint solution must be not only continuous, but also smooth (continuous in the first derivative or continuous for the probability density function).

The "generator" of ideas considers possible solutions to the problem faster than the analyst, since he does not check them immediately. At a certain point in time (let us denote it τ_i), the volume of the problem solved by him, and the speed of the solution may turn out to be equal to how an analyst would solve it, but much later, at the moment of time τ_a. If at this moment you transfer the task from the "generator" of ideas to the analyst, the overall solution of the problem will be reduced by an amount $(\tau_a - \tau_i)$. In a sense, this transfer of the solution from the "generator" of the idea to the analyst models "insight" in the group solution of the problem. Thanks to this "insight," the probability distribution shifts along the time axis to the left – depicted by the "Collab" line in *Fig. 6*. With the selected distribution parameters, the time value will be as follows: $\tau_a \sim 7.8$, and $\tau_i \sim 1.8$, and, consequently, the time to solve the problem can be reduced by an amount equal to 6, i.e., the average time to solve the problem is reduced by more than half ($t = 10$).

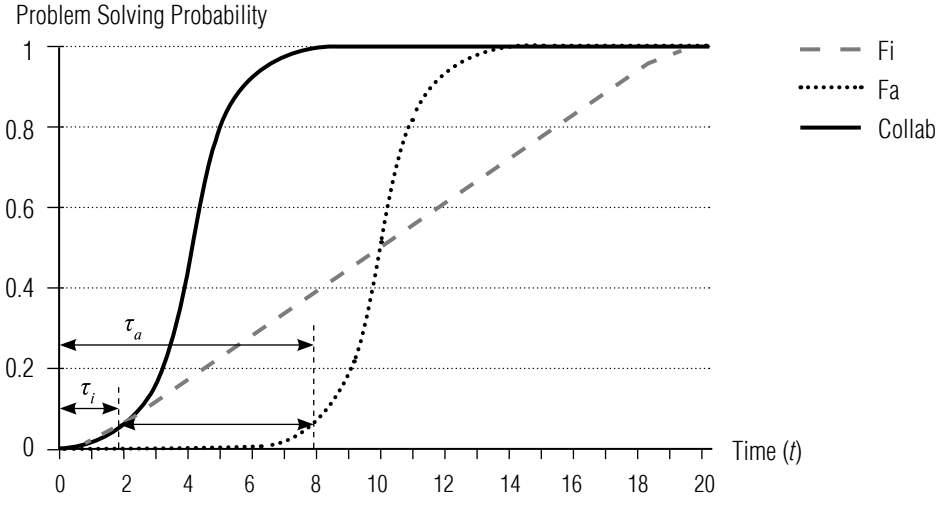

Fig. 6. Solving the problem as a result of collaboration.

Note that the creative specialist who "generates" ideas participates in solving the problem less time than the analyst (in the above case, more than four times). This suggests that for the effective use of collaboration in creative activity, it is advisable to use one "generator" of ideas to work with several analysts. The practice of managerial activity, in which the role of a creative specialist is often played by the head of the department, confirms this – employees, of whom there are always several, bring the ideas expressed by one manager to a complete form.

Thus, the technologies of collective intelligence, in addition to the technical competencies of group activity, should take into account the ability to be an analyst or a "generator" of ideas, and such abilities can change places in a person depending on the field of knowledge. A person engaged in intellectual activity alone is forced to play both roles, often postponing research in order to look at it from the other side later. It is not difficult to understand that such an approach will always lose out to teamwork, if, of course, the abilities and competencies of a person are taken into account when collaborating. When organizing scientific or research activities, it is very important to take into account how the participant solves problems – as a "generator" of ideas, or as an analyst, in order to integrate it more effectively into teamwork.

It can be shown (using a similar probabilistic approach) that synergy is manifested not only at the initial stage of solving the problem, but also at its completion. For example, when preparing research reports, different research participants often read the same text, reviewing and making their corrections. This is not because the competence of the author writing the text is less than the competence of the reviewers – a look from the outside allows you to better see the shortcomings. In addition, the labor costs for the examination, as a rule, are an order of magnitude less than the labor costs for preparing the initial document, which allows you to attract several people with different competencies and experience to work at once. The division of group work participants into those who create a document and those who review it is the basis of the method of evolutionary coordination [32] and can be used in the activities of various organizations requiring intellectual work, including the search for solutions [33].

5. Accounting for competencies and collaboration in business processes

The technologies described above make it possible to solve the problem of business processes in which creative intellectual activity plays the main role. It is necessary, on the one hand, to consider the competencies of the participants in the process, and on the other hand, to organize joint work on solving problems. In fact, we are talking about mapping the model (classifier) of competencies and group activity options into the stages of the business process. *Figure 7* shows an example of such mapping. Collective intelligence technologies are a link between information systems that automate the organization's business processes and employees of the organization who not only have certain competencies, but also solve group tasks.

In normal activities, when a small number of standard competencies are needed to participate in a business process, an employee of the organization is selected in such a way that his competencies correspond to the business process, possibly after appropriate training. When it comes to intellectual activity, the number of necessary competencies increases significantly, and they relate not only to the professional field, but also to organizational, creative abilities, which are not easy (and in some cases impossible) to teach. At the same time, group work becomes an important element, and it cannot be ignored in the management of business processes.

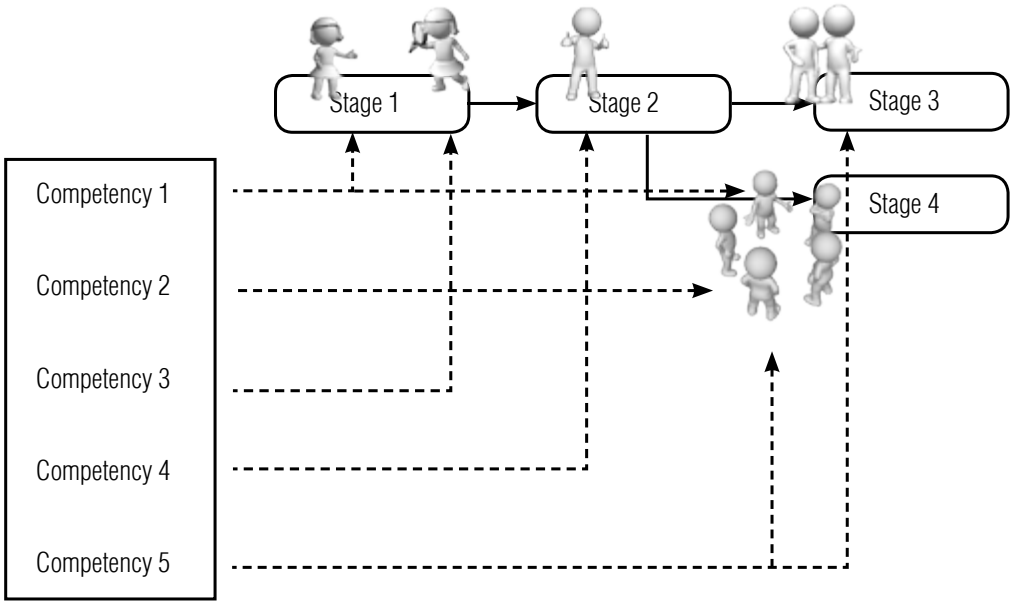

Fig. 7. Mapping of the competence model to the business process, considering group activities.

A person's competencies must necessarily be evaluated in the process of real activity, and the assessment should not serve to punish or reward employees, but to more accurately distribute them in creative work and for training. As business processes are implemented, the quality of information about employee competencies also increases. The relationship between competencies and business processes, in fact, is the relationship between implicit and explicit knowledge in the organization. It is in this connection that we should talk about effective knowledge management. The task of collective intelligence technologies is precisely to maximize the use of human intellectual capital when working with explicit knowledge or with the organizational capital of the company.

Collective intelligence technologies are finding more and more applications in various fields today. For example, the paper [34] explores the possibility of using collective intelligence technologies in online MOOC communities. The authors have shown that in educational communities, as they develop, the role of facilitators (undergraduates, teachers) decreases and the role of interaction with peers increases. The work [35] is devoted to the study of the possibility of using collective intelligence technologies in predictive analysis. Many studies are devoted to the possibilities of collective intelligence technologies in the organization of scientific [36] and expert [37] activities.

The role of competencies and the need to develop technologies for managing them in the new economy is not yet sufficiently understood. This is partly because there is no theoretical basis for human intellectual activity management technologies. At business forums today, the need for human capital development is increasingly being discussed since innovation is becoming one of the main activities of the company, but so far scientific research concerns the subject as a whole, and not specific technologies.

Deming [38, p. 87], who promoted idea of cooperation for the effective organization of corporate work, cited the orchestra as a refer-

ence example: "Musicians do not play solo, but listen attentively to each other. They gather to support each other... Thus, each of the 140 musicians of the Royal Philharmonic Society in London supports the other 139 colleagues. The sound of the orchestra is evaluated by listeners; in this case, the role is played not by the fame of the performers, but by what they get as a result". Unlike musicians listening to the sound of colleagues, the integration of professionals in the field of intellectual activity is facilitated by electronic communications, which allows us to talk about the creation of a single network mind.

Conclusion

Thus, it can be said that it is the technologies of collective intelligence, based on a competence-based approach and taking into account the synergy from group work that will make it possible to manage business processes in conditions of creative activity, which is increasingly in demand by organizations. South Korean schools have been teaching children for a long time, seating them around round tables. This is done intentionally to teach schoolchildren to group work from childhood. As companies' activities become more and more creative, it is group work, considering the specific competencies and organizational characteristics of employees that will be able to reduce the uncertainty in completing tasks. Business process management based on collective intelligence technologies will require the introduction of a competence-based approach, and the measurement of competencies will need to be carried out continuously within the framework of feedback. The measurement of competencies will allow you to adjust the business process management system to changing conditions, change or retrain employees. Organizations that will be the first to establish such business process management systems will gain competitive advantages in the field of innovative development. ■

Acknowledgements

The article was prepared based on the results of research carried out at the expense of budgetary funds under the state assignment of the Financial University under the Government of the Russian Federation.

References

1. Slavin B.B. (2017) From commodity economy to human economy. *Economics and management: problems, solutions*, vol. 7, no. 8, pp. 79–84 (in Russian).
2. Male S.A., Bush M.B., Chapman E.S. (2010) Perceptions of competency deficiencies in engineering graduates. *Australasian Journal of Engineering Education*, vol. 16, no. 1, pp. 55–68. https://doi.org/10.1080/22054952.2010.11464039
3. Loseva O.V. (2009) Methodology for assessing the state and analysis of the dynamics of human intellectual capital development in the organization. *Izvestiya Penza State Pedagogical University*, vol. 16, no. 12. pp. 75-71 (in Russian).
4. Spady W.G. (1978) The concept and implications of competency-based education. *Education Leadership*, pp. 16–22.
5. Homer M. (2001) Skills and competency management. *Industrial and Commercial Training*, vol. 33, no. 2, pp. 59–62. https://doi.org/10.1108/00197850110385624
6. Darnton G. (2002) Modelling requirements and architecting large-scale on-line competence-based learning systems. Proceedings of the *IEEE International Conference on Advanced Learning Technologies (ICALT 2002), Kazan, Russia*, pp. 170–174.

7. Altukhova N.F., Danilina O.M. (2008) On the issue of competencies in the context of corporate knowledge management. *Bulletin of the University*, vol. 21, no. 11, pp. 9–16 (in Russian).
8. Polanyi M. (2009) *The tacit dimension*. Chicago: University of Chicago Pres. https://doi.org/10.1007/s11016-010-9328-0
9. Nonaka I., Takeuchi H. (2011) *The company is the creator of knowledge*. Moscow: Olymp-Business (in Russian).
10. Slavin B.B. (2016) Technologies of collective intelligence. *Problems of management*, no. 5, pp. 2–9 (in Russian).
11. Wechsler D. (1971) Concept of collective intelligence. *American Psychologist*, vol. 26, pp. 904–907.
12. Fisher K., Fisher M.D. (1997) *The distributed mind: Achieving high performance through the collective intelligence of knowledge work teams.* New York: Amacom.
13. Weiss A. (2005) The power of collective intelligence. *netWorker*, vol. 9, pp. 16–23. https://doi.org/10.1145/1086762.1086763
14. Levy P. (1997) *Collective intelligence: Mankind's emerging world in cyberspace*. Cambridge: Perseus Books.
15. Heylighen F. (2007) The global superorganism: An evolutionary-cybernetic model of the emerging network society. *Social Evolution & History*, vol. 5, no. 1, pp. 57–117.
16. Heylighen F. (2014) The concept of the global brain. *The Birth of the collective mind: On the new laws of the network society and the network economy and their impact on human behavior.* Moscow: Lenand (in Russian).
17. Kittur A., Lee B., Kraut R.E. (2009) Coordination in collective intelligence: The role of team structure and task interdependence. Proceedings of the *27th International Conference on Human Factors in Computing Systems, Boston, MA, USA, April 4–9, 2009*, pp. 1495–1504. https://doi.org/10.1145/1518701.1518928
18. Chorost M. (2011) *World Mind.* Moscow: Eksmo (in Russian).
19. Zetssu K., Kiyoki Y. (2006) Towards knowledge management based on harnessing collective intelligence on the Web. Proceedings of the *15th International Conference on Managing Knowledge in a World of Networks (EKAW'06), Podebrady, Czech Republic, October 2006*, pp. 350–357. https://doi.org/10.1007/11891451_31
20. Luo S., Xia H., Yoshida T., Wang Z. (2008) Toward collective intelligence of online communities: a primitive conceptual model. *Journal of Systems Science and Systems Engineering*, vol. 18, no. 2, pp. 203–221. https://doi.org/10.1007/s11518-009-5095-0
21. Woolley A., Aggarwa I., Malone T. (2015) Collective intelligence and group performance. *Current Directions in Psychological Science*, vol. 24, no. 6, pp. 420–424. https://doi.org/10.1177/0963721415599543
22. Leimeister J.M. (2010) Collective intelligence. *Business & Information Systems Engineering*, no. 4, pp. 245–248. https://doi.org/10.1007/s12599-010-0114-8
23. Gregg D.G. (2010) Designing for collective intelligence. *Communications of ACM*, vol. 53, no. 4, pp. 134–138. https://doi.org/10.1145/1721654.1721691
24. Alag S. (2008) *Collective intelligence in action*. Greenwich: Manning Publications Co.
25. Malone T.W., Laubacher R., Dellarocas C. (2009) *Harnessing crowds: Mapping the genome of collective intelligence.* MIT Sloan Research Paper No. 4732-09. https://doi.org/10.2139/ssrn.1381502
26. Buecheler T., Sieg J., Fuchslin R., Pfeifer R. (2010) Crowdsourcing, open innovation and collective intelligence in the scientific method: a research agenda and operational framework. Proceedings of the *12th International Conference on the Synthesis and Simulation of Living Systems (Artificial Life XII), Odense, Denmark, 19-23 August 2010*, pp. 679–686. https://doi.org/10.21256/zhaw-4094
27. Bothos E., Apostolou D., Mentzas G. (2009) Collective intelligence for idea management with Internet-based information aggregation markets. *Internet Research*, vol. 19, no. 1, pp. 26–41. https://doi.org/10.1108/10662240910927803

28. Gruber T. (2008) Collective knowledge systems: Where the social web meets the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, pp. 4–13. https://doi.org/10.1016/j.websem.2007.11.011

29. Lykourentzou I., Vergados D., Kapetanios E., Loumos V. (2011) Collective intelligence systems: Classification. *Journal of Emerging Technologies in Web Intelligence*, vol. 3, no. 3, pp. 217–226. https://doi.org/10.4304/jetwi.3.3.217-226

30. Altshuller G.S. (1969) *Algorithm of invention*. Moscow: Moskovsky rabochy (in Russian).

31. Alterman R., Hirsch K. (2017) A more reflective form of joint problem solving. *International Journal of Computer-Supported Collaborative Learning*, vol. 12, pp. 9–33. https://doi.org/10.1007/s11412-017-9250-1

32. Protasov V. (2011) Method of evolutionary coordination of solutions. Computer and mathematical models. *Mining information and analytical bulletin*, vol. 1, no. 12, pp. 360–379 (in Russian).

33. Protasov V.I., Slavin B.B. (2017) Improving the tools of electronic democracy using technologies of collective intelligence. *Information Society*, no. 2, pp. 37–44 (in Russian).

34. Garreta-Domingo M., Sloep P.B., Hernandez-Leo D., Mor Y. (2018) Design for collective intelligence: pop-up communities in MOOCs. *AI & Society*, vol. 33, no. 4, pp. 91–100. https://doi.org/10.1007/s00146-017-0745-0

35. Kenneth J., et al. (2008) The promise of prediction markets. *Science*, vol. 320, pp. 877–888. https://doi.org/10.1126/science.1157679

36. Yu C., Chai Y., Liu Y. (2018) Literature review on collective intelligence: a crowd science perspective. *International Journal of Crowd Science*, vol. 2, no. 3, pp. 64–73. https://doi.org/10.1108/IJCS-08-2017-0013

37. Slavin B. (2014) Modern expert networks. *Open systems*, no. 7, pp. 30–33 (in Russian).

38. Deming E. (2006) *New Economy*. Moscow: Eksmo (in Russian).

About the author

Boris B. Slavin

Dr. Sci. (Econ.);

Professor of the Department of Business Informatics, Financial University under the Government of the Russian Federation, 49, Leningradsky Prospekt, Moscow 125993, Russia;

E-mail: bbslavin@gmail.com

ORCID: 0000-0003-3465-0311

DOI: 10.17323/2587-814X.2022.2.49.61

New energy efficiency metrics for the IT industry

Rafael R. Sukhov [a]
E-mail: r.sukhov@uptimetechnology.ru

Maxim B. Amzarakov [a]
E-mail: m.amzarakov@uptimetechnology.ru

Evgeny A. Isaev [b]
E-mail: is@itaec.ru

[a] INO Uptime Technology
Address: 7, Marshala Rybalko st., Moscow 123060, Russia

[b] Institute of Mathematical Problems of Biology RAS – the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences
Address: 1, Professor Vitkevich st., Pushchino 142290, Russia

Abstract

Reducing the technogenic impact of human activity on the ecology of the planet is a problem that is increasingly moving from a theoretical category into a practical one. The environmental situation is serious and requires more attention. One of the significant factors of the negative impact of humans on their environment is the emissions of harmful substances that occur during the production of electricity. The technical development of humanity and the widespread introduction of information technologies are characterized by an explosive growth in the number of electronic devices and the amount of data transmitted over information networks. This contributes to an increase in the need for computing resources for storing and processing this data, and as a result, the need for electricity is also increasing. Over the past 15–20 years, computing equipment has increased its computing power many times. The number of servers in operation is currently estimated at many millions of units, and the total energy consumption of the server park is becoming very significant in the structure of energy costs in all developed countries. In this article, we will analyze a way to reduce energy costs in the operation of servers and data centers, the application of which has a high potential for saving energy. We will give an example of a new way to evaluate the efficiency of IT equipment using a new factor – the server idle coefficient (SIC).

Keywords: energy efficiency, energy consumption, server, data center

Citation: Sukhov R.R., Amzarakov M.B., Isaev E.A. (2022) New energy efficiency metrics for the IT industry. *Business Informatics*, vol. 16, no. 2, pp. 49–61. DOI: 10.17323/2587-814X.2022.2.49.61

Introduction

The development of global (international) and regional (at the level of certain countries) social networks, Internet services, and the widespread introduction of information technologies in all sectors of the economy lead to the need to increase the efficiency of the use of computing resources. Currently, this is expressed in the consolidation of server equipment in specialized places of operation – data processing centers, which allows us to reduce costs due to the deep optimization of power supply and cooling of server equipment, as well as in the development and production of servers with improved characteristics in terms of power consumption and computing power.

Modern data centers that use innovative ways of power distribution and cooling are already approaching the theoretical limits of energy efficiency. Further technological developments in the engineering systems of data processing centers will slightly increase energy efficiency [1], while significantly increasing their cost.

On the one hand, according to the available studies [2], it can be concluded that, e.g. in the UK, approximately 10% of total electricity production is consumed by commercial and government data centers and IT systems located in them.

On the other hand, modern servers have come close to the limits of compactness and energy efficiency. The current "silicon" technological basis will not provide a significant reduction in energy consumption with comparable computing power in the near future.

And, despite the active development in recent decades of so-called "green" energy – that is, energy that uses alternative energy sources to traditional ones, operating on oil, extracted natural gas and coal, traditional energy sources during combustion emit carbon dioxide into the atmosphere, which leads to an increase of the greenhouse effect and global warming.

Currently, more than two-thirds of the energy sources in world production are traditional ones that cause significant harm to the environment (*Fig. 1*) [3].

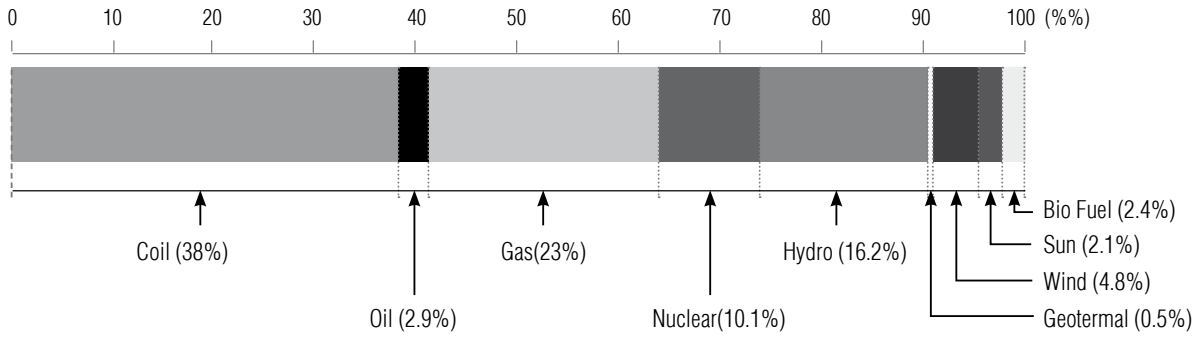

Fig. 1. Share of electricity generation by energy sources in the world.

Consolidation of IT resources improves efficiency both by optimizing maintenance costs and by reducing costs associated with power supply and subsequent additional cooling costs. The emergence of a new industry – data centers – is a logical development of the IT industry as a whole.

A modern data processing center (DPC) is a high-tech enterprise that provides continuous and reliable power supply to servers. The main resource that is managed, distributed and supplied by the data center is electricity, the efficiency of which determines the overall efficiency of the data center in particular and the IT industry as a whole.

Understanding the existing limitations pushes researchers around the world to look for new ways to reduce energy costs in the IT industry.

In the work of a data center, two of the most significant consumers of electricity can be distinguished: server equipment and auxiliary engineering systems (air conditioning, power distribution and uninterruptible power supply, etc.). For each type of consumer there are various energy efficiency metrics that, to one degree or another, make it possible to give a qualitative or quantitative assessment of each of the consumers. However, the authors of this article are not aware of a single metric that would allow both of them to be comprehensively combined and make it possible to assess the impact on the final energy efficiency of the data center operation.

That is why it becomes important to create a single metric that would allow us to evaluate the efficiency of using electricity in a data center when performing calculations, regardless of which processor the server uses, or what cooling technologies are used in the data center [1, 2].

However, the task of measuring server energy efficiency is not as simple as it seems at first glance.

1. Current energy efficiency indicators

The efficiency of a data center, in terms of energy costs for maintaining the operation of server equipment, is evaluated by using the power usage effectiveness (PUE) coefficient [4]. This coefficient appeared in 2007 and has firmly entered the everyday life of specialists. It allows you to instantly assess the energy efficiency of a data center as an object of engineering infrastructure.

PUE is calculated as the ratio of the total data center energy consumption (including all energy costs, both IT and support costs) to the energy costs of the data center server equipment, i.e. PUE shows how much electricity the data center consumes to ensure that the server equipment works properly:

$$\text{PUE} = \frac{P_{total}}{P_{IT}}, \text{ where} \qquad (1)$$

P_{total} – the total amount of energy consumed by the data center;

P_{IT} – the amount of energy consumed by all IT equipment in the data center in the same time.

According to the Uptime Institute data, the PUE coefficient decreased sharply from 2006 to 2013 [5], however, after 2013, the PUE coefficient remains approximately at the same level (*Fig. 2*) and fluctuates at the level of 1.5–1.7.

Each watt of electricity consumed by the server is associated with energy costs for its "delivery": transmission, conversion, cooling, lighting, etc. Currently, this additional cost required to keep the server up and running is 0.5 to 0.8 watts for every watt the server consumes.

In the professional environment such an effect is called a cascading effect (*Fig. 3*).

At the moment, it is not expected that the energy efficiency of the data center can be significantly improved. Individual data center projects show phenomenally low PUE values of 1.06–1.1. However, it should be noted that

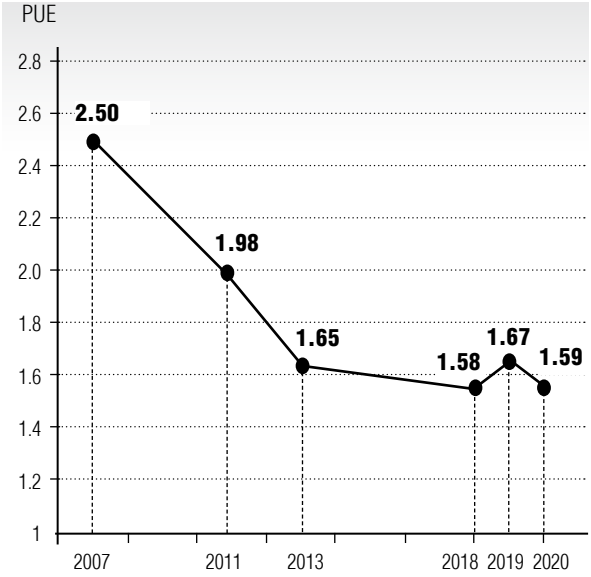

Fig. 2. Decrease in the dynamics of improving the energy efficiency of data centers [5].

such values are achieved under very limited conditions, while using harsh, complex and interdependent operating conditions of engineering systems [6] and IT equipment. In most cases they are difficult to achieve or practically unrealizable [7].

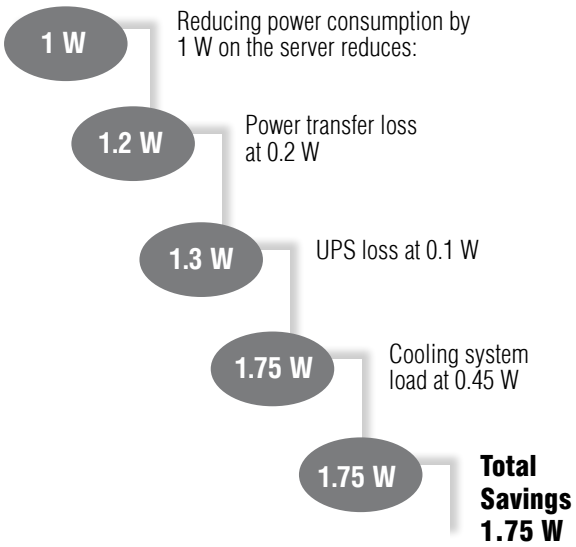

Fig. 3. Cascading effect.

Another component of energy costs, which has a high weight in total energy consumption, lies directly in the IT equipment.

Almost since the advent of the first computers, there has been a relentless struggle to reduce the size and power consumption of computers. And in this area amazing results have been achieved. For example, you can look at the results achieved by AMD [8]. Over the past six years, AMD has improved the power efficiency of its mobile processors by 31.7x (*Fig. 4*).

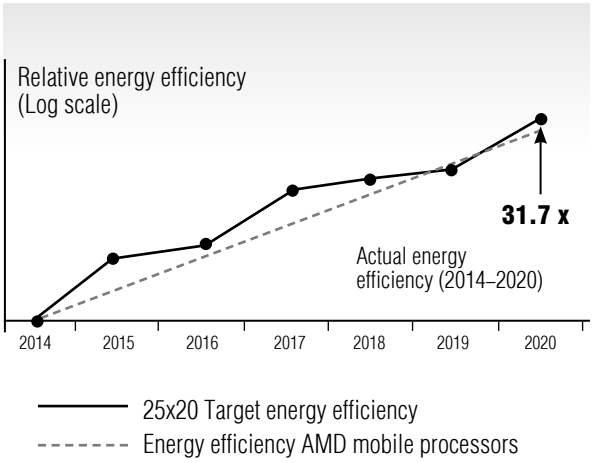

Fig. 4. Increasing the energy efficiency of mobile processors [8].

We observe an interesting effect here. In the last 15–20 years, the average power consumption of the processor itself and the computer (server) as a whole has been decreasing insignificantly, and the processor performance, i.e., computer (server) is growing significantly. Formally speaking, the computer processor becomes much more energy efficient, because the performance of one operation requires significantly less energy than was required before.

And, arguing in this vein, it can be assumed that in this case, the overall energy consumption of the IT industry and the data center industry should be reduced, which is a favorable factor for reducing resource consumption.

However, this does not happen. The main reason, in our opinion, lies in the fact that all the increase in processor performance is "spent" on meeting the needs of society in the best possible way, for example, making video content of better quality, delivering this content to the user more quickly, generating new content, attracting new users, creating new IT services that are increasingly using neural networks, built in their turn also on server clusters, etc.

So far, we do not see that there is any limit to this "arms race," but at the same time, can we find ways to reduce energy consumption which will be less dependent on the above-mentioned drivers that generate demand for IT services?

2. Energy management

Based on the data given in the article [1], we can take as a basis the energy distribution in the average server, shown in *Table 1.*

The table shows that the main consumer is the server processor – this is almost 64% of the total server power consumption. At the same time, it is also known that, on average, depending on the nature of the calculations performed, server processors are busy with useful work from 2 to 30 percent of the time [9, 10]. From this, we can conclude that while the processor consumes a significant part of the energy in the server, in fact the processor most of the time does not perform useful work.

For server manufacturers and other participants in the IT industry, this fact is not a secret, and in order to minimize energy losses during server downtime, as well as to reduce the risk of processor overheating and the occurrence of negative effects in semiconductors (tunnel effect), they have been building special server management systems into server management systems for a long time: algorithms and methods for managing energy saving which allow you to flexibly configure the operation of all elements of the server (processor, RAM, permanent memory, video card, etc.) and thereby reduce unproductive losses of the server as a whole. Research in the direction of how to make server energy management systems more efficient is ongoing.

Examples of such studies include:

- Berkeley laboratory study: Comparing server energy use and efficiency using small sample sizes (comparison of energy consumption and efficiency of servers on the example of servers of typical sizes of small format) [11];

Table 1.

Distribution of energy costs in the server

Component	Consumption, W	Consumption, %
Processor	115	63.89%
RAM	15	8.33%
Hard disks	2	13.33%
Network communication	5	2.78%
Cooling	8.87	4.93%
Power supply	11.13	6.18%
Other energy costs	1	0.56%

- Report at the international conference on high-performance computing, data transfer, storage and analysis: Energy-aware data transfer algorithms (energy-saving data transfer algorithms) [12];
- Report at the 9th International Conference on Applied Energy: Development of a simple power consumption model of information technology (IT) equipment for building simulation (development of a simple power consumption model of information technology (IT) equipment for building simulation) [13].

And in the future, we will see new generations of servers and telecommunications equipment which, thanks to technologies created on the basis of such research, will have better characteristics than today.

As the main areas that would improve energy efficiency, the following can be distinguished:

- measurement of energy consumption;
- control of energy consumption;
- energy management.

3. New trends in efficiency

As we noted earlier, the energy efficiency of a server depends on how much it is loaded with useful work, performing calculations.

There are many different metrics for measuring the energy efficiency of computing. The simplest metric is the measurement of the amount of energy expended to produce a floating-point calculation [14].

For personal computers, the SPEC metric is widespread [15]. This calculates the amount of energy spent on performing typical actions on a computer running different operating systems.

Energy Star (an organization under the Environmental Protection Agency EPA, USA) has developed a whole program for measuring efficiency – SERT [16], including the measurement of power consumption in the production of various operations.

SUN Microsystems proposed the SWAP (space, wattage and performance) metric [17].

However, all the metrics mentioned above have one common drawback, namely, the need to determine the performance of the server. For the SWAP metric, useful work is determined directly by indicating the complex actions performed. For the SPEC metric, performance is determined by the performance of known programs. For SERT, this is the performance of specialized software execution.

And these restrictions significantly narrow the possibilities of using such metrics, which in their turn does not make it possible to extend this or that technique to all types and types of servers and computing.

But on the part of society, represented by the state, there is a serious request for the search and implementation of new mechanisms and technologies that allow additional energy savings when using servers.

And in this sense, the state performs a very important function, namely, it sets rules and standards, sets metrics, or boundary values, the achievement of which becomes mandatory, thereby stimulating the industries involved in the search for new solutions and their implementation in technologies and equipment.

A widespread example of energy efficiency improvement in the world is the introduction of energy efficiency classes in industrial and household electrical equipment, as well as the establishment of control dates when restrictions begin to operate, or a complete ban on the circulation of equipment with low energy efficiency indicators, as introduced in Russia in the form of a state industry standard (GOST) [18], or on the example of the provisions of the EU Commission [2, 19].

Another illustrative example of the implementation of an energy efficiency program is the LEAP program (Lower Energy Acceleration Program) – the program for the develop-

ment of lower energy consumption [20]. This program is an initiative run by the Dutch government and is part of such an EU program. Within the framework of this program, studies are carried out aimed at finding areas where significant improvements are possible in terms of energy efficiency.

In particular, the report prepared by Certios and WCoolIT by order of the Netherlands Enterprise Agency (Netherlands Industrial Agency) is very interesting. The report is called "LEAP Track 1 "Powermanagement" Pilot analysis" – LEAP Stage 1 Pilot analysis of energy saving [9].

Before proceeding directly to such metric as the server idle coefficient (SIC), it will be useful to describe the criteria used in calculating the metric.

The authors of the SIC metric do not claim to define performance or useful work. However, the operation of any computing device is characterized by the performance of "parasitic calculations" (for the processor – the NOOP instruction). From the authors' point of view, any computational action is useful, regardless of what kind of calculation is performed and how "useful" it is to the consumer. Nevertheless, most of the time, the processor does not perform computational operations, but is in standby mode, in which the NOOP instruction is executed. The energy expended to perform this operation is considered a direct waste.

The idea of the study was to try to understand how efficiently the processor time is used, how much they affect the energy savings implemented in the hardware of servers or operating systems, the ability to manage energy savings, and also whether it is possible to present the actual energy efficiency of the server in an equally understandable way for all kinds.

For the study, statistics were taken from a pool of servers operating with a real load and different profiles of the load itself. We tested all the main power saving modes built into servers and some operating systems in various load modes.

The data obtained was analyzed and presented in a very visual way, demonstrating the real situation with the energy consumption of these servers. *Figure 5* shows the dependence of energy consumption depending on the load on the server. An example of server operation with power management enabled is given. The direct dependence of power consumption on processor load is clearly visible. We also see that this server has explicit maximums and minimums of the payload on the processor associated with the specifics of the applications running on this server.

The energy efficiency assessment was performed using a methodology to measure server processor uptime and idle time, which correlated with energy monitoring data at these points in time.

In this case, the SIC coefficient was proposed as the final server efficiency coefficient – the server idle coefficient.

$$\mathrm{SIC} = \frac{E_{total}}{E_{total} - E_{idle}}, \tag{2}$$

$$\mathrm{SIC\%} = 100\% \cdot \frac{E_{total}}{E_{idle}}, \tag{3}$$

where

E_{total} – total energy consumed by the server;

E_{idle} – energy consumed by the server during idle time.

Indicator (2) is interpreted by analogy with the PUE coefficient, i.e., the closer the value is to one, the more energy the server spends on useful work out of the total amount of energy spent.

The indicator (3) is the percentage of power consumption when the server was idle to the total amount of power spent. That is, the closer this indicator is to 100%, the less time this server performs useful work.

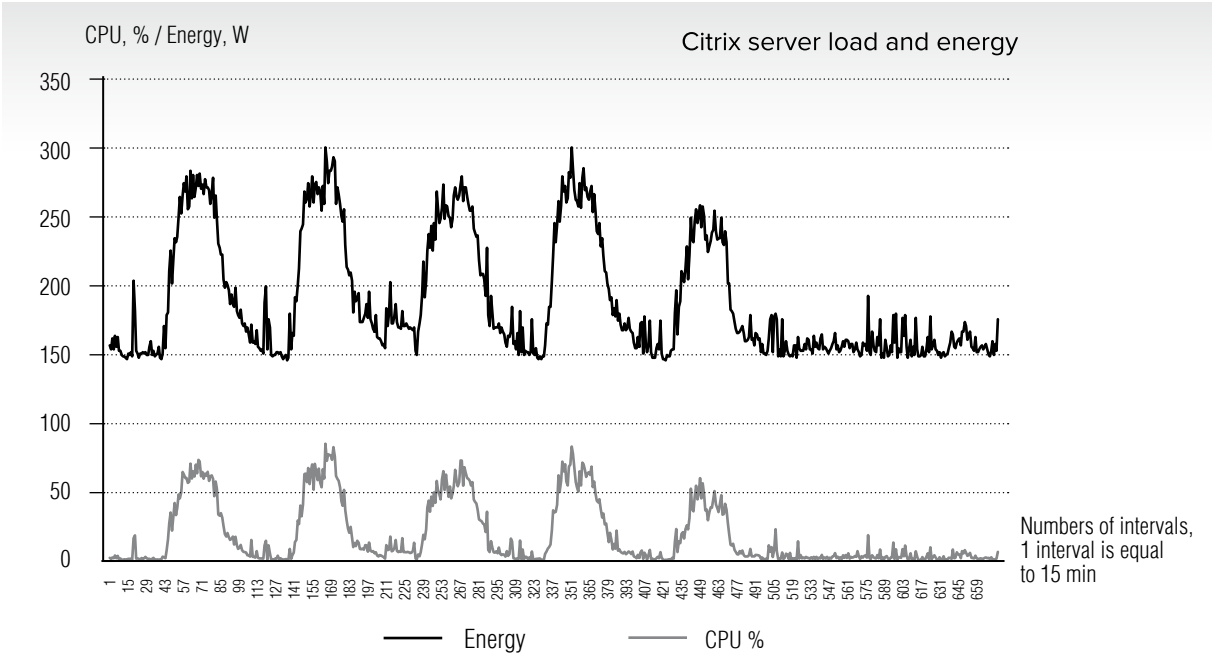

Fig. 5. Graphs of the dependence of the consumed energy of the processor on the load [9].

The study noted that:

- SIC % indicators in different groups of servers varied from 34% to 91%;
- there is a gap in the knowledge of the technical staff responsible for the operation of servers about the role of virtualization in power supply management;
- there are still strong prejudices regarding a significant decrease in the performance of systems configured to use dynamic power saving modes;
- most server cluster operators do not have any clear rules and policies regarding server power management. And where these policies exist, they most often override in favor of maximum server performance.

The obtained research data show a very high potential for reducing inefficient energy consumption.

As a practical example, we can analyze the data presented in *Fig. 6.*

Graphs of *Fig. 6* shows data on the server processor load and its power consumption in two operating modes: the upper graph is the high-performance mode; the lower graph is the power saving mode under the control of the server operating system.

The received data was processed and presented in the total values of the energy consumed in different modes and the SIC was calculated for the power saving mode enabled:

- total energy consumption for the period: 24.5 kW;
- total energy spent during idle periods: 8.43 kW;
- Average CPU idle time: 60.4%.

Using formulas (2, 3), the SIC coefficient is calculated as a percentage and as a ratio of the amount of total energy spent to the energy spent on computing needs:

$$SIC\% = 8.43/24.4 = 34.4\%;$$

$$SIC = 1.5;$$

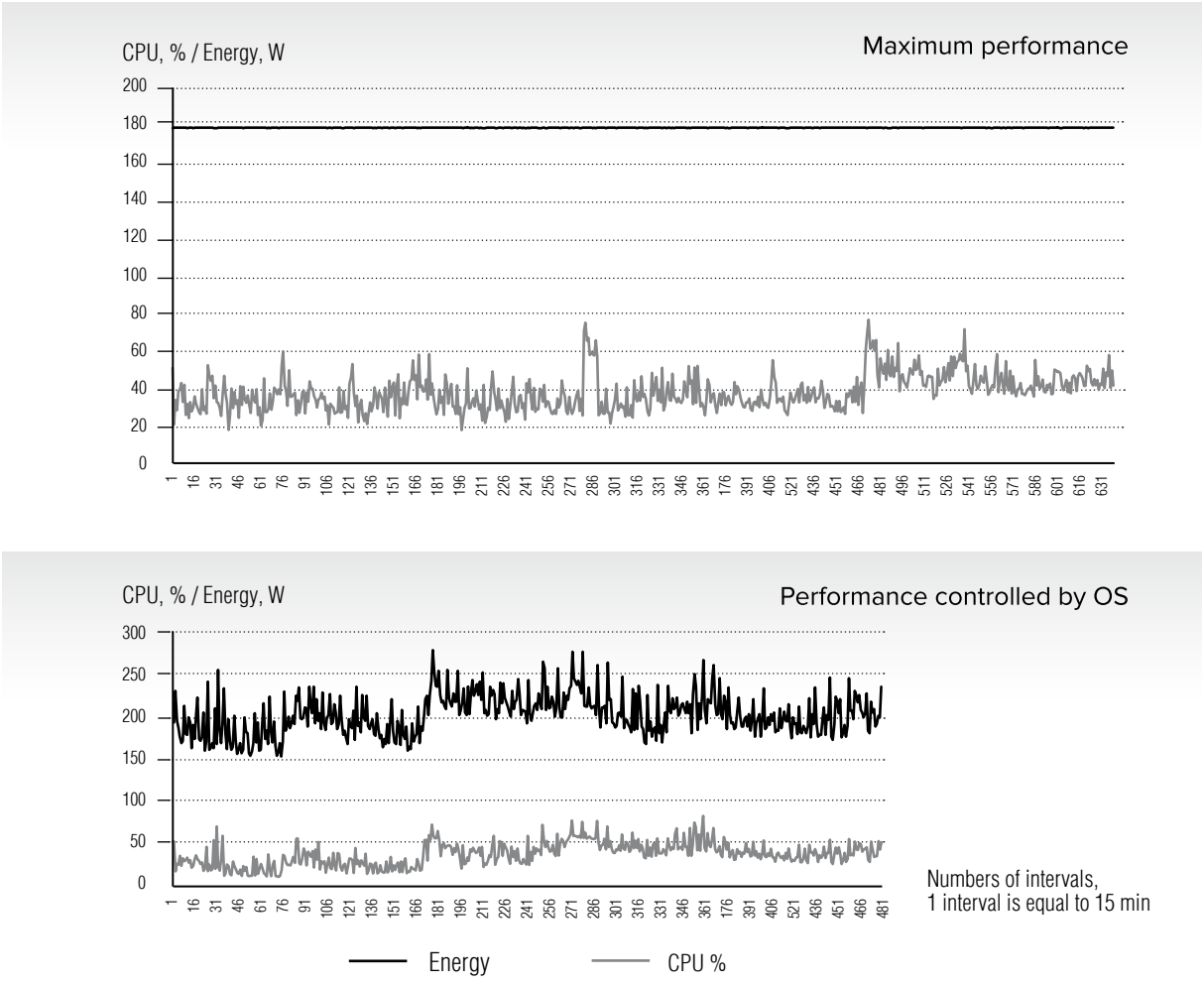

Fig. 6. Server load and different power saving modes [9].

and in power saving off mode:

- SIC% = CPU idle time % = 60.4%
- SIC = 2.53.

Thus, the SIC coefficient gives a visual representation of the efficiency of the server.

This approach seems to be interesting and promising, as it allows you to evaluate the efficiency of server hardware without reference to its real energy consumption parameters. This makes it possible to evaluate the effectiveness of the complex: server – software – clustering – dynamic load management.

The development of this approach will certainly require a deeper study and the development of a unified methodology and data extraction mechanisms for the evaluation and subsequent interpretation of the data obtained.

4. Evaluation of the practical benefits of using the server downtime factor

Improving the idle factor value is possible in two ways. The first direction is to improve the scheduling of instruction execution by the operating system and software. That is pos-

sible, for example, due to better distribution of calculations between servers. The aim to increase the useful time of computing power is the task for developers of information systems and software.

The second direction – reducing the power consumption of server equipment in the absence of calculations – is a task for developers of processors and server equipment.

It can be argued that the widespread introduction of the SIC metric makes it possible to create a single basic environment for a comparable and reliable assessment of the efficiency of modern servers, which, in combination with government regulation, will give a great incentive to software developers to increase the payload of computer processors and to equipment developers to increase the energy efficiency of their devices.

The SIC coefficient also makes it possible to apply new methods for the practical implementation of energy efficiency programs in data centers with minimal investment costs for the implementation of such practices.

Switching to server power monitoring using the SIC coefficient allows you to perform these procedures in real time, without spending resources on costly and time-consuming measurements of the PUE coefficient or its derivatives.

Automated SIC background calculation on each server and aggregation into a single analysis system can make it possible to almost instantly identify server hardware that is being used inefficiently.

Based on the cautious estimates of the authors of the study [9] in the possibility of energy savings of at least 10% for highly loaded servers when switching to dynamic server power management, we estimate the amount of potential energy savings for a small typical data center with the following characteristics:

Total number of servers in operation: 2 000 pcs.;

Average power consumption per server: 200 W;

Influence of cascade effect (PUE coefficient) data center: 1.7.

Suppose that, as a result of controlling the SIC coefficient, we managed to reduce the power consumption of one server by 10% on average, we will get:

$$P_{savings} = 200\ (\text{W}) \cdot 10\% = 20\ (\text{W}) - \text{for one server;}$$

Taking into account the total number of servers, equal to 2000, we get:

$$P_{total\ savings\ IT} = 20\ (\text{W}) \cdot 2000 = 40\ (\text{kW}) - \text{for total IT load,}$$

Thus, we see that the seemingly insignificant 20 W savings in power consumption per server on the scale of our example, in the amount of 2000 servers, allow us to obtain a total reduction in electrical power consumption in the amount of 40 kW.

Taking into account the PUE coefficient, we get:

$$P_{total\ savings} = 40\ (\text{kW}) \cdot 1.7 = 68\ (\text{kW}) - \text{for data center,}$$

Reducing the power consumption of auxiliary engineering systems necessary for the normal functioning of the servers adds another 28 kW of electrical energy to the overall savings.

Below we will move from consumed electricity measured in kW·h to consumed electricity measured in kW·h and calculate the amount of electrical energy savings per month:

$$P_{total\ savings} = 68\ (\text{kW}) \cdot 720\ (\text{hours}) = 48\ 960\ (\text{kW·h}),$$

where 720 hours – is the average number of hours per month.

Next, we calculate the amount of electricity savings per year:

$$P_{total\ savings} = 68\ (kW) \cdot 8640\ (hours) = 587\ 250\ (kW \cdot h),$$

where 8 640 hours – is the average number of hours per year.

If we assume that in Moscow the cost of one kWh of electricity in 2021 is 7.11 rubles, then for the year we will get a total savings in cash:

$$S_{total\ savings} = 587\ 250\ (kW \cdot h) \cdot 7.11\ (rub.) = 4175\ (mln\ rub.).$$

The given example of a data center, in comparison with the giants of the industry, has a small capacity, about 700 kW, but this example clearly shows what potential there is in reducing the overall energy consumption of data centers and the IT industry.

Conclusion

This article discusses the issue of the main trends in the field of increasing the efficiency of using electricity in the operation of server clusters and systems.

The energy efficiency of the calculations performed at the level of individual devices, based on the number of operations per unit of power, and the reduction in operational losses of engineering systems of data centers [21] are the main driver for improving the energy efficiency of the IT industry as a whole.

It is shown that a 10% reduction in energy consumption on servers reduces the required power from 700 to 632 kW and provides significant savings in the cost of paying for consumed electricity.

It is necessary to pay close attention to new ways of improving the energy efficiency of the IT industry, namely of changing approaches to managing server systems. Combining single devices – servers into clusters and systems with simultaneous management of processor capacity loading, as well as dynamic management of the power supply of server elements, in addition to the main ways to increase energy efficiency, provides another powerful tool for management and control.

The introduction into practice of new methods for evaluating the efficiency of server hardware, such as the server idle coefficient (SIC), can give a qualitatively new assessment of the efficiency of calculations, regardless of how energy-efficient the server processor is by itself. ■

References

1. Amzarakov M.B., Sukhov R.R., Isaev E.A., Amzarakova A.M. (2019) Energy efficiency of the Data Processing Center as a combination of engineering and IT infrastructure. *Instruments and Systems: Monitoring, Control, and Diagnostics*, no. 12, pp. 47–52 (in Russian).
2. BCS, the Chartered Institute for IT (2021) *A new European roadmap to cleaner, greener data centres*. Available at: https://www.bcs.org/content-hub/a-new-european-roadmap-to-cleaner-greener-data-centres/ (accessed 28 February 2022).
3. IEA (2021) *World gross electricity production by source, 2019*. Available at: https://www.iea.org/data-and-statistics/charts/world-gross-electricity-production-by-source-2019 (accessed 28 February 2022).
4. Wikipedia, the free encyclopedia (2022) *Power usage effectiveness*. Available at: https://en.wikipedia.org/wiki/Power_usage_effectiveness (accessed 28 February 2022).
5. Lawrence A. (2020) *Data center PUEs have been flat since 2013*. Uptime Institute. Available at: https://www.datacenterdynamics.com/en/opinions/data-center-pues-have-been-flat-2013/ (accessed 28 February 2022).

6. Delta Power Solution (2014) *Overview of green energy strategies and methodologies for modern data centers*. Available at: https://www.deltapowersolutions.com/en-us/mcis/white-paper-overview-of-green-energy-strategies-and-techniques-for-modern-data-centers.php (accessed 28 February 2022).

7. Super Micro Computer, Inc. (2019) *Supermicro second annual green data center report finds opportunity for saving millions in energy costs, and reductions in E-Waste*. Available at: https://www.supermicro.com/en/pressreleases/supermicro-second-annual-green-data-center-report-finds-opportunity-saving-millions (accessed 28 February 2022).

8. Advanced Micro Devices, Inc. (2022) *AMD 25x20 energy efficiency initiative*. Available at: https://www.amd.com/en/technologies/25x20 (accessed 28 February 2022).

9. Harryvan D., Verzijl M., Amzarakov M. (2020) *LEAP Track 1 'Powermanagement' Pilot analysis*. Netherlands Enterprise Agency. Available at: https://amsterdameconomicboard.com/app/uploads/2020/10/LEAP-Track-1-'Powermanagement-Pilot-analysis.pdf (accessed 28 February 2022).

10. Meisner D., Gold B.T., Wenisch T.F. (2009) PowerNap: eliminating server idle power. *ACM SIGARCH Computer Architecture News*, vol. 37, no. 1, pp. 205–216. https://doi.org/10.1145/2528521.1508269

11. Coles H.C., Qin Y., Price P.N. (2014) *Comparing server energy use and efficiency using small sample sizes*. Lawrence Berkeley National Laboratory. Available at: https://buildings.lbl.gov/publications/comparing-server-energy-use-and (accessed 28 February 2022).

12. Alan I., Arslan E., Kosar T. (2015) Energy-aware data transfer algorithms. Proceedings of the *International Conference for High Performance Computing, Networking, Storage and Analysis (SC'15)*. Article No.: 44, pp. 1–12. https://doi.org/10.1145/2807591.2807628

13. Cheung H., Wang S., Zhuang C. (2017) Development of a simple power consumption model of information technology (IT) equipment for building simulation. *Energy Procedia,* vol. 142. pp. 1787–1792. https://doi.org/10.1016/j.egypro.2017.12.564

14. TOP500.org (2021) *Green500 results*. Available at: https://www.top500.org/lists/green500/ (accessed 28 February 2022).

15. *Standard Performance Evaluation Corporation*. Available at: http://www.spec.org/index.html (accessed 28 February 2022).

16. Fogle R. *How to measure server efficiency with SERT*. Energy Star. Available at: https://www.energystar.gov/products/ask-the-experts/how-to-measure-server-efficiency-with-sert- (accessed 28 February 2022).

17. Phys.org (2005) *Sun introduces new metric for server efficiency*. Available at: https://phys.org/news/2005-12-sun-metric-server-efficiency.html (accessed 28 February 2022).

18. GOST R 51749-2001 (2002) *Energy saving. Energy-consuming equipment for general industrial use*. Available at: http://docs.cntd.ru/document/1200012993 (accessed 28 February 2022).

19. Commission Regulation (EU) 2019/424 of 15 March 2019 (2019) *Official Journal of the European Union,* no. L 74, pp. 46–66. Available at: https://eur-lex.europa.eu/eli/reg/2019/424/oj (accessed 28 February 2022).

20. The Amsterdam Economic Board (2020) *LEAP – Lower Energy Acceleration Program*. Available at: https://amsterdameconomicboard.com/en/results/lancering-lower-energy-acceleration-programme-leap/ (accessed 28 February 2022).

21. Amzarakov M.B., Sukhov R.R., Isaev E.A. (2014) The modular data center: a holistic view. *Business Informatics,* vol. 29, no. 3, pp. 7–14.

About authors

Rafael R. Sukhov

Financial manager, INO Uptime Technology, 7, Marshala Rybalko st., Moscow 123060, Russia;

E-mail: r.sukhov@uptimetechnology.ru

ORCID: 0000-0002-8124-137X

Maxim B. Amzarakov

Director, INO Uptime Technology, 7, Marshala Rybalko st., Moscow 123060, Russia;

E-mail: m.amzarakov@uptimetechnology.ru

ORCID: 0000-0001-6229-8592

Evgeny A. Isaev

Cand. Sci. (Tech.);

Senior Research Fellow, Institute of Mathematical Problems of Biology RAS – the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, 1, Professor Vitkevich st., Pushchino 142290, Russia;

E-mail: is@itaec.ru

ORCID: 0000-0002-3703-447X

DOI: 10.17323/2587-814X.2022.2.62.73

Trusted artificial intelligence: Strengthening digital protection

Sergey M. Avdoshin
E-mail: savdoshin@hse.ru

Elena Yu. Pesotskaya
E-mail: epesotskaya@hse.ru

HSE University
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

Abstract

This article is devoted to aspects associated with the up-coming need for mass implementation of neural networks in the modern society. On the one hand, the latter will fully expand the capabilities of state institutions and society delegated to perform numerous tasks with higher efficiency. However, a significant threat to democratic institutions obliges society to set out the concept of reliable artificial intelligence (AI). The authors explore a new concept of a trusted AI necessary for the scientific and international community to counter improper future digital penetration. Explaining to what extent digital transformation is mandatory, the authors emphasize the numerous dangers associated with the applications of artificial intelligence. The purpose of the article is to study the potential hazards of neural networks' abuse by the authorities and the resistance to them with reliance on the trusted AI. Studying various aspects of digital transformation and the use of artificial intelligence technologies, the authors formalize the dangers associated with the emergence and propose an approach to the use of digital protection technologies that can be trusted.

Keywords: neural networks, digital protection, trust, artificial intelligence, society

Citation: Avdoshin S.M., Pesotskaya E.Yu. (2022) Trusted artificial intelligence: Strengthening digital protection. *Business Informatics*, vol. 16, no. 2, pp. 62–73. DOI: 10.17323/2587-814X.2022.2.62.73

Introduction

Undoubtedly, these days the general trend is around the rapid introduction of modern digital technologies into multiple processes within society, where artificial intelligence (AI) is playing a central role [1, 5]. Numerous reforms are noted within both mechanisms already in place and those just emerging to bring humanity to a new stage in its development. Spheres that are actively implementing digital technologies in order to modernize processes and accelerate economic development are no exception.

McCarthy coined the concept of artificial intelligence (AI) in 1955 [6, 7]. AI is a system's ability to interpret data, to learn from such data, and to use gained data to achieve specific goals and tasks through flexible adaptation [6].

We take note that the discussion on the definition of AI has not yet led to a clear result satisfying all stakeholders of AI technologies. Since 2016, the artificial intelligence industry has broken out with the support of cloud computing and big data. Today cloud based artificial neural networks and deep learning form the basis of most applications we know under the label of AI.

The Artificial Neural Network is to some extent modeled on the structure of the biological brain. With these networks, various problems can be solved in a computer-based way. It consists of an abstracted model of interconnected neurons, whose special arrangement and linking can be used to solve computer-based application problems in various fields such as statistics, technology or economics [8].

AI's analytical and cognitive tools allow technology owners to analyze significant amounts of data, meaning they can immediately detect and effectively respond to changes in the agenda. Relying on complex mathematical algorithms, it is possible to increase the level of transparency, optimize internal processes of interdepartmental interaction and stimulate innovative activities, ultimately establishing a higher level of trust [9].

The consequences of the mass introduction of AI are expected to be beneficial for society as a whole. This means that the issue of creating and subsequent implementation of a centralized digital ecosystem aimed at improving the interconnection and interaction of all stakeholders (government, business, associations and individuals) is on the agenda for many companies and organizations [9–11]. The main role in this transformation will be assigned to AI.

The field for AI introduction is truly vast. However, nowadays, there is a noticeable discord, undermining whether the widespread use of AI is timely [12]. On top of issues with privacy, hacker attacks, technological singularity, etc., already widely scrutinized, there is concrete evidence of an equally vital danger [5, 13–15]. The discussion here lies around the accumulation of digital power in the hands of a narrow group of people. It can definitely be argued that a set of modern algorithms opens up almost limitless possibilities. The detailed scenario reflects the concept of the emergence of digital dictatorship in the modern world and the need for digital protection to regulate the unfair use of technology.

Thus, the concept of trusted AI has been proposed as a countermeasure to the unethical use of neural networks [6, 12]. The concept's framework is supposed to find the golden mean between the progressiveness of AI application strategies and the protection of

ethical and moral aspects of human life. However, the existing legislative base is negligible and international consensus is absent. Therefore, the delegation of any tasks to neural networks seems to be of high risk now. Not only would the traditional model of public administration be challenged, but also core human values might be threatened [6, 7, 12]. Moreover, understanding the immense gap between the principles of international law and reality, where these principles are constantly violated, we have to realize the insufficiency of only defining abstract principles. Specific counteractions to prevent the use of neural networks against society should be determined. A comprehensive regulatory system is needed to encourage technological progress and define concrete steps to combat rights violations.

Thus, the purpose of this article is to study the potential hazards of neural networks' abuse by technology owners and ways to resist them based on the concept of trust. To avoid problems that can harm a person by distorting, stealing or leaking data, it is necessary to make sure that the results of AI work can be trusted. In this paper, potential problems related to issues of trust, confidentiality and reliability are investigated. The concept of "trusted artificial intelligence" is considered, as well as the phenomenon of digital protection. Section 1 examines in detail the nature of violations of citizens' rights from the point of view of four different spheres of life: political, social, cultural and economic. We point out the irreversibility of the process of implementing neural networks in the context of digital transformation and describe the idea of trusted AI. Section 2 is devoted to discussing possible response strategies and proposals for establishing digital protection. Our study concludes with a summary of key findings.

1. Prospects and methods of artificial intelligence

1.1. New possibilities of artificial intelligence

Scientific and technological progress cannot be stopped. Under the pressure of the rapid development of the IT sector of the world economy, countries are forced to meet the ever-changing demands of business or risk facing a real digital abyss in management [9]. Artificial intelligence has become a part of our lives as smart systems are used in many areas, from client analytics and search engines to voice assistants and medical research. In the medical field, systems that recognize pathologies from video recordings of endoscopic examinations are being developed; in transport – autopilots and traffic management systems; in finance – systems that identify customers or identify suspicious transactions that may indicate tax evasion or money laundering. It is safe to say that further global economic development and progress directly depend on how effectively various industries learn to use artificial intelligence. However, with the development of technology, the problem of trusted artificial intelligence has become more acute, as any problem can have serious consequences. Users want to be sure that the model has a high degree of accuracy and that its results are fair and easily interpreted. For example, incorrectly calibrated sensors of a car equipped with autopilot can cause an accident. Errors in the control of the infrastructure that AI relies upon can lead to a leak of patients' personal data, incorrect medical diagnosis or identity theft. Regarding business and industry, a low level of AI software can lead to delays in transportation, damaging the supply chain.

In today's agenda, special attention is paid to the social aspect of AI's mass introduc-

tion. Neural networks' computational and analytical capabilities, far superior to human performance, open up new horizons for public institutions. In addition, AI does not have a limited reserve of endurance and is always available. Consequently, significant amelioration is expected in traffic systems, healthcare, maintaining public order and public services personalization, including education [15]. Impressive progress has already been achieved in providing public services to citizens and legal entities. For now, the scope of AI's implementation is quite limited. The main categories are processing of requests (social payments, migration, citizens' questions, etc.), filling out and searching for documents, translating texts and drafting documents [4].

Significant successes are predicted for neural networks in the field of economic management. Plans are devoted mainly to the improvement of resource allocation and logistics efficiency. It is necessary to restructure and optimize supply channels, warehouse systems, and recycling [15]. Neural networks will be crucial for the promising concept of the "smart city," with control of CCTV cameras, electricity grids, water supply, transport systems, etc. delegated to them [7, 16].

Note that significant changes have occurred under the influence of the COVID-19 pandemic, in particular, in the field of teaching and learning. Academic institutions are moving to digital technologies to provide their students with more resources. Thanks to technology, students now have more opportunities to learn and improve skills at their own pace and on an individual trajectory, now having the opportunity to pass control stages using online tests. Online proctoring services are gaining more and more popularity, in which the subject's face is identified and analyzed to predict his emotions. In addition, aspects such as a phone, a book, or the presence of another person are detected. This combination of models creates an intelligent rule-based inference system that can determine whether there has been any cheating during an exam or test.

Here, the question about the correctness of the system and the adequacy of the assessment of behavior may arise. Any failures and abuses are fraught with negative consequences in terms of academic integrity, discrediting the idea of both offline and online learning. Among the key risks, it is worth highlighting a violation of confidentiality, compromised availability or compromised accounts, leaks of personal data, or distortion of results.

Big data is necessary for the successful development of machine learning models. The quantity, quality and availability of big data affect the efficiency and accuracy of the models being trained. Therefore, many companies are interested in continuous data collection about their consumers. Many systems collect information that is not subject to disclosure: videos and photos from video cameras, speech recordings and financial transactions. Unreasonable use of this information, errors in the model or data theft, can cause threats to the security of individuals or even enterprises and government organizations.

1.2. Potential problems and threats of digital penetration

Data leakage for artificial intelligence is especially dangerous due to the fact that big data usually carries a lot of confidential information from which you can get information about the object that was attacked. Simultaneously, data leakage can occur at any stage of development: training or using a ready-made model.

Violation of confidentiality is another important detail since it is personal information that acts as a catalyst for any digital transformation, becoming the basis for learning models. The process of developing many neural networks is practically inseparable from relying on the collected data, including speech, Internet activity, images, financial flows, medical indicators, etc. Thus, the issue of access to big data turns out to be one of the most significant problems associated with the integration of neural networks into society [13]. Any third party will be aware of the potential risks of using information collected by technology owners regarding user information, their right to privacy and the protection of their personal data [17, 18]. This aspect is widely discussed within society, as there is a direct threat to human health and life behind it. The relevance of this problem has been raised by prominent scientists and statesmen many times. Also, related reports and studies have been repeatedly presented in international discussions (IEEE, EU Committees, OECD, etc.) [6]. The core idea is that AI has only those "ethical values" that have been defined by the developer.

By publicly guaranteeing transparency, full audit and objectivity through the introduction of neural networks, technology owners in the era of digitalization are able to perform any manipulation. This possibility arises directly from the lack of understanding between society users of digital technologies of the structure and the principles of algorithms. Complex AI models perform colossal calculations which cannot be fully understood even by the creators [5, 7]. Thus, for most users, the process of neural networks will be opaque [15]. Scientists refer to it as a "black box" problem. Taking advantage of this phenomenon, unscrupulous developers can use neural networks at their discretion.

Another problem with training data is its low availability. Often, small amounts of data belong to different persons who have no reason to trust each other or the developer, and it is impossible to compile one dataset of sufficient size. If a dataset of the required volume exists, it may still be unavailable if the data contained in it is confidential. Even if suitable data can be found, it is necessary to ensure that they reflect the real state of affairs. In particular, they should not contain hidden biases, as, for example, happened to some facial recognition systems: due to the imbalance of data, they, for example, coped much better with the recognition of light-skinned men than dark-skinned women.

The data that a ready-made AI system works with may also be of poor quality: they may come from unreliable sources or contain information with a high degree of uncertainty. In addition, the databases that the system interacts with may be at risk if the system itself is hacked. For example, a biometric data verification system may be subject to several types of attacks in order to force it to accept an attacker as the owner. Noise is added to the processed data so that the already trained model identifies the object in the photo in the wrong class. Such attacks can be used in computer vision, for example, forcing the model to incorrectly identify road signs.

The bias of the model can have an extremely negative impact on the results of using digital technologies. Among the possible causes of bias are the uneven distribution of data in the training sample, algorithmically embedded preferences and a biased attitude toward individual groups of individuals. Even the classic spectrum of potential crisis cases is huge, starting with ageism and sexism when hiring, and ending with racism when identifying potential criminals [7].

Moreover, it is unacceptable to exclude situations where the bias of the model is a meaningful policy of its creators dictated by their interests, which is even more dangerous. Without proper control, developers can gain serious power over society. Simultaneously, it is difficult to overestimate the degree of destructiveness of the consequences of discrimination: systematic violations of the rights of certain social groups will lead to the definition of AI as an inhumane mechanism.

To avoid problems that can harm a person by distorting, stealing, or leaking data, it is necessary to ensure that the results of AI work can be trusted. Thus, there is a need for the concept of "trusted artificial intelligence": an AI system in respect of which the user can be sure that it is capable of performing the tasks qualitatively [19].

2. Results and discussion: Implementation of trusted artificial intelligence

It is worth admitting that the introduction of artificial intelligence seems to be a very profitable process. However, it is trust that is the key factor in the use of AI, since the rejection of this technology by the masses will exclude all potential benefits [5, 8, 13]. Consequently, the establishment of rational, trusting relationships, excluding excessive trust or its absence, will make it possible to achieve benefits for all of society. The essence of this idea is reflected in the concept of trusted AI.

Trust in AI at the physical level means confidence in the correct operation of all its physical components, such as sensors, and in the quality of the data received by the system. Trust in the infrastructure surrounding AI means confidence in the security of the data with which AI interacts, and in control over access to the system itself. Trust at the application level means confidence in the correct operation of the software.

If an AI system is considered trusted, that is, trust in it is manifested at all three levels, then such a system can be allowed to solve problems with a hugely positive outcome, since the user can be confident in the results of its work.

The trust issue is quite complex. It directly depends on various features of the human psyche [13, 20, 21]. Nevertheless, researchers emphasize that trust is the desire and willingness of an individual to depend on the other party's actions to extract some benefit, despite the potential risks from being in a vulnerable position. This phenomenon is based on the coincidence of the moral values of the parties, which allows a person to predict the further actions of the party subject to trust, and confidence in its sufficient competence [20]. Additionally, the cumulative nature of trust makes development of relationships dependent on the first experience. Therefore, strategies for building a trusted AI should be developed before its mass introduction in strict accordance with ethics and robustness.

Next, we will present several key approaches that appear to be the most important for implementing trusted artificial intelligence for each stakeholder.

A. International community

The first stage is to develop a legislative framework regulating the work of AI in each industry. Already various international committees and organizations are busy drafting general recommendations and guidelines for action [21]. While some attempts to

control the technological agenda were made earlier [22], the OECD document "Principles on Artificial Intelligence" is considered the reference point in the discussion around ethics and competence of AI [23, 24]. Having underscored the initial outlines and core principles, the authors provided the basis for more comprehensive documents in the future [23, 25, 26]. Uniting the most vital criteria of trusted AI, they must encourage leading countries in this field to start elaborating their policies [20].

Key principles here include but are not limited to constant human oversight, resilience, accountability, privacy and transparency. Future conventions might also have notions about non-discrimination and fairness, perseverance of social and environmental well-being, attention to mitigating circumstances and the introduction of the so-called "right to explanation" [23, 25–27]. It is essential to underscore that while government agencies will be obliged not only to constantly adapt the legal sphere but also to expand it, international conventions must stand for the announced principles. This will let the authorities develop a balanced system of norms, delineating the areas of regulation between soft and hard laws if these principles are followed [7]. Moreover, this process has to be original. Issues of AI control need to be solved based on the existing legislative structure and not contrary to it. Authorities will give ethical reasons to trust AI if bias is successfully avoided. However, despite the cornerstone importance of future standards, they only point to the importance of compliance with the law in matters of safety and quality, while maintaining abstract rules.

The concept of trusted AI is illustrated in *Fig. 1* and includes many aspects.

B. Scientific community

First of all, it is worth focusing on the importance of scientists and developers to protect neural networks. Otherwise, the basic requirements of trusted intelligence would be undermined, and the situation would potentially shift towards citizens' digital dependence from third parties. However, the evolution of defense methods is almost synchronous with a similar process for AI attacks, which can be illustrated by the following list of the most potential ones:

1. Privacy breaches are one of the most likely issues among the expected ones. Leaks of personal data can occur at almost any point in the neural network's operation, from training to outputting results. Individuals' details are extremely lucrative to certain third parties in many matters. The focus of defense is on privacy-enhancing technologies (PET). The most notable of them is OPAL (the Open Algorithms project). It gives algorithms remote-controlled access to information instead of sending anonymized data. Consequently, only the aggregated result will be returned to the model developers. Owing to the full operations' recording, the entire learning phase is audited [28, 29].

2. Another scenario is data poisoning. The idea here is to inject false information in the training sample, either devaluing the results of the entire system or pushing the neural network towards making wrong decisions beneficial to a third party. Such technologies can both cause significant damage to the reputation of an individual and affect the behavior of the masses (for example, during an election) [20]. No specific solution has yet been proposed here. However, despite all its threatening potential, the risks here are still minimal since these technologies are at an early

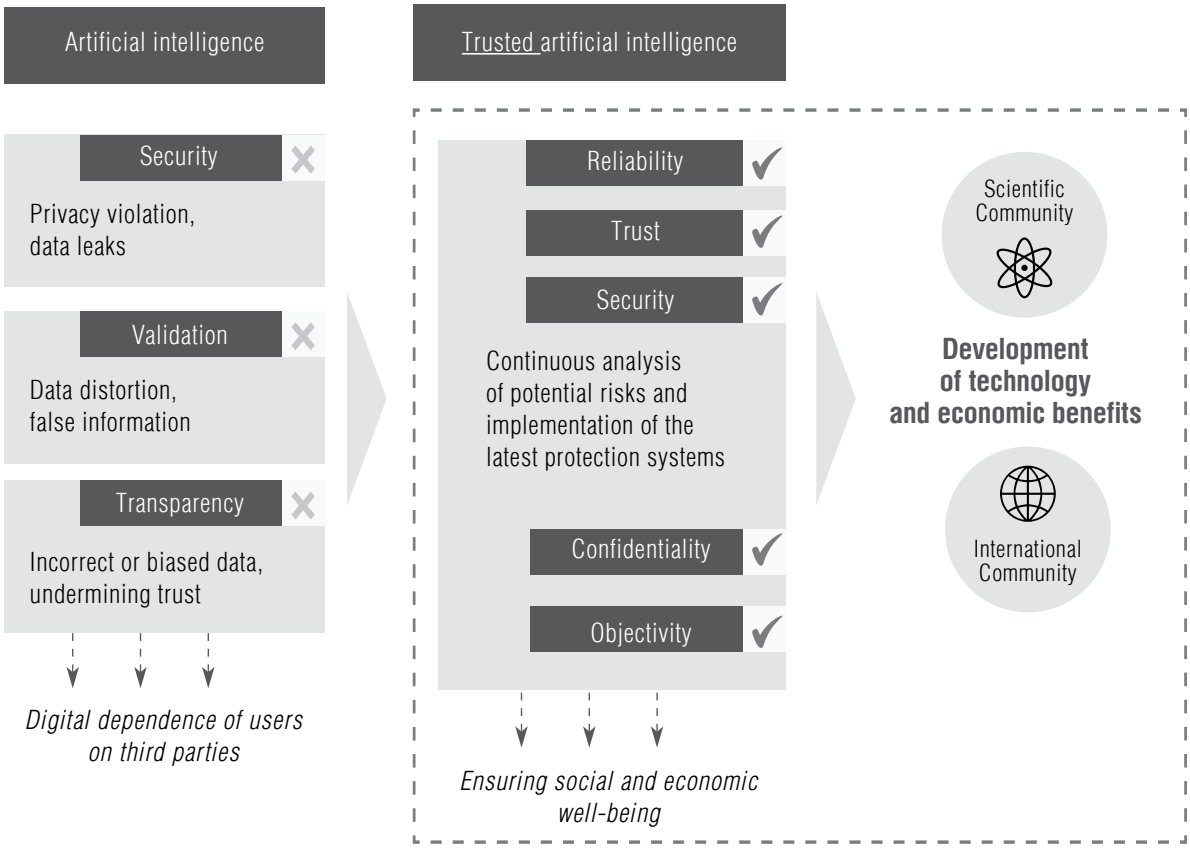

Fig. 1. The concept of using trusted AI.

development stage which does not imply any substantial threats. There are already several models in the literature for combating data poisoning, almost all of them focused on human control and protection.

3. The so-called evasion attack also causes serious concerns. Unlike data poisoning, which is used during the training phase, this threat appears at the stage of applying AI. It is possible to get a radically new network response by imperceptibly modifying the input values. Even the slightest transformations can lead to sufficient consequences [7]. There are several mechanisms for countering it, but the most widely described is adversarial training, which implies that developers include intentionally incorrect data during training, bolstering the model to ignore potential noise in the future [17, 30].

4. The intellectual value of an already trained neural network is obvious, especially when using big data collected by the government. Model extraction can potentially mean large leaks of personal data. Such a violation of confidentiality can lead to unpredictably catastrophic outcomes (for example, medical records) [7]. Since model inversion mainly exists in scientific articles and abstract models, methods of countering it are still theoretical. The private aggregation of teaching ensembles (PATE) is among the most recognized ones. The concept's idea is to separate data into several sets, each training a separate neural network. Then these independent models, called

"teachers," are combined to train the neural network named "student" by voting, not giving the latter access to the original data [18, 31].

5. Several metrics allow detection of model biases with sufficient efficiency, including equal opportunity, disparate impact, difference in means and normalized mutual information. Developers might discover the imperfections of neural networks and make appropriate alterations using these methods. In turn, the revealed bias can be mitigated by either pre-processing, in-processing or post-processing algorithms [32].

Lastly, an important aspect of the united scientific community's policy is to present specific technologies that would allow the asserted principles to be implemented more effectively. Even now, the gap between the algorithms employed and the abstract principles is obvious. Therefore, in-depth research is constantly taking place, exploring possible scenarios and tools.

Several approaches have already been presented to counter the above issues, one of which encourages us to scrutinize blockchain, a continuous chain of blocks connected in reverse order through hash sums. Each block, in addition to its hash and the hash of the previous block, contains some information. Thus, the blockchain is called a distributed public registry providing data storage and transmission with high robustness and almost zero chance of interference. Smart contracts (the program code ensuring the fulfillment of all the established rules) in conjunction with AI will guarantee the reliability of the final results. Therefore, combining blockchain technology with AI will create a decentralized system that maintains information of any value and provides it for neural network training. As a result, not only is data security guaranteed, but ethical concerns are also addressed owing to a comprehensive history of operations that rules out external interference or pre-programmed bias [24, 33]

Conclusion

The world community has to prepare itself for the up-coming challenges on the inevitable road to a digital society. The latter, indeed, is becoming a crucial stage in improving both governments' operations and human life. The ubiquitous integration of digital technologies and the creation of decentralized ecosystems can open new horizons, eliminating a number of current issues. The key role of neural networks in this process forces us to pay special attention to each potential menace. The establishment and monitoring of trusted AI principles will therefore enable both the scientific community and international organizations to create and regularly update mechanisms to counter risks. Moreover, humanity as a whole will have a chance not to miss out on all the political experience accumulated over countless centuries.

However, the presented technological strategies are nothing more than a theoretical speculation on the topic of future processes. That is why it is vital to constantly adapt strategies to changing realities, to expand the legislative framework and to look for new solutions. At the same time, a strong collective commitment to maintaining democratic institutions is imperative.

Nevertheless, AI is already present in our lives. Although the extent of its adoption is relatively modest, its prospects are truly breathtaking. If we meet the upcoming challenges with dignity, the gains listed above will elevate humanity to a completely new level of existence. ■

References

1. Vogt T., Winter P., Nessler B., Doms T. (2021) *Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications*. Vienna: TÜV Austria Holding AG.
2. Kuleshov A., Ignatiev A., Abramova A., Marshalko G. (2020) Addressing AI ethics through codification. *2020 International Conference Engineering Technologies and Computer Science (EnT)*, pp. 24–30. https://doi.org/10.1109/EnT48576.2020.00011
3. Harrison T., Luna-Reyes L. (2021) Cultivating trustworthy artificial intelligence in digital government. *Social Science Computer Review*, vol. 40, no. 2, pp. 494–511. https://doi.org/10.1177/0894439320980122
4. OECD (2014) *Recommendation of the council on digital government strategies*. Paris: OECD Publishing. Available at: https://www.oecd.org/gov/digital-government/Recommendation-digital-government-strategies.pdf (accessed 22 September 2021).
5. Chakraborty A., Alam M., Dey V., Chattopadhyay A., Mukhopadhyay D. (2018) *Adversarial attacks and defences: A survey*. Working paper arXiv: 1810.00069. https://doi.org/10.48550/arXiv.1810.00069
6. Haenlein M., Kaplan A. (2019) A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, vol. 61, no. 4, pp. 5–14. https://doi.org/10.1177/0008125619864925
7. Wei J. (2018) Research progress and application of computer artificial intelligence technology. *MATEC Web of Conferences*, vol. 176, Article Number 01043. https://doi.org/10.1051/matecconf/201817601043
8. Mijwil M., Esen A., Alsaadi A. (2019) *Overview of neural networks*. Available at: https://www.researchgate.net/publication/332655457_Overview_of_Neural_Networks (accessed 22 September 2021).
9. Sibai F.N. (2020) AI crimes: A classification. *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, pp. 1–8. https://doi.org/10.1109/CyberSecurity49315.2020.9138891
10. Jastroch N. (2020) Trusted artificial intelligence: On the use of private data. In: *Product Lifecycle Management Enabling Smart X. PLM 2020* (eds. F. Nyffenegger, J. Ríos, L. Rivest, A. Bouras). IFIP Advances in Information and Communication Technology, vol. 594. https://doi.org/10.1007/978-3-030-62807-9_52
11. Nemitz P. (2018) Constitutional democracy and technology in the age of artificial intelligence. *Phil. Trans. R. Soc. A.*, vol. 376. http://doi.org/10.1098/rsta.2018.0089
12. Misra S.K., Das S., Gupta S., Sharma S.K. (2020) Public Policy and Regulatory Challenges of Artificial Intelligence (AI). In: *Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation. TDIT 2020* (eds. S.K. Sharma, Y.K. Dwivedi, B. Metri, N.P. Rana). IFIP Advances in Information and Communication Technology, vol. 617. https://doi.org/10.1007/978-3-030-64849-7_10
13. Pavlutenkova M. (2019) Electronic government vs digital government in context of digital transformation. *Monitoring of Public Opinion: Economic and Social Changes Journal*, no. 5, pp. 120–135 (in Russian). https://doi.org/10.14515/monitoring.2019.5.07
14. Kochetkov A.P., Vasilenko I.A., Volodenkov S.V., Gadzhiev K.S., Kovalenko V.I., Soloviev A.I., Kirsanova E.G. (2021) Political Project for Russia: Prospects for implementation in the context of challenges and risks of digitalization of society. *Vlast' (The Authority)*, vol. 29, no. 1, pp. 317–331 (in Russian). https://doi.org/10.31171/vlast.v29i1.7963
15. Williams M., Valayer C. (2018) *Digital government benchmark – study on digital government transformation*. European Union. Available at: https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/document/report-digital-government-benchmark-study-digital-government-transformation (accessed 22 September 2021).

16. Carter D. (2020) Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role? *Business Information Review*, vol. 37, no. 2. pp. 60–68. https://doi.org/10.1177/0266382120923962

17. Kamat G. (2014) *Algorithms for private data analysis. Lecture 14 – Private ML and stats: Modern ML*. Available at: http://www.gautamkamath.com/CS860notes/lec14.pdf (accessed 22 September 2021).

18. Hinnefeld J., Cooman P., Mammo N., Deese R. (2018) *Evaluating fairness metrics in the presence of dataset bias*. Working paper arXiv: 1809.09245. https://doi.org/10.48550/arXiv.1809.09245

19. National Standard of the Russian Federation (2021) *Artificial intelligence systems. Methods for ensuring trust. General. GOST R 59276-2020* (in Russian).

20. Tinholt D., Carrara W., Linden N. (2017) *Unleashing the potential of Artificial Intelligence in the Public Sector*. Capgemini. Available at: https://www.capgemini.com/consulting/wp-content/uploads/sites/30/2017/10/ai-in-public-sector.pdf (accessed 22 September 2021).

21. Sharma G.D., Yadav A., Chopra R. (2020) Artificial intelligence and effective governance: A review, critique and research agenda. *Sustainable Futures*, vol. 2, Article ID 100004. https://doi.org/10.1016/j.sftr.2019.100004

22. McCormick T.R., Min D. (2020) *Principles of Bioethics*. University of Washington. Available at: https://depts.washington.edu/bhdept/ethics-medicine/bioethics-topics/articles/principles-bioethics (accessed 22 September 2021).

23. Lindgren I., Veenstra A.F. (2018) Digital government transformation: a case illustrating public e-service development as part of public sector transformation. Proceedings of the *19th Annual International Conference on Digital Government Research: Governance in the Data Age, Delft, The Netherlands*, pp. 1–6.

24. IEEE (2021) *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent System*s, *Version 2*. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Available at: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf (accessed 22 September 2021).

25. Bradul N.V., Lebezova E.M. (2020) Conceptualization of Smart Government: A scientometric approach. *Upravlenets (The Manager)*, vol. 11, no. 3, pp. 33–45. https://doi.org/10.29141/2218-5003-2020-11-3-3 (in Russian).

26. Thiebes S., Lins S., Sunyaev A. (2021) Trustworthy artificial intelligence. *Electron Markets*, vol. 31, pp. 447–464. https://doi.org/10.1007/s12525-020-00441-4

27. Falco G., Viswanathan A., Caldera C., Shrobe H. (2018) A master attack methodology for an AI-based automated attack planner for smart cities. *IEEE Access*, vol. 6, pp. 48360-48373. https://doi.org/ 10.1109/ACCESS.2018.2867556.

28. Bellamy R. et al. (2018) *AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*. Working paper arXiv: 1810.01943. https://doi.org/10.48550/arXiv.1810.01943

29. Sarpatwar K. et al. (2019) Towards enabling trusted artificial intelligence via Blockchain. In: Calo, S., Bertino, E., Verma, D. (eds) *Policy-Based Autonomic Data Governance. Lecture Notes in Computer Science*, vol. 11550, pp. 137–153. Springer, Cham. https://doi.org/10.1007/978-3-030-17277-0_8

30. Montjoye Y.D., Farzanehfar A., Hendrickx J., Rocher L. (2017) Solving artificial intelligence's privacy problem. *Field Actions Science Reports*, Special Issue 17, pp. 80–83. Available at: https://journals.openedition.org/factsreports/pdf/4494 (accessed 22 September 2021).

31. DataCollaboratives.org (2021) *Open Algorithms (OPAL) Project*. Available at: https://datacollaboratives.org/cases/open-algorithms-opal-project.html (accessed 22 September 2021).

32. Salah K., Rehman M.H.U., Nizamuddin N., Al-Fuqaha A. (2019) Blockchain for AI: Review and open research challenges. *IEEE Access*, 2019, vol. 7, pp. 10127–10149. https://doi.org/10.1109/ACCESS.2018.2890507

33. Baker-Brunnbauer J. (2021) Management perspective of ethics in artificial intelligence. *AI and Ethics*, vol. 1, pp. 173–181. https://doi.org/10.1007/s43681-020-00022-3

About the authors

Sergey M. Avdoshin

Cand. Sci. (Tech.);

Professor, School of Computer Engineering, HSE Tikhonov Moscow Institute of Electronics and Mathematics (MIEM HSE), National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: savdoshin@hse.ru

ORCID: 0000-0001-8473-8077

Elena Yu. Pesotskaya

Cand. Sci. (Econ.);

Associate Professor, School of Software Engineering, Faculty of Computer Science, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: epesotskaya@hse.ru

ORCID: 0000-0003-2129-4645

DOI: 10.17323/2587-814X.2022.2.74.88

On assigning service life for technical systems under inflation

Sergey A. Smolyak
E-mail: smolyak1@yandex.ru

Central Economics and Mathematics Institute, Russian Academy of Science
Address: 47, Nakhimovsky Prospect, Moscow 117418, Russia

Abstract

A technical system is used by an enterprise that is a typical market participant to perform specific work. During operation, the operating characteristics of the system deteriorate. In case of a possible failure of the system, it is decommissioned and this causes losses for the enterprise. It turns out that it is beneficial to assign a certain service life to the system after which (if no failure has occurred) it is subject to decommissioning. We are solving the problem of optimizing this assigned service life. Usually, when solving it, inflation is not taken into account, and the optimality criteria are the average costs per unit of time and other indicators that do not fully reflect the commercial interests of the enterprise owning the system. Using the principles and methods of valuation, we build a mathematical model and propose formulas that allow us, taking into account inflation, to find the optimal assigned service life of the system and at the same time estimate the market value of the work performed by the system and calculate the change in the market value of the system with age. Moreover, in this problem, the optimality criterion is the ratio of the expected discounted costs to the expected discounted volume of work performed by the system. We show that such a criterion maximizes the market value of the enterprise owning the system. We give examples of using the constructed model. The results obtained can be used both for solving other optimization problems of the reliability theory and for practical valuation of some types of machinery and equipment.

Keywords: technical system reliability, failure, assigned service life, optimization criterion, valuation theory, market value

Citation: Smolyak S.A. (2022) On assigning service life for technical systems under inflation. *Business Informatics*, vol. 16, no. 2, pp. 74–88. DOI: 10.17323/2587-814X.2022.2.74.88

Introduction

The establishment of the service life of machinery and equipment has been, at least since the 20th century, a topical task for both economists and technical specialists. A review of the relevant literature would require too much space, so we do not provide it. Let us just mention that back in the 1920s and 1930s, a number of articles were published on the economic aspects of optimizing the service life of machines that have not lost their relevance today. However, in their research, the economists usually ignored the stochastic nature of the machine operation process, while technical specialists focused on the issues of their reliability and repair, ignoring the impact of physical deterioration on their performance and (sometimes) on operating costs. Economists considered the purchase and utilization of machines as an investment project and accordingly relied on the theory of evaluating the efficiency of investment projects. At the same time, technical specialists, when setting the assigned service life of technical systems, chose the optimality criterion, often arbitrarily, without proper justification (for more details, see the next section). It is significant that almost all known works on this problem did not take into account the influence of inflation. Meanwhile, in the course of practical valuation of real assets (buildings, construction facilities, machinery and equipment), appraisers take into account not only inflation, but also physical deterioration and reliability of the assets being evaluated, and the appropriate valuation standards include the general principles of such assessment.

In this regard, it seems essential to combine technical and economic aspects within a single optimization model, taking into account both reliability and physical deterioration of technical systems, and relying (unlike existing economic and mathematical models) on the theory of valuation. So, the purpose of this article is to develop a model regarding the tasks of assigning service life to technical systems that are subject to stochastic failures.

1. Statement of the problem and main concepts

The objects of the study in this article are technical systems, usually machines and equipment that:

- are used by enterprises participating in the market;
- are subject to stochastic failures;
- are unrepairable, i.e. are not subject to major or medium repairs (in the case of failure, such systems are decommissioned and disposed of).

The subjects of our study are assigned service lives and market values of technical systems of different ages. We present the information related with valuation of technical systems and their reliability in accordance with valuation standards [1] and the book [2]. The main operational characteristics of a technical system (depending on its technical conditions), are:

- operational productivity (the volume of work performed by the system in a small unit of time);
- operating costs rate (operating costs related with use of the system for its intended purpose for a small unit of time);
- hazard rate (probability of failure of the system within a small unit of time).

The reliability theory considers the task of establishing the assigned service life of an unrepairable technical system. We will deal with the justification of optimality criterion of such service life. For this reason, we limit our considerations to the following situation, which is quite simple.

An enterprise acquires on the market and uses an unrepairable technical system of a cer-

tain type (for example, a machine of a certain model). During the operation, the system is degraded and its operational characteristics deteriorate, despite the ongoing maintenance (we include the costs of such operations into overall operating costs). This provision needs some comments.

Many specialists, as in the times of the "plan oriented economy," believe that such characteristics of machines and equipment as their productivity and annual operating costs must remain unchanged throughout the entire service life at the level provided for by the project. The permanent performance of the technical system is also accepted in existing publications on the problems of optimizing the assigned service life and the timing of scheduled repairs of the systems. However, an analysis of the operation of machines and equipment used for various purposes shows that with increasing age their characteristics tend to deteriorate, and overhauls compensate such deterioration only partially (for more details, see [3]). Researchers often try to explain this point by saying that only the cost and duration of maintenance grow with increasing age. Nevertheless, the data available indicate that for construction machines their productivity per hour decreases with age, fuel costs per 1 km of vehicle mileage increase and duration of the processing cycle of machine tools increases.

For this reason, it seems unreasonable to consider the operational characteristics of technical systems as permanent over time. Further, we assume that the technical condition of the system is determined by its age (operating time), so its operational characteristics may be presented as certain functions of age.

Since technical systems are freely traded on the open market and have some usefulness for enterprises participating in the market, they become objects of valuation. There are many bases of value, but the main one is the market value. We will not give a detailed definition of this concept, referring to the valuation standards [1]. However, it should be noted that the market value of an object of valuation as of a certain date (valuation date) reflects the price of this object in a transaction (real or hypothetical) made on this date between independent and economically rational typical market participants. At the same time, the market value of the object reflects both its usefulness for typical market participants and the contribution of the object to the market value of an enterprise (a typical market participant) that possesses this object.

For valuation of objects, three approaches are used (individually or in various combinations):

- using a *comparative (market) approach*, the value of an object is determined relying on the prices of transactions with similar objects, adjusted for differences in the characteristics of objects, date and conditions of the transaction;
- with the *cost approach*, the value of the object is estimated by the costs incurred during creating or receiving this object;
- with the *income approach*, the value of the object is found, taking into account the flow of benefits that the owner will receive from using the object.

Since a technical system is a plant and equipment, its market value may differ depending on whether it is evaluated "in place" or "for removal." New systems are usually sold by their manufacturers or dealers in the primary market and are similar to each other. While the price spread is small, the market value of a new object "for moving to the place of operation" may be determined using a comparative approach, which is not difficult in practice. The market value of the same technical system "at the place of its operation" is higher, since it includes additional costs for the transportation of the purchased object and its installation. We consider the market value of a new object "in place" equal to a known value K.

Used technical systems are sold on the secondary market. Such objects do not have exact analogues. Buyers do not have the opportunity to assess their technical condition, and sellers do not inform buyers (or even do not know) the history of their operation. As a result, the prices of machines or equipment of the same type and the same age have a very large dispersion, so that valuation of used objects presents significant difficulties for appraisers.

As an analogue of a used technical system, it is reasonable to consider a similar new system (we will call it a "new analogue"). The market value of a new analogue of the system is called the reproduction value of the object being valued. A decrease in the market value of a used system compared with its reproduction value is called depreciation of the appraised object, and the ratio of these values, expressed in fractions of a unit or as a percentage, is the coefficient/percentage of goodness or relative value (Percent Good Factor, PGF). In the valuation related literature, there are a number of formulas and tables describing (not always correctly) the dependence of the PGF on the age of the system [3].

A system that can no longer be used for its intended purpose or is ineffective is to be decommissioned and disposed of. The market value of such a system is called scrap value. This value is usually determined relying on the information about prices of machines sold "for scrap" or calculated as the cost of separate elements of the machine that are suitable for further use (including scrap metal), less costs of dismantling and delivery of such elements to the place of their further use. The scrap value of machinery and equipment is small in comparison with their reproduction value – the ratio of these values (relative scrap value) for machinery and equipment is usually from 0.03 to 0.09.

Methods of valuation of machinery and equipment are described in small sections of valuation standards and textbooks. As a rule, these issued take into account neither the deterioration of the operational characteristics of systems with increasing age, nor the probabilistic nature of their service life. The exception is the methods used in national accounting systems [4], but they are focused on assessing large groups of assets and inadequately take into account the degradation of machinery and equipment [5]. At the same time, it is possible to take into account these factors adequately relying on basic principles of valuation and valuation practice by using models and methods of the reliability theory.

One of the main characteristics of technical systems' reliability is the hazard rate. We will consider it as dependent on age of the technical system (t) and denote it by $p(t)$. This value has a simple "physical meaning": if the system has worked without failure for t years, then the probability of its failure in the time interval (t, $t + dt$) is $p(t)dt$. In this case, the (random) moment of failure has a distribution with the density $p(t)e^{-P(t)}$, where $P(t) = \int_0^t p(x)dx$. As is known (see, for example, [2, 6]), the probability of failure–free operation of a technical system during the period t is $e^{-P(t)}$, and the average uptime (average time to failure), unless it is specifically limited, is – $\int_0^\infty e^{-P(t)}dt$.

It is essential that the failure of a technical system leads to the failure of not only the system itself, but also of other assets of the enterprise. At the same time, the damage to the enterprise is, generally speaking, stochastic. Methods for determining such damages (losses) are described in the relevant literature, for example [7]. We will assume that the average value of the damage is known, and its value takes into account the cost of the elements of the failed system and other assets of the enterprise that are suitable for further use.

In order to reduce losses because of the failure of the system, the system is assigned a certain service life (operating time) S, upon reaching which the object must be decommissioned regardless of its technical conditions. We want to find the optimal value of the period S. Selecting S also determines the average remaining service life (average uptime) of objects of different ages.

Let us consider a technical system that has survived to the age of t. Let $T(t)$ be the average remaining life of its operating. To find it, we note that the remaining service life of the object will be $x-t$ if it fails at the age of $x < S$, and will be equal to $S-t$ if no failure occurs within the designated period. Considering that the probability of failure-free operation of an object that has survived to the age of t, during the entire designated period is equal to $e^{P(S)-P(t)}$, and failure during the time interval $(t, t+dt)$, $t < S$, is possible with probability $p(t)e^{P(S)-P(t)}dt$, we get:

$$T(t)=\int_t^S (x-t)p(x)e^{P(t)-P(x)}dx+(S-t)e^{P(t)-P(S)}= \\ =\int_t^S e^{P(t)-P(x)}dx. \qquad (1)$$

For many types of technical systems, it is assumed that the uptime has the Weibull distribution, in which the hazard rate $p(t)$ increases proportionally to some degree of age. The values of the degree index for various types of objects were calculated by a number of authors, including for the purposes of national accounting, and the recommended values for a number of engineering facilities are given in [6]. It should be noted that for some types of machines (freezers, refrigerators, vacuum cleaners, microwave ovens, video recorders, washing machines, electric heating appliances, small cars, equipment for car repair and maintenance, railway wagons), this indicator turns out to be close to one [8–10]. Later, as an example, we will consider exactly such objects. Their uptime has a Rayleigh distribution, $p(t)=t/\omega^2$, where ω is the scale parameter. In this case, the average remaining service life of an object of age t will be

$$T(t)=e^{t^2/2\omega^2}\int_t^S e^{-x^2/2\omega^2}dx= \\ =\omega\sqrt{2\pi}\left[\Phi\left(\frac{S}{\omega}\right)-\Phi\left(\frac{t}{\omega}\right)\right]e^{t^2/2\omega^2}.$$

At $t = 0$ the formula for the average full service life of the technical system is:

$$T_m=T(0)=\omega\sqrt{2\pi}\left[\Phi\left(\frac{S}{\omega}\right)-\frac{1}{2}\right]. \qquad (2)$$

At $S=\infty$, when the service life is not assigned, the average uptime of the system will be equal to $\omega\sqrt{2\pi}$.

In the reliability theory, various criteria are usually used to solve optimization problems. The examples are: the average number or cost of repairs over the service life, average repair costs per unit of time [11], average costs per unit of time for the inter-repair cycle [12], the ratio of average lifetime costs to average service life [13–15], total discounted costs or an annuity equivalent to them [16, 17]. These criteria have a number of common disadvantages.

1. Selection of the criterion is of a formal mathematical nature and is focused on the use of the measured characteristics of the system (primarily, related with cost and time). The interests of a particular business using a technical system are not taken into account [12].

2. Comparing options of applying technical systems using costs is correct only if they give identical results [18]. However, if the performance of a system changes with age, then options with various service lives assigned will have differences in the value and variances of results obtained over time, i.e. volumes of work performed.

3. Calculating indicators such as total or average lifetime costs, does not take into account the impact of inflation. This could be justified if we were talking about systems with a short (up to 1–2 years) service life. However, in con-

struction, engineering and transport industries these terms exceed 5–15 years, and in such cases, consideration of inflation is essential.

Taking into account the points mentioned above, we consider that when solving the task it is more appropriate to apply general principles of valuation focused on maximization of the market value of enterprises.

At the same time, unlike works on reliability theory, it turns out to be possible to take into account not only inflation, but also the impact of technical systems' degradation on their performance[1] and operating costs, as well as the scrap value of the systems. Taking into account inflation matters is related with special considerations which the next section will be devoted to.

2. Group inflation

Selecting economically rational solutions for managing technical systems under conditions of inflation is associated with significant difficulties. The fact is that this requires a forecast of the economic characteristics of the system for its entire service life, which implies, at a minimum, forecasting prices for products produced by the systems (goods, works or services), as well as for various resources consumed. Unfortunately, specialists in the management of technical systems are not able to develop such forecasts. That is why in mathematical models of optimization of technical solutions focused on practical application, inflation is usually not taken into account. This also applies to the task of assigning the service life for a technical system. Meanwhile, it is possible to find out how inflation affects the solution of this problem, relying on the theory of valuation and the practice of valuation activities. It is proposed to take into account only the most important and measurable characteristics of inflation, neglecting all the others.

Analysis of the prices in primary and secondary markets, as well as the experience of evaluating used machines and equipment show that the prices of used machines usually change synchronously with the prices of similar new ones. This is due to the fact that the "economically rational" buyer of a used machine (on the secondary market) compares the planned purchase with the alternative of purchasing its new analogue on the primary market.

In this regard, it seems natural to assume that in conditions when the market values of new technical systems of some kind are growing, the market value of used systems of this type are growing in the same proportion. This phenomenon is called group inflation [3]. However, it is not easy to give a strict definition of group inflation, since all used systems of the same type, unlike new ones, have no analogues, are in different technical condition, their quantity and composition on the secondary market are constantly changing. Therefore, the concept of growth of systems values on the secondary market becomes uncertain.

Nevertheless, the essence of group inflation can be explained by a conditional example relating to a group of machines of the same model, the technical condition of which changes over time. Let us imagine a market in which at some time T_1 there is a set of C_1 of M new and used machines of this group, which are in different states, and the costs of all these machines are known. Now let us assume that at a later moment T_2, a set of C_2 of M machines of the group also appeared on the market, and each i-th of them is in exactly the same technical condition as the i-th machine from the set of C_1. In other words, in this situation, all the

[1] In the works on the problem of optimizing the assigned service life and the timing of scheduled repairs of technical systems, their performance is usually considered as unchanged.

machines of the group available on the market at the time T_1, as if "moved" to a later date without changing their technical conditions. At the same time, due to inflation, the market values of all the machines will change. However, under group inflation, they will change proportionally, and it turns out that the ratio of the value of a used machine to the cost of the same new one (the PGF) will depend only on the condition of this machine, but not on whether it is valued at time T_1 or at time T_2. This statement will be the basis for the definition of the concept of group inflation. We will say that for a certain type of a technical system, group inflation takes place in a certain time interval if the PGFs of these machines depend only on the condition of the systems, but not on what date (in the specified time interval) their value was estimated. The rate of group inflation is defined as the growth rate of the market value of new machines. Therefore, for the analysis and short-term forecast of this rate, it is sufficient to use open and accessible information about the prices of manufacturers or dealers in the period close to the valuation date.

Group inflation assumption significantly simplifies the solution of the problem of optimizing the assigned service life of a technical system, as we will see in the next section. However, the validity of this assumption is also confirmed by other arguments.

Usually, the technical condition of used machines sold is unknown to appraisers[2]; they only know their age. Therefore, they have to characterize the condition of the machine by its age (as in this article). Then, for an approximate calculation of PGF, they form a sample of the market prices of machines of different ages in a certain base period (for example, in the current year) and build a regression dependence $F(t)$ of machines' prices on their age (t). At the same time, the value of $F(t)$ will reflect on this basis the average market value of machines of age t that were (or could be[3]) presented on the market in the base period, and $F(0)$ – the market value of new machine in this period. The goodness factors are found by the formula: $k(t) = F(t)/F(0)$. In the absence of group inflation, similar dependencies constructed for machines of the same model according to data from different years could differ significantly (this would mean, for example, that under the influence of inflation, the prices of older machines decrease or grow more slowly than the prices of younger ones). However, this phenomenon is usually not observed.

On the contrary, under conditions of group inflation, the function $k(t)$ does not depend on which period the prices of machines belong to. Particularly, Russian appraisers usually use the dependence of PGF of machines on the age, built on the market data of previous years by other authors. By doing so, they are essentially also assuming group inflation assumption. Approximately the same situation appeared in some US states where machinery and equipment are subject to taxation. There, regression dependences of PGF on age are built annually, though for quite wide groups of machines (for example, agricultural or construction machinery and equipment). An analysis of relevant publications (for example, [19–21]) shows that these dependencies change slightly from year to year, which also indicates group inflation.

[2] It would seem that quantitative assessments of the technical condition of machines and equipment are provided by automated diagnostic systems. However, these systems do not assess either the hazard rate of the machine or other important characteristics for market participants (for example, the condition of the frame of a truck or the body of a passenger car significantly affecting their market value).

[3] This stipulation is related with the fact that the constructed dependence allows us to calculate $F(t)$ for machines of any age t, whereas in the base period machines of some ages were not available on the market.

It is important to note that the assumption of group inflation allows us, when constructing the dependence $k(t)$, to include in the sample the prices of machines that have developed in different years (while simultaneously calculating the market value of new machines in these years). This makes it possible to significantly increase the sample size and improve the accuracy in calculating dependence $k(t)$.

Of course, in other situations, group inflation can be understood in a different way, using the operating time or other objectively measurable characteristics instead of age. In such cases, the following model will need to be adjusted accordingly.

3. Optimization model

To optimize the assigned service life for a technical system belonging to some type, it is necessary to set a certain optimality criterion. For this purpose, we will solve another problem that, at first glance, is quite far from the theory of reliability: we will evaluate the market values of technical systems of different ages "in place." As the valuation date, we will take the moment of making a decision on the assignment of the service life for a system. In addition, we assume that for an object of this type for a short period (close to the valuation date) there is a group inflation with a known rate i.

We will also consider as known the following characteristics of the system: market value of a new object K, the scrap value U, the average (expected) damage caused from the failure of the system L, hazard rate $p(t)$ of an operable system of age t, its operational performance $Q(t)$ and the intensity $C(t)$ of its operating costs[4].

The assigned service life of the system to be optimized is denoted by S, while market value of an operable system of age t "on the spot" is denoted by $V(t)$.

Note that the work performed by the system has some utility for market participants and is measured in certain physical units (cubic meters of displaced soil, decaliters of spilled liquid, the number of conventional cans of canned food, etc.). If so, then, according to the valuation standards [1], it also has its market value. Another thing is that the owners of machines usually do not know it, and professional appraisers rarely evaluate the market value of work (except, perhaps, construction and repair works). The unknown market value of a unit of work performed by the system at the valuation date is denoted by B.

In general, the benefits from the use of an object for a period are defined as market value of works performed by the system during this period, minus costs incurred. Depending on the method of measuring values and costs, the following types of benefits are distinguished:

- if values and costs are measured in prices of the corresponding period (i.e. including inflation), the benefits are called *nominal*, while if prices of a certain fixed date are used – they are *real*;
- if income tax is taken into account as a part of costs, the benefits are called *after-tax*, otherwise they are *pre-tax*.

In this paper, the benefits of using the technical system are considered as nominal and pre-tax. In this case, the intensity of benefits from using an operable system of age t (i.e. the volume of benefits the system brings in a small unit of time) will be $BQ(t) - C(t)$ at the valuation date.

[4] Strictly speaking, the productivity and operating costs of a system depend on variances in demand for the products of enterprises using the system, and, consequently, on the need for the work performed by the system. In our model, $Q(t)$ and $C(t)$ are average (for typical enterprises – owners of the systems) values of appropriate characteristics relevant to the average mode of applying operable systems of age t.

To evaluate a technical system, we will use the anticipation of benefits principle, which underlies the income approach to property valuation and is mentioned (but not disclosed) in the valuation standards. We will give its most detailed formulation following [3].

Market value of an object at the valuation date is equal to the sum of the discounted by this data benefits from its use anticipated in the subsequent forecast period (including market value of the object at the end of the period), if the object is used most effectively, and not less than it otherwise.

A number of comments should be made to this formulation.

1. The duration of the forecast period can be chosen arbitrarily, since market value of an object does not depend on this choice.

2. Inclusion of the object value at the end of the period in the total benefits for the forecast period can also be treated as the benefit from the sale of the object at market value. Thus, the sale of an object is considered as one of the ways to use it.

3. In the conditions of stochasticity, the term "expected" is treated as an expectation (in [1] – "probability-weighted"), and the risks associated with the uncertainty of benefits are not taken into account in the discount rate (we will call such a rate "risk-free").

4. The type of discount rate is determined by the type of benefits: nominal (real) benefits should be discounted at the nominal (real) rate, and after-tax (pre-tax) benefits – at the after- (pre-) tax rate[5]. Therefore, in this article, when evaluating a technical system, we use the nominal pre-tax risk-free discount rate, denoted by r.

5. Valuation standards [1] require that, when assessing any type of object value, you specify the premise regarding the way of its use. Here and further, the way of using the object is assumed to be the most effective (in valuation standards terminology – "highest and best"), so that the appropriate premise is omitted.

Turning to the definition of unknown values $V(t)$, first we note that market value of a new technical system is a known value $V(0) = K$. In addition, a system reached the designated service life S is disposed of, so that its market value is $V(S) = U$.

Let us consider a technical system that has reached the age of $t < S$. To evaluate its market value $V(t)$ at the valuation date, we apply the anticipation of benefits principle regarding this system, choosing a forecast period of infinitesimal duration dt. At the same time, we take into account that in our model all possible ways of using a system differ only in the assigned service life S.

If a failure occurs during dt, which is possible with a probability of $p(t)dt$, then the average damage for the enterprise is L, and the system will be disposed of. In the opposite case, the system will bring benefits $[BQ(t) - C(t)]dt$, and at the end of the period its age will be equal to $t + dt$. A technical system of this age at the assessment date has market value $V(t + dt)$. However, under group inflation at a rate of i during dt, it will grow by $1 + i\,dt$ times and become equal to $(1 + i\,dt)\,V(t + dt)$. Considering the probabilities of both cases, we get:

$$V(t) \geq p(t)dt \cdot (-L) + \\ + \left[1 - p(t)dt\right]\left\{\left[BQ(t) - C(t)\right]dt + \right. \\ \left. + (1 - rdt)(1 + idt)V(t + dt)\right\}, \qquad (3)$$

moreover, equality is achieved here with the most efficient use of the system, i.e. with the optimal value of S.

Let us now introduce the "real"[6] discount rate into consideration:

[5] Appraisers determine discount rates based on data on nominal pre-tax yields of financial instruments observed in the market.

[6] The term "real" is taken in quotation marks, since in formula (4) the nominal rate is reduced by the rate of group inflation, and not by the rate of general inflation in the country, as is required when calculating the proper real rate.

$$\rho = r - i. \quad (4)$$

Using it, after simple transformations of inequality (3) we find:

$$BQ(t) - C(t) - \rho U - (L+U)p(t) \le \le [\rho + p(t)][V(t) - U] - [V(t) - U]'.$$

Let us multiply this inequality by $e^{-\rho t - P(t)}$ and integrate by t from $t = s$ to $t = S$. Considering that $V(S) = U$, we get:

$$\int_s^S [BQ(t) - C(t) - \rho U - (L+U)p(t)] e^{-\rho t - P(t)} ds \le \le [V(s) - U] e^{-\rho s - P(s)}, \quad (5)$$

or:

$$V(s) \ge U + BQ_\Sigma(s,S) - C_\Sigma(s,S), \quad (6)$$

where:

$$\begin{cases} Q_\Sigma(s,S) = \int_s^S Q(t) e^{-\rho(t-s) - P(t) + P(s)} dt, \\ C_\Sigma(s,S) = \int_s^S [C(t) + \rho U + (L+U) \cdot \\ \quad \cdot p(t)] e^{-\rho(t-s) - P(t) + P(s)} dt. \end{cases} \quad (7)$$

Substituting $s = 0$, $V(0) = K$ into formula (6) and expressing B from it, we get:

$$B \le \frac{K - U + C_\Sigma(0,S)}{Q_\Sigma(0,S)} = Z. \quad (8)$$

In this case, the equalities in (5), (6) and (8) are achieved with optimal S.

It follows that with the optimal assigned service life, the minimum of the Z index, which has the form of a fraction, is achieved. Let us find out the economic meaning of its numerator and denominator using formulas (7).

At first glance, it is obvious. The denominator of the fraction is the expected discounted amount of work performed by the system during the entire assigned service life, and the numerator is the expected discounted costs of its acquisition and use (less disposal value of the system)[7], including damage from failures. So, the fraction Z represents the expected specific (per unit of work) discounted costs (ESDC).

In fact, this term is not quite correct, since the values $Q(t)$, $C(t)$, $p(t)$ included in formulas (7) reflect not the dynamics of the characteristics of an operable system over its service life, but the characteristics of operable systems of different ages at the assessment date. Only when all prices in the country grow at the same rate i throughout the life of the system, the numerator and denominator of the fraction Z can be interpreted as the expected discounted costs and results associated with the acquisition and use of the system until the end of its service life. Nevertheless, the introduced term and its abbreviation ESDC seem convenient and visual, and we will use them.

As can be seen from formula (8), the optimal assigned service life should correspond to the minimum value of the ESDC. Moreover, this value will be equal to the market value of the unit of work performed by the object B. Such a method of evaluating works' market value is consistent with a cost approach to valuation, although appraisers do not use it in this form.

We also see that the application of the ESDC criterion orients the enterprise to the most efficient use of the technical system and maximizes its market value[8]. A similar criterion of discounted costs per unit for determining the service life of machines in a deterministic situation was previously justified using optimization models (see, for example, [18, 19]). Criteria similar in content were also used in solving

[7] Since the discounted performance and costs relate to an operable (not failed) technical system, for discounting we apply a "real" rate which takes into account the risk of failure depending on the age of the object $\rho + p(s)$.

[8] The market value of the enterprise, assuming that it assigns a different service life to the object, will decrease.

some optimization problems of the reliability theory (see, for example, [23, 24]); however, the productivity and intensity of operating costs of the system were assumed to be constant, and its disposal value was not taken into account.) In contrast to the criterion of the ratio of average costs to average service life, often used in works on the theory of reliability, the ESDC takes into account the different timing of costs and results of the system, but does not consider the time to eliminate the consequences of failures.

The calculation of the optimal value of S and the cost of a unit of work B can be simplified by using the inequality (5). Since the integrand in (5) decreases with the growth of t, it is not difficult to make sure that the maximum in the left part of (5) is reached when S is the unique root of the equation: $BQ(S) - C(S) - \rho U - (L + U)\, p(S) = 0$. In this case, inequality (8) becomes equality, and S and B will be the solution of a system of equations:

$$B = \frac{K - U + C_\Sigma(S)}{Q_\Sigma(S)} = \frac{C(S) + \rho U + (L+U)\, p(S)}{Q(S)}. \quad (9)$$

The dependence of the cost of the technical system on its age now follows from inequality (6), which at optimal S becomes equality:

$$V(s) = U + BQ_\Sigma(s,S) - C_\Sigma(s,S). \quad (10)$$

Substituting in (10) B from the first equality (9), we get:

$$V(s) = U + (K - U)\frac{Q_\Sigma(s,S)}{Q_\Sigma(S)} + \left[C_\Sigma(S)\frac{Q_\Sigma(s,S)}{Q_\Sigma(S)} - C_\Sigma(s,S)\right]. \quad (11)$$

This equality, in fact, is one of the modifications of the well-known Lvov formula [25], which was used in the USSR to evaluate the efficiency of new equipment and set prices for new technology, and is currently used in the machinery and equipment valuation (for the history of the Lvov formula and its modifications, see [3, 26, 27]). Unlike other modifications of the Lvov formula, (11) takes into account the scrap value of the system, as well as the risk of its failure and is applied to the used object valuation.

The characteristics of machines and equipment used in their valuation include average full and residual service life. They can be calculated for the systems under consideration. It would seem that for these purposes, formulas (1) and (2) can be used, linking the specified terms with the assigned service life (S). However, this would not be quite correct, since when deriving formula (1), it was assumed that until the system is disposed of, its assigned service life does not change. Such an assumption would be justified if, starting from the valuation date, group inflation with a constant rate would take place for the system. However, the service life of a system can be tens of years, while the rate of price growth for machines can fluctuate considerably. In this case, the optimal values of the assigned service life of the system will also change. Under such conditions, the values of $T(t)$ and T_m calculated by formulas (1) and (2) can no longer be called the average residual and full service life of the object. Rather, they reflect such terms calculated under the assumption of the group nature and the constant rate of inflation throughout the designated service life of the system. However, similar stipulation should be made regarding other technical and economic standards (for example, standards for frequency of repairs or safety margin factors), because they all were developed for specific economic conditions and therefore will be economically justified until these conditions change significantly.

4. Experimental calculations

According to the model constructed, experimental calculations were carried out in several variants. Time was measured in years and fractions of a year. In all variants, $V(0) = K = 100$, $Q(0) = 1$ were accepted. The market values of

technical systems of different ages at the same time numerically coincide with their PGF. It was assumed that the dependencies $C(t)$ and $Q(t)$ are linear, and the uptime of the systems has a Rayleigh distribution:

$$Q(t) = 1 - \alpha t,\ C(t) = C_0(1 + \beta t),\ p(t) = t/\omega^2,$$

where α and β – (basic) rates of productivity decline and growth of operating costs, respectively, 1/year;
ω – parameter od the Rayleigh distribution, years.

The optimal assigned service life of the system (S) and market values of units of work performed by the system on the valuation date (B) were found by numerical solution of the system of equations (9).

The basic variant was adopted, with the following values of parameters: $U = 7$; $C_0 = 100$; $L = 100$; $\rho = 0.1$; $\alpha = 0.01$; $\beta = 0.02$; $\omega = 8$.

By varying the parameters of this variant, it is possible to identify their influence on the optimal value S of the assigned service life of the system (*Figs. 1–4*).

Note also that the optimal S decreases slightly with the growth of the scrap value U and increases with the increase in the discount rate (especially strongly – with small damage values L).

The dependences of the average service life calculated by formula (2) on the intensity of operating costs at the beginning of operation (C_0) and ω at different damage values are shown in *Figs. 5–6*.

As one can see, new systems that require relatively large (compared to their value) operating costs have shorter average service lives. With a decrease in the failure rate (an increase in ω), the average service life of a system increases, but only up to a certain extent (the "economic" service life of similar objects that are not subject to failures).

The average market values of used objects as a function of age can be calculated using the formula (10). However, it is difficult to compare the appropriate graphs, since the average and assigned

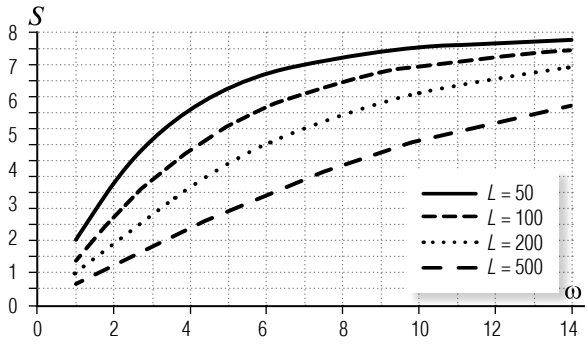

Fig. 1. Influence of ω on the optimal assigned service life under different L.

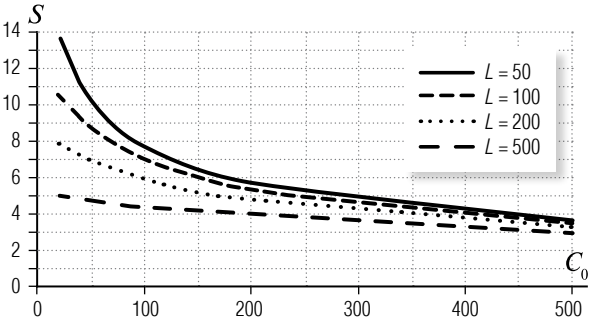

Fig. 2. Influence C_0 on the optimal assigned service life under different L.

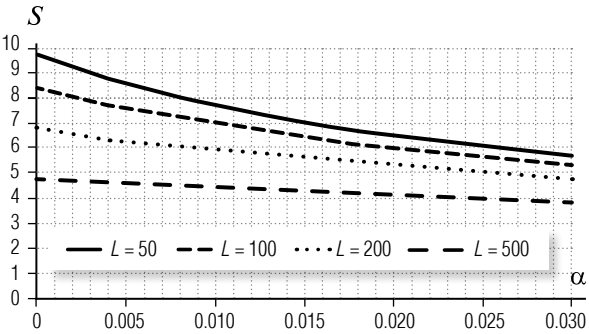

Fig. 3. Influence α on the optimal assigned service life under different L.

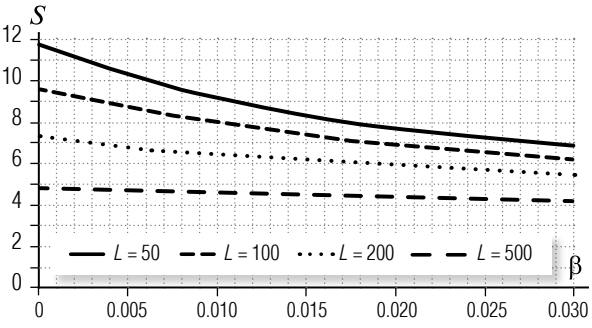

Fig. 4. Influence β on the optimal assigned service life under different L.

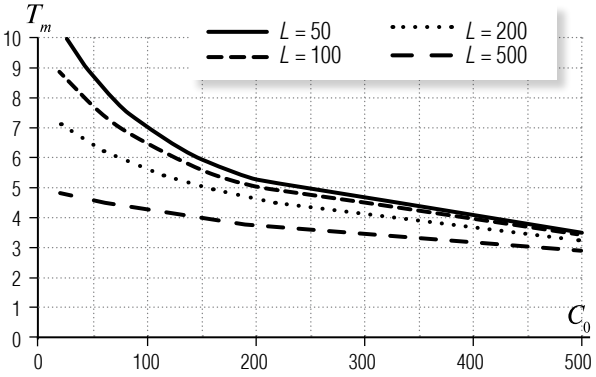

Fig. 5. Influence C_0 on the average service life of the system under different L.

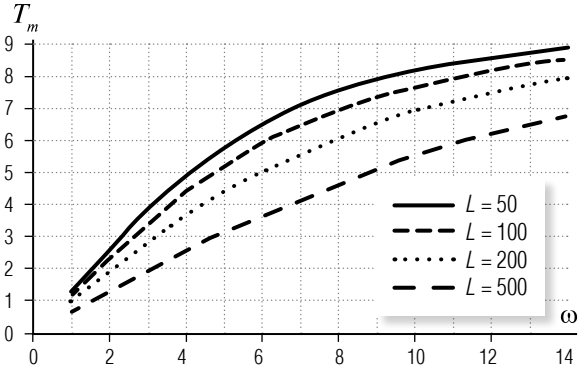

Fig. 6. Influence ω on the average service life of the system under different L.

service life for different variants vary significantly. However, the situation changes if, on the basis of these calculations, we plot the dependence of the average PGF on the relative age (τ) – the ratio of the age of the object (t) to the average period of their service (T_m).

Four variants of such dependencies are presented in *Fig. 7*. In all these variants, it was assumed that $K = 100$; $U = 7$; $Q_0 = 1$; $\rho = 0.1$; $\alpha = \beta = 0.01$. The values of other parameters of systems, as well as assigned and average service lives for these variants are summarized in *Table 1*.

Note that, despite significant differences in the main parameters of the system, the appropriate graphs turn out to be quite close. At the same time, they give lower PGF than when using the hyperbolic model adopted in the Russian system of national accounts for valuation of machinery and equipment.

Table 1.

Main parameters of technical systems by variants

Variant	1	2	3	4
C_0	20	100	40	300
L	100	200	200	500
ω	10	10	5	5
S	13.36	7.44	4.94	2.78
T_m	10.26	6.80	4.24	2.64

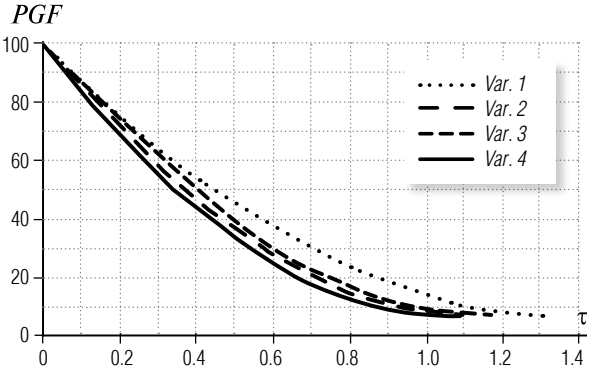

Fig. 7. Dependence of average Percent Good Factor (*PGF*) of a technical system on the relative age (τ) by variants.

Conclusion

The traditional approach to setting the assigned service life of technical systems does not fully meet the interests of market participants. To solve this problem, it is proposed to focus on maximizing the market value of the enterprise – owner of the object and to rely on the theory of valuation. It turns out that this approach makes it possible to consider inflation quite simply. In addition, the assigned service life of the object should provide the minimum expected unit discounted costs – the ratio of the expected discounted costs of acquiring and using the system and eliminating the consequences of its failures to the expected discounted amount of work performed by the system. At the same time, the mathematical

model we developed allows us to estimate the market value of the work performed by the system and establish the dependence of the market value of the object on its age. The proposed approach can also be used in a situation where the degradation of an operable object is described by random processes, as well as when solving other optimization problems of the reliability theory, for example, to optimize the schedule of overhauls. ■

References

1. *International Valuation Standards* (2019) London: International Valuation Standards Council.
2. Gnedenko B.V., Belyaev Yu.K., Soloviev A.D. (1965) *Mathematical methods in the reliability theory*. Moscow: Nauka (in Russian).
3. Smolyak S.A. (2016) *Machinery and Equipment Valuation (secrets of the DCF method)*. Moscow: Option (in Russian).
4. *System of National Account 2008* (2009) New York: European Commission, International Monetary Fund, Organization for Economic Co-operation and Development, United Nations, World Bank.
5. Smolyak S.A. (2021) Assessment of machinery depreciation taking into account the provisions of the systems of national accounts. *Proceedings of the Institute for System Analysis of the Russian Academy of Sciences*, vol. 71, no. 1, pp. 44–54 (in Russian). https://doi.org/10.14357/20790279210105
6. Ostreykovsky V.A. (2003) *Reliability theory: Textbook for universities*. Moscow: Vysshaya schkola (in Russian).
7. *Methodical recommendations for assessing damage from accidents on hazardous production facilities (Regulatory Document 03-496-02)* (2010) Moscow: Closed Joint Stock Company «Scientific and Technical Center for Research on Industrial Safety Problems» (in Russian).
8. Lutz J., Hopkins A., Letschert V.E., Franco V., Sturges A. (2011) Using national survey data to estimate lifetimes of residential appliances. *HVAC&R Research,* vol. 17, no. 5, pp. 726–736. https://doi.org/10.1080/10789669.2011.558166
9. Erumban A.A. (2008) Lifetimes of machinery and equipment evidence from Dutch manufacturing. *Review of Income and Wealth*, vol. 54, no. 2, pp. 237–268. https://doi.org/10.1111/j.1475-4991.2008.00272.x
10. Nomura K., Suga Y. (2018) Measurement of depreciation rates using microdata from disposal survey of Japan. In: *The 35th IARIW General Conference, Copenhagen, Denmark*.
11. Alekseev V.V., Khomenko I.V., Prokhorsky R.A. (2011) Models of planning repairs and replacement of elements in the life cycle of complex technical systems. *Bulletin of the Voronezh Institute of the Ministry of Internal Affairs of Russia*, no. 3, pp. 94–102 (in Russian).
12. Van Horenbeek A., Pintelon L., Muchiri P. (2010) Maintenance optimization models and criteria. *International Journal of System Assurance Engineering and Management*, vol. 1, no. 3, pp. 189–200.
13. Barlow R. and Hunter L. (1960) Optimum preventive maintenance policies. *Operations Research*, vol. 8, no. 1, pp. 90–100. https://doi.org/10.1287/opre.8.1.90
14. Jiang R. (2018) Performance evaluation of seven optimization models of age replacement policy. *Reliability Engineering & System Safety*, vol. 180, pp. 302–311. https://doi.org/10.1016/j.ress.2018.07.030
15. Smith D.J. (2011) *Reliability, maintainability and risk: Practical methods for engineers*. Elsevier Ltd, Eighth edition.
16. Aven T. (1983) Optimal replacement under a minimal repair strategy. *Advances in Applied Probability*, vol. 15, no. 1, pp. 198–211. https://doi.org/10.2307/1426990
17. Weersink A., Stauber S. (1988) Optimal replacement interval and depreciation method for a grain combine. *Western Journal of Agricultural Economics*, vol. 13, no. 1, pp. 18–28. https://doi.org/10.22004/ag.econ.32156
18. Vilensky P.L., Livchits V.N., Smolyak S.A. (2015) *Evaluation of the investment projects efficiency: theory and practice*. Moscow: PolyPrintService (in Russian).

19. State of Colorado. The Department of Local Affairs (2021) *2021 Personal Property Factors & Tables*. Available at: https://drive.google.com/file/d/1FZl9Da3QebSCDSVTPpZEX0H-6t0c9MNd/view (accessed 22 May 2022).

20. Utah State Tax Commission. Property Tax Division (2021) *2021 Recommended Personal Property Valuation Schedule*. Available at: https://propertytax.utah.gov/personal-property/val_schedule_2021.pdf (accessed 22 May 2022).

21. Montana Department of Revenue (2022) *2022 Personal Property Depreciation Schedules and Trend Tables*. Available at: https://mtrevenue.gov/?mdocs-file=59493 (accessed 22 May 2022).

22. Livshits V.N., Smolyak S.A. (1990) Models of the dynamics of economic wear of equipment. *Economics and Mathematical Methods*, vol. 26, no. 5, pp. 871–882 (in Russian).

23. Christer A.H., Goodbody W. (1980) Equipment replacement is unsteady economy. *Journal of the Operational Research Society*, vol. 31, pp. 497–506.

24. Dohi T., Ashioka A., Osaki S., Kaio N. (2001) Optimizing the repair-time limit replacement schedule with discounting and imperfect repair. *Journal of Quality in Maintenance Engineering*, vol. 7, no. 1, pp. 71–84. https://doi.org/10.1108/13552510110386973

25. Lvov D.S. (1969) *Economic problems of improving the quality of industrial products*. Moscow: Nauka (in Russian).

26. Smolyak S.A. (2018) On D.S. Lvov's formula for the valuation of machines. *Property relations in the Russian Federation*, vol. 205, no. 10, pp. 19–25 (in Russian). https://doi.org/10.24411/2072-4098-2018-10101

27. Smolyak S.A. (2020) Generalized D.S. Lvov's formula for machines subject to degradation. *Proceedings of the Institute for System Analysis of the Russian Academy of Sciences*, vol. 3, pp. 3–12 (in Russian). https://doi.org/10.14357/20790279200301

About the author

Sergey A. Smolyak

Dr. Sci. (Econ.);

Principal Science Researcher, Central Economics and Mathematics Institute, Russian Academy of Science, 47, Nakhimovsky Prospect, Moscow 117418, Russia;

E-mail: smolyak1@yandex.ru

ORCID: 0000-0001-5287-4285