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

Vol. 15 No 1 – 2021

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

Editor-in-Chief:
Y. Koucheryavy

Deputy Editor-in-Chief
E. Zaramenskikh

Computer Making-up:
O. Bogdanovich

Website Administration:
I. Khrustaleva

Address:
28/11, build. 4, Shablovka Street
Moscow 119049, Russia

Tel./fax: +7 (495) 772-9590 *26311
http://bijournal.hse.ru
E-mail: bijournal@hse.ru

Circulation:
English version – 150 copies,
Russian version – 150 copies,
online versions in English and Russian – open access

Printed in HSE Printing House
3, Kochnovsky Proezd, Moscow, Russia

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 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).

Editor-in-Chief: Dr. Yevgeni A. Koucheryavy.

EDITORIAL BOARD

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.2021.1.7.18

Using out-of-sample Cox–Snell residuals in time-to-event forecasting

Ekaterina V. Rumyantseva
E-mail: evrumyantseva@hse.ru

Kirill K. Furmanov
E-mail: kfurmanov@hse.ru

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

Abstract

The problem of assessing out-of-sample forecasting performance of event-history models is considered. Time-to-event data are usually incomplete because the event of interest can happen outside the period of observation or not happen at all. In this case, only the shortest possible time is observed and the data are right censored. Traditional accuracy measures like mean absolute or mean squared error cannot be applied directly to censored data, because forecasting errors also remain unobserved. Instead of mean error measures, researchers use rank correlation coefficients: concordance indices by Harrell and Uno and Somers' Delta. These measures characterize not the distance between the actual and predicted values but the agreement between orderings of predicted and observed times-to-event. Hence, they take almost "ideal" values even in presence of substantial forecasting bias. Another drawback of using correlation measures when selecting a forecasting model is undesirable reduction of a forecast to a point estimate of predicted value. It is rarely possible to predict the timing of an event precisely, and it is reasonable to consider the forecast not as a point estimate but as an estimate of the whole distribution of the variable of interest. The article proposes computing Cox–Snell residuals for the test or validation dataset as a complement to rank correlation coefficients in model selection. Cox–Snell residuals for the correctly specified model are known to have unit exponential distribution, and that allows comparison of the observed out-of-sample performance of a forecasting model to the ideal case. The comparison can be done by plotting the estimate of integrated hazard function of residuals or by calculating the Kolmogorov distance between the observed and the ideal distribution of residuals. The proposed approach is illustrated with an example of selecting a forecasting model for the timing of mortgage termination.

Key words: forecasting; event-history analysis; Cox–Snell residuals; censoring.

Citation: Rumyantseva E.V., Furmanov K.K. (2021) Using out-of-sample Cox–Snell residuals in time-to-event forecasting. *Business Informatics*, vol. 15, no 1, pp. 7–18.
DOI: 10.17323/2587-814X.2021.1.7.18

Introduction

There are various problems in statistics and data analysis that require modeling the time at which a certain event occurs. It can be the timing of a credit default in financial applications, time until death or recovery of a patient in survival analysis, the age of a woman at first marriage or age of the mother at first birth in social and demographical research. Such problems are considered in a branch of statistics called event-history analysis. This branch has important peculiarities that distinguish event-history analysis from more traditional statistics. One of these peculiarities is data censoring.

Every study lasts for a finite period of time, and the event of interest does not necessarily occur within this period. More than that, there may be objects under study that never face the event: some debtors pay off the loan, so that default never occurs for them; some women never give birth to a child. As a result, the only thing that is known about these objects is that the time-to-event exceeds a certain value which is the duration of a timespan between the start of waiting for the event and the end of the study. Such observations are called right censored.

Techniques for analyzing censored data are quite well known; the classical textbook is [1]. Measuring accuracy of time-to-event forecasts is a less developed field of study. Partly this can be explained by the fact that event history models were commonly constructed not for forecasting but for academic purposes like testing hypotheses about the efficiency of a medical treatment or social policy, revealing the individual attributes that are correlated with the duration of unemployment, etc.

The past decade has shown a growing interest in forecasting. On the one hand, nowadays event history models often find purely practical applications: financial risk assessment [2, 3], predicting the length of crowdfunding campaigns [4]. On the other hand, the expansion of machine learning and, in particular, the widespread use of cross-validation procedures has given rise to assessing the quality of statistical models by their out-of-sample predictive accuracy [5, 6]. The special thing is that commonly used accuracy measures like the mean squared error or the mean absolute percentage error are inapplicable when dealing with censored data.

The article proposes an approach to model selection that is based on combination of (1) concordance coefficients for actual and predicted timings and (2) Cox–Snell residuals. Both concordance coefficients and residuals are calculated for a test sample. The proposed approach is illustrated with an example of building a forecasting model for the timing of mortgage prepayment.

The next section describes the basic concepts that have to be defined or explained because of peculiarities of the event-history analysis. Section 2 contains a review of measures of predictive accuracy that can be applied to censored time-to-event data. Section 3 is devoted to Cox–Snell residuals and their use for model assessment. The use of Cox–Snell residuals for model selection is illustrated with a real data example in section 4, which is followed by a conclusion.

1. Probabilistic model of event occurrence

Time of event occurrence is modeled as a nonnegative random variable that can be either

discrete or continuous according to the nature of the process under study and available data. Here we consider only the continuous case. The distribution of this random variable can be characterized by the following functions that play a central role in event history analysis.

Survival function $S(t)$ reflects the probability that time-to-event exceeds the value of the argument:

$$S(t) = P(T > t).$$

The term refers to actuarial and medical applications where the event of interest is death of an insured person or a patient, so that the value of the survival function is the probability of survival until time t.

Hazard function $h(t)$ reflects changes in the probability of event occurrence over time:

$$h(t) = \lim_{\substack{\Delta \to 0 \\ \Delta > 0}} \frac{P(t < T \le t + \Delta \mid T > t)}{\Delta}.$$

Intgrated hazard function $H(t)$ (also called cumulative hazard) does not have a clear interpretation but plays an important role in this article:

$$H(t) = \int_0^t h(s)ds.$$

Te terminology differs from one area of application to another, and the same functions are known under different names. Survival function is sometimes called reliability function, and the hazard function is also known as the mortality intensity rate or the force of mortality.

Typically, there are two aspects of event history that a researcher is interested in.

The first aspect is a relation between the probability of an event's occurrence in the near future and the time of waiting for the event. This relation is conveniently represented by the hazard function.

The second aspect is a relation between the probability of an event's occurrence and explanatory variables (covariates). There are a variety of regression models that link the distribution of time with covariates; we refer an interested reader to books [1, 7] for a detailed review. Four event-history models that are used in this article for illustration purposes are briefly described below.

Lognormal and generalized gamma regressions are special cases of an accelerated failure-time model, which means that they have a linear form representation:

$$\ln T = x'\beta + \varepsilon.$$

Here x' denotes the row vector of explanatory variables, β is the column vector of estimated coefficients, and ε stands for a random error. Apart from covariates vector x' includes a unit element that corresponds to an intercept term, so that $x'\beta = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k$. Lognormal and generalized gamma regressions differ only in assumed distribution of a random error.

Gompertz and Cox regression are proportional hazard models which means that they assume the hazard function to be proportional to covariates:

$$h(t; x', \beta) = h_0(t) \exp(x'\beta).$$

Here $h_0(t)$ denotes the so-called baseline hazard (the hazard that corresponds to zero effect of explanatory variables). The Gompertz model assumes that the hazard exponentially grows or declines with time: $h_0(t) = e^{\gamma t}$, where γ is an estimated parameter of the baseline distribution. Cox regression does not impose restrictions on the baseline hazard which is estimated via the nonparametric technique. As well as other considered models, the Cox regression restricts a functional form of relation between the time-to-event distribution and explanatory variables, but (in contrast to other models) it permits any kind of dependence between the probability of event occurrence and time.

Estimating procedures for these models are implemented in statistical packages and described in textbooks [1, 7]. The only thing

important for our purposes is that each of these regressions allows a researcher to obtain estimates of survival and integrated hazard functions for arbitrary values of explanatory variables x'. This means the possibility to predict the event's occurrence under the given conditions.

2. Predictive accuracy measures: A review

Measures based on averaging prediction errors. This group includes the most widely used metrics of forecasting accuracy for the models with quantitative response: mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE) etc. Although there are examples of their application in event-history analysis [2, 5, 8], these examples are mere exceptions. In most cases, data are subject to censoring so that the differences between actual and predicted timing of events are not precisely known and cannot be averaged. This problem can easily be solved under the assumption that the prediction errors follow a certain parametric family of distributions. In this case, one may estimate the parameters of this distribution via maximum likelihood and compute the corresponding mean. However, we have not found this approach in academic literature. A possible explanation is that researchers avoid making additional assumptions

Papers [2, 8] consider calculating the mean absolute error only for uncensored observations that contain exact timings of event occurrence. This approach has a substantial drawback because censoring depends on those timings. The longer the observer has to wait for an event, the greater is the probability that the observation period will end before the occurrence and the observation will be censored. Consequently, exact timings will be known mostly in those cases when they are small, and the mean absolute error will take into account only these observations. As a result, a model that predicts early occurrence will be preferred to a model that gives unbiased forecasts. One can say that excluding censored observations leads to sample truncation which is no less troublesome than censoring [1].

Rank correlation coefficients and concordance indices. Harrell's concordance index [9], or *C*-index, is probably the most widely used predictive accuracy measure in event history analysis. Let random variables T_1 and T_2 denote times of event occurrence in two randomly chosen independent observations, and $\hat{T}_1$ and $\hat{T}_2$ denote corresponding predictions. Harrell's *C*-index is defined by the following expression:

$$C = P(\hat{T}_1 < \hat{T}_2 \mid T_1 < T_2). \quad (1)$$

One advantage of this coefficient is its clear interpretability: if time-to-event differs in two cases, then *C* equals the probability that a model predicts a greater value in the observation with a greater actual time. The largest possible value of *C* is one and it is achieved when rankings of actual and predicted values are completely concordant, so that when the event of interest occurs earlier, the model always predicts earlier occurrence too. The lower bound for *C*-index is zero which means complete discordance of actual and predicted rankings (the earlier the event happens, the longer is predicted waiting time).

There are various estimators for the concordance coefficient in the presence of censoring. One of them is a statistic originally proposed by Harrell et al. [9], another example is Uno's estimator [10] that is gaining growing popularity. Apart from the concordance index, Somers' *D* correlation coefficient can be used for the same purpose [11, 12].

The mentioned metrics share a drawback. They measure only the association between rankings of actual and predicted values, which means that they assess a model's ability to distinguish the cases where an event occurs relatively early from those where the event occurs relatively late. It is not a predictive accuracy in the sense that it does not reflect the differ-

ence between real and predicted values. Suppose that the model provides forecasts that are exactly ten times greater than observed values. One would hardly call this forecast accurate, but the concordance index or any other rank correlation coefficient would achieve its highest value of one because the predicted ranking is perfectly concordant with the actual one.

Sometimes it is a model's ability to detect objects with relatively long time-to-event that draws the attention of a researcher. It can be so if, for example, the aim of the study is revealing markers of early recovery or death of patients [13]. However, in many practical cases the analyst is interested in the absolute value of an explained variable. Paper [14] provides an example of such a study in medicine, but such interest seems to be more common in financial applications where times of defaults and prepayments determine cash flow [2, 3, 15].

Another class of measures used for evaluating the predictive power of event-history models consists of ***classification metrics*** that evaluate accuracy of binary prediction (whether an event occurs in a certain period of time or not). This class has been actively developed in the last decade [16, 17] and deserves attention, but we do not review these metrics here, because they represent a substantially different approach to forecasting. However, Harrell's C-index can be considered also as a classification metric [18].

3. Cox–Snell residuals and their application to predictive power assessment

Consider a sample $(T_1, x_1'),...,(T_n, x_n')$, where T_i denotes time to event and x_i' is the vector of explanatory variables in observation i. Let $\hat{H}(t; x')$ denote an estimate for the integrated hazard function of T_i random variables (it can be obtained from some regression model). A Cox–Snell residual [19] in observation i is defined as follows: $r_i^{CS} = \hat{H}(T_i; x_i')$.

If the estimate $\hat{H}(t; x')$ coincides with the true function $H(t; x')$, then the Cox–Snell residuals follow exponential distribution with unit mean. In this case, the integrated hazard function of the residuals is $H^{CS}(t) = t$.

Below we describe the visual test that is commonly used for regression diagnostics. It is based upon the Cox–Snell residuals and is performed in three steps.

1. Estimate the regression model and compute Cox–Snell residuals for each observation.

2. Compute the estimate of the integrated hazard function of the residuals $\hat{H}^{CS}(t)$. If some observations on time-to-event are censored, the corresponding residuals are also censored, which should be taken into account. We use the Nelson–Aalen technique [20, 21] to estimate the integrated hazard from censored data.

3. Plot the estimate of the integrated hazard against residuals r_i^{CS}. Further we refer to this plot as the Cox–Snell residuals plot. In case of a correctly specified model, the integrated hazard estimate approximately lies on the line $H(t) = t$ (*Figure 1a*). An example for incorrect model specification is presented in *Figure 1b*.

This test is widely known and described in textbooks; Cox–Snell residual plots are presented by researchers to assess the goodness-of-fit of models they use [7, 22, 23]. We have not found examples of using Cox–Snell residuals for evaluating forecasting accuracy; the possible reasons are discussed in the conclusion to this article. Further Cox–Snell residuals calculated for the test sample are called out-of-sample residuals because they characterize the model's performance outside the sample used for estimation.

Out-of-sample residuals can be used to detect the systematic prediction bias. Consider a case where a fitted hazard function is c times greater than the true hazard: $h(t; x') = c\, h(t; x')$. Then the Cox–Snell residuals are also c times greater that the integrated hazard function: $r_i^{CS} = \hat{H}(T_i; x_i') = cH(T_i; x_i')$. As a result, a residual

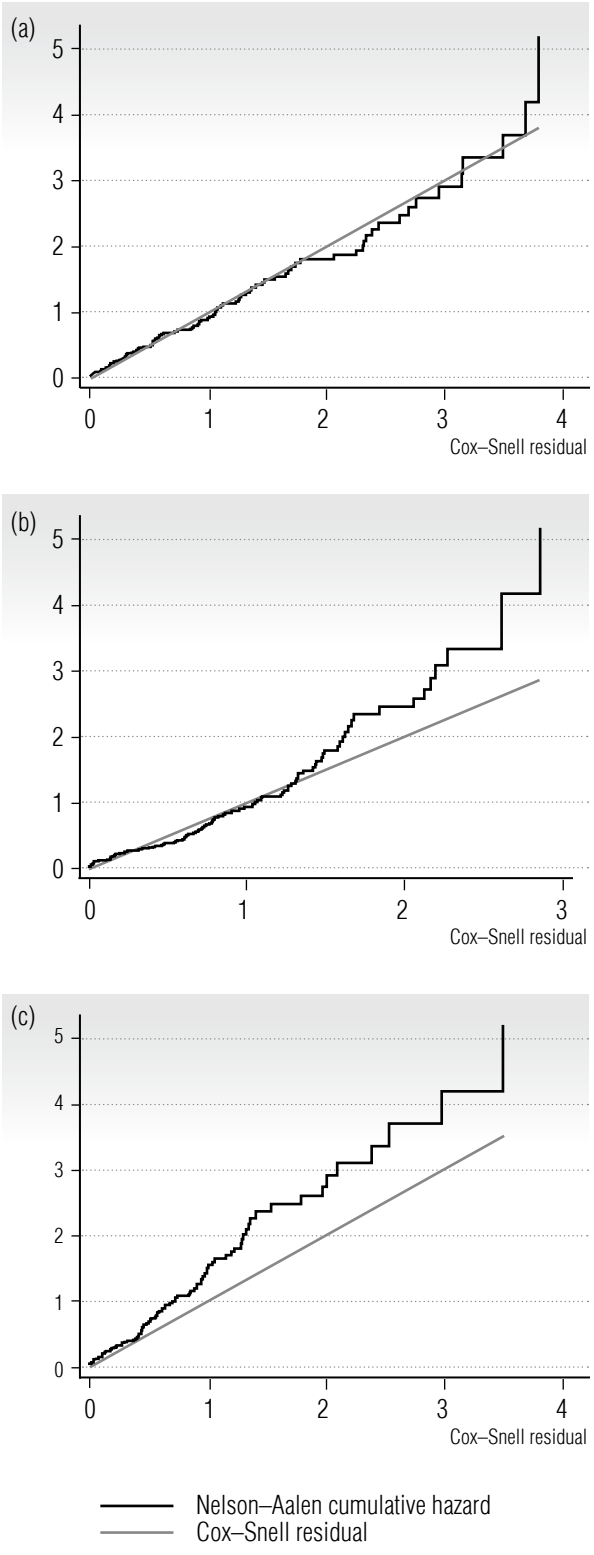


Fig. 1. Examples of Cox–Snell residuals plot for
(a) a correctly specified model;
(b) a model based on inappropriate time–to–event distribution;
(c) a model with systematic prediction bias

plot is higher or lower (depending on the value of c) than the line $H(t) = t$ (*Figure 1c*). This situation is practically impossible when assessing the regression performance in the training sample and is unlikely even when examining out-of-sample predictions if the test sample was selected at random. Such prediction bias is more likely to be found when performing external validation, evaluating the model's performance with new data.

The residual plot is a preferable tool for manual model selection, while automatic selection requires numerical measure for goodness-of-fit. Further we use the Kolmogorov distance between the survival estimate for the Cox–Snell residuals $S^{CS}(t) = \exp(-\hat{H}^{CS}(t))$ and the corresponding function for unit exponential distribution $S(t) = e^{-t}$ as such a metric:

$$KD = \sup_{t\in[0;t_{\max}]} \left|\hat{S}^{CS}(t) - e^{-t}\right|. \qquad (2)$$

Here $t_{\max}$ denotes the largest time in the sample (it can be censored). We consider only the set $[0; t_{\max}]$, because the Nelson–Aalen method does not allow estimating the right tail of the distribution. The corresponding survivor function at $t_{\max}$ has not reached zero yet, and there is no data to estimate it for the values of the argument greater than $t_{\max}$.

The next section contains an example of using this metric for predictive model selection.

4. Example: Modeling mortgage prepayment

The example uses data from a large mortgage agency. The data contain more than 280,000 observations on mortgage contracts concluded from 2001 to 2013. The explained variable is the time between a conclusion of a contract and its prepayment. Observations are right-censored due to the following reasons:

✦ The end of the observation period: we use the data gathered on 1 January 2014, these data do not contain information on exact payment date for the loans that were not paid by that date.

✦ The termination of the contract: if the prepayment did not occur, then the observation is treated as right-censored as if there is a possibility of prepayment after the termination date. This is a statistical trick convenient when objects under study are exposed to mutually exclusive events (like prepayment and payment in time in our example). It is convenient to suppose that both events happen but only the first of them is observed.

✦ Mortgage default. This is another event that prevents prepayment.

This dataset is used for estimation of several event-history models that differ in (a) assumed distribution of the explained variable and its relation to the covariate vector (lognormal and generalized gamma regressions, Cox and Gompertz models), and (b) the set of explanatory variables ("short" and "long" models). Both "short" and "long" models include attributes of a loan, of a main borrower and of a subject of mortgage. The "short" model accounts for the interest rate of the loan, the credit term, payment-to-income ratio, age of the main borrower, type of employment of the main borrower and the number of rooms in the subject of mortgage. The "long" model includes, apart from all the mentioned attributes, sex, marital status and education of the main borrower, the number of co-borrowers, regional effect (measured according to the agency's rating of socio-economic development of regions of Russian Federation that divides all the regions into three groups with low, moderate and high level of development), the type of the mortgage subject (house or apartment), the ratio of living space to total space, the ratio of total amount of planned payments to the price of the mortgage subject, and a loan-to-value ratio.

The whole dataset is randomly partitioned into training and test samples in proportion at a ratio 60:40 respectively.

Plots of out-of-sample Cox–Snell residuals for eight estimated models are given in the Appendix. The distances between observed and theoretical distributions of the residuals are calculated according to equation (2) and presented in *Table 1* among with Harrell's coefficient values.

It is seen from Table 1 that the concordance index is practically the same for different baseline distributions but depends on the choice of covariates: the "long" model stably outperforms the "short" one. The probability of concordance between actual and predicted values is greater for the "long" model by approximately 0.02. An analyst may consider it to be a minor discrepancy but it is stable. Repeating the random split into training and test samples, we found the difference between the values of the *C*-index for "long" and "short" models to be essentially the same. On the contrary, Kolmogorov's distance between ideal and observed distributions of Cox–Snell residuals substan-

Table 1.

Out-of-sample performance of mortgage prepayment models

	"Short" model		"Long" model	
	Harrell's C-index	Kolmogorov distance	Harrell's C-index	Kolmogorov distance
Lognormal	0.593	0.078	0.612	0.066
Generalized gamma	0.593	0.015	0.610	0.009
Gompertz	0.592	0.059	0.608	0.063
Cox	0.593	0.007	0.609	0.014

tially depends on the choice of the baseline distribution: generalized gamma regression and the Cox model provide better goodness-of-fit than lognormal and Gompertz regressions both in the case of "short" and "long" models.

A natural question is: do predictions obtained from these models really differ? *Figure 2* presents plots of estimated survivor functions obtained from the mentioned models for a loan with typical values of explanatory variables:

- ✦ gender of the main borrower: male;
- ✦ marital status of the main borrower: married;
- ✦ education of the main borrower: higher education or doctoral degree;
- ✦ employment type of the main borrower: private sector employee,
- ✦ number of co-borrowers including the main borrower: 2;
- ✦ age of the main borrower: less than 35 years;
- ✦ payment-to-income ratio: from 20% to 35%;
- ✦ type of the mortgage subject: apartment;
- ✦ location of the mortgage subject: moderately developed region;
- ✦ number of rooms in the mortgage subject: 2;
- ✦ annual interest rate: less than 11,5% (low-risk interest rate);
- ✦ ratio of living space to total space of the mortgage subject: from 50% to 70%;
- ✦ loan-to-value ratio: from 50% to 70%;
- ✦ credit term: more than 180 months (long term loan);
- ✦ ratio of total amount of planned payments to the price of the mortgage subject: from 1 to 1.82.

The probability of survival in the next several years for a newly made loan is shown in *Figure 2a*, while *Figure 2b* presents the same probability for a five-year old loan (the conditional survival function with respect to condition $\{T > 5\}$). The horizontal axis represents days after the start of the loan term, and the vertical axis represents the probability that the loan is not paid off until that time. Survival curves obtained from the Cox model ("S_cox" line) and from the generalized gamma regression ("S_gamma") practically coincide in both figures. The curve estimated from the Gompertz regression ("S_gomp") deviates from the others and shows a greater risk of prepayment. On the contrary, the lognormal curve ("S_ln") predicts the lowest probability of prepayment, and deviates from gamma and Cox curves for an "old" loan but not in the first years of a mortgage credit. Plots of out-of-sample Cox-Snell residuals that are given in Appendix, show that the gamma model correctly predicts the distribution of prepayment times and the Cox regression performs almost as well. Residuals from the Gompertz regression have a distribution that is completely different from exponential, and the lognormal model provides residuals that lack goodness-of-fit in the right tail of the distribution.

Hence, if one chooses a model on the basis of Harrell's index, then the lognormal regression is chosen which leads to underestimation of the probability of prepayment. Taking into account out-of-sample Cox–Snell residuals, an analyst would prefer either generalized gamma regression or the Cox model, both resulting in similar forecasts that agree with the distribution of prepayment times in the test set. From *Figure 2* it follows that the estimated probability of prepayment substantially depends on the choice of forecasting model. The difference is especially large for a long-term forecast horizon: the distance between the survival curves in *Figure 2b* exceeds 20 percentage points in the right side of the plot. However, the difference between two-year-ahead forecasts already exceed 5 percentage points.

Conclusion

Cox–Snell residuals and the corresponding visual test for model specification are well-known and covered in both research papers and textbooks. However, these residuals do not seem to be a popular tool for assessing predictive accuracy. This is an important point that naturally raises a question: are they really useful? The example considered in a previous sec-

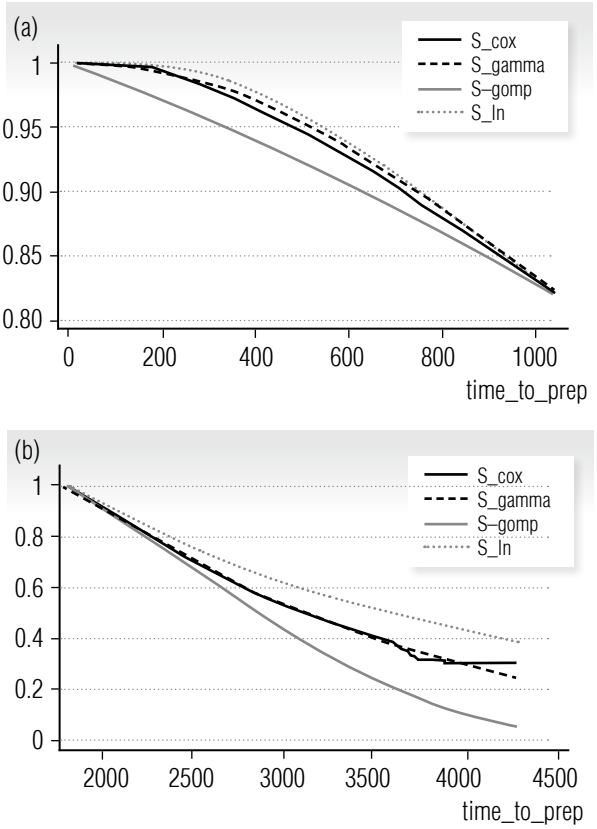


Fig. 2. Estimated survival functions for (a) a new loan; (b) a five–year loan

tion was presented to demonstrate the benefits of examining out-of-sample residuals, but now it is time to make a substantial caveat.

Cox–Snell residuals alone cannot be regarded as a measure of predictive error. They only allow us to compare the actual distribution in the data with the distribution that should be observed according to the model. This comparison can be performed either on the training sample (as is usually done), or on the test sample (as we suggest for evaluating forecasting performance). Note that a model with no covariates, that yields the same prediction for all observations, may produce practically ideal exponential Cox–Snell residuals if a proper distribution for an explained variable is chosen. One need not think that such model would outperform a regression that includes informative covariates but has improper baseline distribution that would be detected by examining the residuals plot.

On the contrary, rank correlation and concordance coefficients are a useful tool for selecting a set of explanatory variables, but do not characterize the ability of a model to predict the whole distribution of event timings. More common problems of regression analysis do not often require such prediction, and an analyst is interested mostly in a point forecast that can be obtained by estimating the mean or median of an explained variable. However, a time of event occurrence can almost never be characterized by a single number and reducing the forecast to a point estimate is counterproductive. Quantiles are of no less importance than the mean, and as the quantile function uniquely characterize the distribution of a random variable, one can say that the whole distribution is important. Often researchers do not pay much attention to distribution selection and prefer to use the Cox model that does not require parameterizing a baseline hazard. The example presented in this article demonstrates that a relatively simple parametric model can outperform Cox regression even on a very large sample. The Cox model is prone to overfit, and it restricts the functional form of relation between hazard and covariates, so that the shape of the hazard function is essentially the same for all values of explanatory variables. Parametric models can account for other forms of dependence.

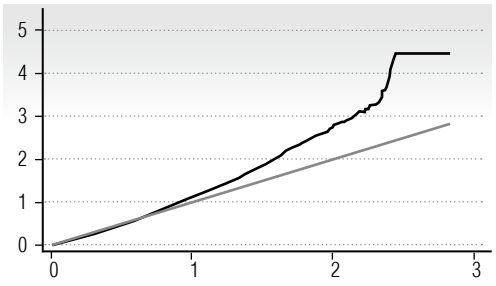
Examining the concordance index and the Cox–Snell residuals plot, a researcher can assess both the choice of explanatory variables and the selected baseline distribution. This combination gives a full picture of the model's out-of-sample performance. Of course, it would be more convenient to have a single metric that would allow unambiguous ranking for a set of models according to their forecasting accuracy, but at the moment such a metric seems to be unavailable. ■

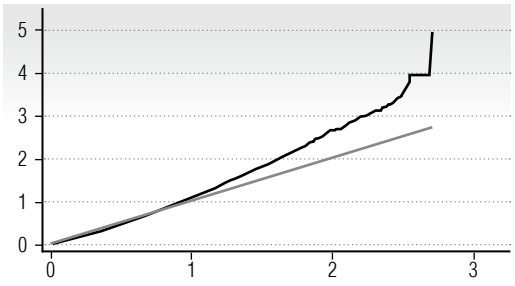
Appendix

Out-of-sample Cox–Snell residuals plots for mortgage prepayment models:

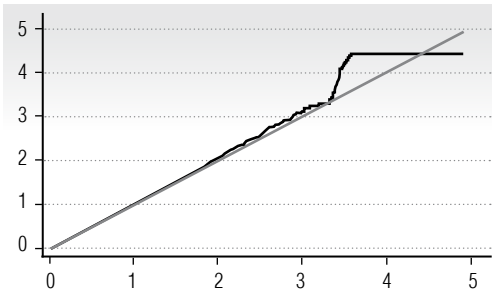
«Short» model	«Long» model

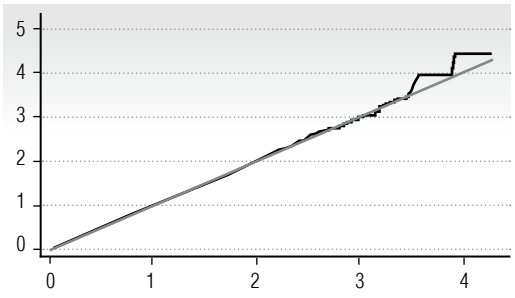
Lognormal regression



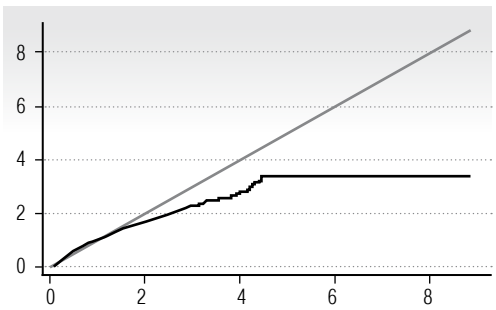


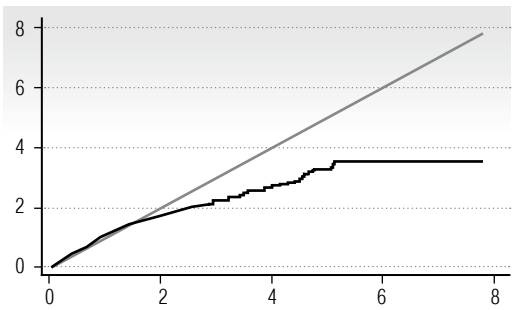
Generalized gamma regression



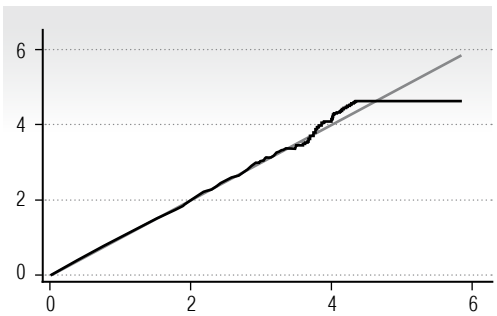


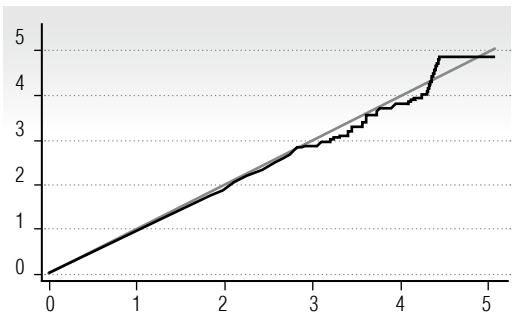
Gompertz regression





Cox model





Nelson–Aalen cumulative hazard Cox–Snell residual

References

1. Klein J.P., Moeschberger M.L. (2005) *Survival analysis: Techniques for censored and truncated data*. Second edition. Springer.
2. Zhang J., Thomas L.C. (2012) Comparison of linear regression and survival analysis using single and mixture distribution approaches in modelling LGD. *International Journal of Forecasting*, vol. 28, no 1, pp. 204–215. DOI: 10.1016/j.ijforecast.2010.06.002.
3. Zamisniy P., Kozlov A. (2018) Using credit history data for estimating the probabilities of default and early termination. *Bank Crediting*, no 4. pp. 4–11 (in Russian).
4. Rakesh V., Lee W.-C., Reddy C.K. (2016) Probabilistic group recommendation model for crowdfunding domains. Proceedings of the *9th ACM International Conference on Web Search and Data Mining (WSDM 2016), San Francisco, California, USA, 22–25 February 2016*, pp. 257–266. DOI: 10.1145/2835776.2835793.
5. Ameri S., Fard M.J., Chinnam R.B., Reddy C.K. (2016) Survival analysis based framework for early prediction of student dropouts. Proceedings of the *25th ACM International Conference on Information and Knowledge Management (CIKM 2016), Indianapolis, Indiana, USA, 24–28 October 2016*, pp. 903–912. DOI: 10.1145/2983323.2983351.
6. Wang P., Li Y., Reddy C.K. (2019) Machine learning for survival analysis: A survey. *ACM Computing Surveys*, February, article no 110. DOI: 10.1145/3214306.
7. Cleves M.A., Gould W.W., Gutierrez R.G., Marchenko Y.U. (2010) *An introduction to survival analysis using Stata*. Third Edition. College Station, Texas: Stata Press.
8. Dirick L., Claeskens G., Baesens B. (2017) Time to default in credit scoring using survival analysis: a benchmark study. *Journal of Operational Research Society*, vol. 68, no 6, pp. 652–665. DOI: 10.1057/s41274-016-0128-9.
9. Harrell F.E. Jr., Califf R.M., Pryor D.B., Lee K.L., Rosati R.A. (1982) Evaluating the yield of medical tests. *Journal of the American Medical Association*, vol. 247, no 18, pp. 2543–2546. DOI: 10.1001/jama.1982.0332043004703.
10. Uno H., Cai T., Pencina M.J., D'Agostino R.B., Wei L.J. (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, vol. 30, no 10, pp. 1105–1117. DOI: 10.1002/sim.4154.
11. Newson R.B. (2010) Comparing the predictive powers of survival models using Harrell's C or Somers' D. *The Stata Journal*, vol. 10, no 3, pp. 339–358.
12. Somers R.H. (1962) A new asymmetric measure of association for ordinal variables. *American Sociological Review*, vol. 27, no 6, pp. 799–811.
13. Martinez-Romero J., Bueno-Fortes S., Martín-Merino M., de Molina A.R., De Las Rivaz J. (2018) Survival marker genes of colorectal cancer derived from consistent transcriptomic profiling. *BMC Genomics*, no 19, article no 857. DOI: 10.1186/s12864-018-5193-9.
14. Kalinin M.N., Khasanova D.R., Ibatullin M.M. (2019) Possible timing for anticoagulation therapy initiation in ischemic stroke patients with atrial fibrillation: further analysis of the hemorrhagic transformation index. *Neurology, Neuropsychiatry, Psychosomatics*, vol. 11, no 2, pp. 12–21 (in Russian). DOI: 10.14412/2074-2711-2019-2-12-21.
15. Rumyantseva E.V., Furmanov K.K. (2016) Modeling mortgage survival. *Applied Econometrics*, vol. 41, no 1, pp. 123–143 (in Russian).
16. Hung H., Chiang C.T. (2010) Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian Journal of Statistics*, vol. 37, no 4, pp. 664–679.
17. Kamarudin A.N., Cox T., Kolamunnage-Dona R. (2017) Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology*, no 17, article no 53. DOI: 10.1186/s12874-017-0332-6.

18. Heagerty P.J., Zheng Y. (2005) Survival model predictive accuracy and ROC curves. *Biometrics*, vol. 61, no 1, pp. 92–105. DOI: https://doi.org/10.1111/j.0006-341X.2005.030814.x.

19. Cox D.R., Snell E.J. (1968) A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 30, no 2, pp. 248–275.

20. Nelson W. (1972) Theory and applications of hazard plotting for censored failure data. *Technometrics*, vol. 14, no 4, pp. 945–966. DOI: 10.1080/00401706.1972.10488991.

21. Aalen O. (1978) Nonparametric inference for a family of counting processes. *Annals of Statistics*, vol. 6, no 4, pp. 701–726.

22. Arzhenovsky S. (2006) Socioeconomic determinants of smoking in Russia. *Quantile*, no 1, pp. 81–100 (in Russian).

23. Rapakov G.G., Gorbunov V.A. (2015) Time-to-event analysis methods for demographics processing. *Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies*, no 4. pp. 110–120 (in Russian).

About the authors

Ekaterina V. Rumyantseva

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

Senior Lecturer, Department of Applied Economics, National Research University Higher School of Economics,

20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: evrumyantseva@hse.ru

Kirill K. Furmanov

Cand. Sci. (Econ.);

Assistant Professor, Department of Applied Economics, National Research University Higher School of Economics,

20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: kfurmanov@hse.ru

ORCID: 0000-0002-3433-9497

DOI: 10.17323/2587-814X.2021.1.19.29

Risk-based efficiency assessment of information systems

Eugeni A. Isaev[a]
E-mail: is@itaec.ru

Dmitry V. Pervukhin[b]
E-mail: dvperv@gmail.com

Georgy O. Rytikov[b,c]
E-mail: GR-yandex@yandex.ru

Ekaterina K. Filyugina[d]
E-mail: ekaterina.filyugina@mail.ru

Diana A. Hayrapetyan[e]
E-mail: hayrapetyandiana@gmail.com

[a] Institute of Mathematical Problems of Biology, Russian Academy of Sciences (Branch of the Keldysh Institute of Applied Mathematics, Russian Academy of Sciences)
Address: 1, Professor Vitkevich Street, Pushino, Moscow Region 142290, Russia

[b] State University of Management
Address: 99, Ryazansky Prospect, Moscow 109542, Russia

[c] Moscow Polytechnic University
Address: 38, Bolshaya Semyonovskaya Street, Moscow 107023, Russia

[d] Impact Electronics Ltd.
Address: 14/19 build. 8, Novoslobodskaya Street, Moscow 127055, Russia

[e] Epokha Vozrozdeniya Ltd.
Address: 13 build. 2, Rusakovskaya Street, Moscow 107078, Russia

Abstract

The implementation of information systems is aimed at improving the financial performance of a company, creating a transparent reporting system and improving many other competitive factors. However, the acquisition of these benefits does not negate the complexity of making a decision

whether or not to implement a particular IT project. The total cost of ownership of the information system throughout the life cycle is usually not considered in comparison with the expected benefits from the use of the system, due to the uncertainty of such benefits. Comparative certainty of approaches and methods is present only in terms of costs, both for a priori (planned) and a posteriori (actual) assessment. It is possible to determine both capital and operating costs accurately enough. Indirect definition of the positive influence of an information system on the activity of the organization also seems possible. However, there are currently no generally recognized methods for analyzing the expected positive effect of an IT project. At the same time, large companies, in accordance with the requirements of the respective regulators and / or due to internal management considerations, build a risk management system to determine the level of capabilities, losses and to prevent adverse events. This study considers the feasibility of an approach to analyze the effectiveness of the implementation of the information system on the basis of the company's risk reduction, leading to a decrease in economic benefits. It takes into account the internal risks of the information system that occur during the installation of the system, its operation and the termination of work with the system.

Key words: risk assessment; IT project management; information system; implementation of information systems.

Citation: Isaev E.A., Pervukhin D.V., Rytikov G.O., Filyugina E.K., Hayrapetyan D.A. (2021) Risk-based efficiency assessment of information systems. *Business Informatics*, vol. 15, no 1, pp. 19–29.
DOI: 10.17323/2587-814X.2021.1.19.29

Introduction

Modern business conditions imply a regime of fierce competition and increasingly shrinking time for decision-making, which is emphasized in many scientific papers [1]. A large number of empirical studies have been conducted confirming the influence of factors such as organizational architecture, production infrastructure and related business processes on the ability of an enterprise to survive and function effectively [2]. At the same time, one of the key aspects of successful company management is the use of information technology and modern software tools, as well as appropriate methods and models (e.g., convolutional neural networks, significantly accelerating the processing of large data sets [3]).

In continuation of the authors' works in the field of evaluating the effectiveness of implementation of IT projects [4], it is worth mentioning the modern hierarchical system of existing classes of information systems (IS). In terms of enterprise architecture, when classifying IS "top-down", first we should mention the group of systems designed to provide operational analytics (examples include SAP HANA, Lumira, Predictive Analytics). This block is followed by ERP-systems that automate individual business processes and support financial and economic management. Operational translation of market needs into specific production tasks is provided by a block of MES, PLM and SCADA systems, where the first is responsible for production in general, the second for product lifecycle management, and the third for the quality of individual production iterations.

The implementation of each of these systems involves the planning and implementation of an investment project [5]. The ability to assess the feasibility of a project and its relevance becomes a critical task for a company [6, 7]. At the same time, the introduction of information technology, along with certain expectations, is associated with certain risks[1] [8]. The positive consequences of the introduction of IS include a well-ordered organizational structure with transparent and uniform reporting, acceleration of the process of analyzing the company's activity with further adoption of strategically important managerial decisions, and automation of many business processes. The negative effect on the key indicators of the company is caused by the cost of the implementation of the IS and the necessary structural reorganization [9], support of the system during its operation, as well as relevant updates of the software product.

To date, the assessment of the effectiveness of the implementation of the IS is still an issue. Along with well-known approaches such as IE (Information Economics), TEI (Total Economic Impact), REJ (Rapid Economic Justification) and BSC (Balanced Scorecard), the standard numerical indicators ROI (return of investment), NPV (net present value), IRR (internal rate of return), EVA (economic value added), ROV (real options valuations) are used to consider projects. The right choice of IT project management methodology is also necessary [10]. As for this paper, it offers an approach to analyze the effectiveness of the implemented information system based on an assessment of the company's risks before and after the implementation of the IT project. An important feature is that it takes into account the internal risks associated with the information system itself and the likelihood of critical failures in its operation [11].

1. Impact of informational systems on company risks

The risk management system of large companies provides for the formation of a risk map that can be used, among other things, to determine the positive effect of the introduction of information systems [12]. When determining the impact of risks on the company's performance indicators, attention is paid both to the strength of the negative impact of an individual event on certain indicators and the frequency of certain adverse events occurring.

The influence of risks on company indicators can be estimated in a context of two situations – "as is", i.e. before realization of certain measures directed at reducing risks (for example, before introduction of information system), and "as will be", i.e. after realization of corresponding measures. Thus, the effect of the implementation of the measure (implementation of the IS) can be expressed in the form of the difference of values of the same indicators in the situations "as is" and "as will be".

There are many examples that support a positive conclusion that information systems can reduce a company's risks. In particular, one of the most noticeable factors affecting the key indicators of the company is the adoption of balanced and reasonable decisions, including those made at the senior management level [13]. It is management mistakes that can lead to the disruption of existing business processes. Possible sources of economically unjustified manage-

[1] The terms "risk" and "risk management" are used hereinafter within the terminology established in the field of risk management and defined, for example, in standards such as ISO 31000, FERMA, Basel, etc.

ment decisions may be both incorrect data, on the basis of which the decision is made, and human factor [14]. The implemented system allows you to significantly improve the methods of collecting and analyzing data, and reduce (up to complete elimination) their manual processing by employees of the company. The results of the analysis of information using algorithms inherent in the IS, provide impartiality and credibility in the formation of reporting, which increases the reliability of the indicators on the basis of which decisions will be taken.

The success stories of companies that have coped with the problem of illiquid goods (or significantly reduced it) by changing the structure of production and the subsequent reduction of warehouse space [4] show the effectiveness of implemented information technology, using the method described in this paper. The risks that can arise from errors in inventory management are mitigated because the need to purchase a substantial number of materials is reduced. At the same time, the need for rapid interaction within the company and with its contractors increases. Getting data in real time becomes a critical factor, which is enabled by information systems.

Another critical factor that has the potential for a negative outcome is the technological side of production. IT solutions provide an opportunity to fix the chain of manufacturing operations within the company [4]. At the same time there is tracking and standardization of activities on interaction with contractors. This allows for the automatic generation of signals on the management of production facilities.

Such examples are given in order to expand the functionality of companies' risk management systems. Consideration of problematic issues, the solution of which is associated with the implementation of the IS, can be assessed on the basis of changes in key indicators as a result of the introduction of a particular information system. This approach covers the issue of assessing the effectiveness of the IT project. However, at the same time a new problem arises: the fact that the risks are not only outside the information system, their source may be the project itself. Thus, to improve the assessment of the project, it is necessary to adjust for the risks arising, for example, in cases of problems during the implementation of the IS, or critical failures during the use of the system, as well as during its decommissioning [15].

2. Internal risks of information systems

According to the available research on ERP-systems efficiency, 51% of implementation projects experience some unforeseen difficulties in the process of IS installation, 53% of projects demonstrate significant financial difficulties, which exceed the initially approved budget [16], 83% of projects fail to meet the deadline, 42% of projects fail to complete the expected characteristics of the implemented technical solution [17, 18], 40% of projects fail to solve the business task after completion and putting the IS into operation [19, 20]. These statistics indicate that managers and investors recognize the presence of significant risks in the implementation of IS. Typically, implementation projects (as well as subsequent system support) are outsourced to external consultants, but this business model raises issues related to conflicts of interest [21] and blurring responsibility for results, which aggravates the potential overall negative effect of project risks [22].

To reduce risk exposure associated with software, system implementation, use, and decommissioning, a number of approaches have been developed to identify problem-

atic aspects [23]. Some of these approaches are accounted for in ALM-class information systems (e.g., SAP Cloud ALM, Solution Manager), which computerize the lifecycle management process. In general, the risks of IT projects are technical in nature and related to the formulation of requirements, their variability and the speed of updates. At the same time, the most important are the technical aspects, without the implementation of which it is not possible to make the implementation of an information system [24]. Although a group of risks, the identification of which affects the business processes of the company, should be highlighted. The integration of IS with production is often accompanied by a sharp drop in labor efficiency [25, 26]. Thus, the development of proper strategies and measures to solve this problem is required [27].

Identification of IT project risks is an important effort, which has been carried out by ISACA. The result of this activity has become a document COBIT5 [28], which categorizes the risks in practice into 111 categories. Examples include information leakage risks [29], as well as risks associated with the project life cycle, IS architecture, company infrastructure, software vendor selection, as well as the personnel and their competencies.

The description of risks implies their division into areas of influence. The first type is considered strategic, describing missed opportunities to use information technology to improve the company's efficiency [30]. The second type is directly related to the technical implementation of the IT project. The last type is operational, which corresponds to the operation of the system and the technical support services. These risks, as a matter of fact, are divided by the degree of impact on the company's activities in case of their implementation.

Thus, the assessment of risks in the "as it will be" situation, i.e. after the implementation of the system, should include the internal risks of the information system directly related to its implementation and use.

3. Evaluation of information systems' impact on company risks and performance indicators

Using the example of process management, let's consider the issues of evaluating the impact of information systems on the company's risks (and, accordingly, on its key performance indicators), taking into account the internal risks associated with the information systems themselves.

Let's consider a process consisting of several consecutive stages, along the implementation of which there is an increase of some selected performance indicator. If the performance indicator s is the only one, then its values at different stages of the process can be represented as a vector $\overline{\mathbf{s}} = \{s_i\}$, $i = \overline{0, N}$, where N is the number of process stages. At the same time, s_0 represents the initial value of the index (i.e., occurring before the beginning of the first stage), s_i is the value of the indicator after the completion of the i-th stage of the process. Respectively, s_N is the value of the indicator after the last, N-th stage of the process, i.e. after completion of the whole process.

Now let's introduce the concept of the coefficient of dynamics, which characterizes the positive dynamics of the efficiency indicator under consideration as certain stages of the process are implemented. For an arbitrary i-th stage of the process the value of the coefficient of dynamics is determined as: $k_i = s_i / s_{i-1}$, where s_i is the value of the performance indicator after the completion of the i-th stage of the process, and s_{i-1} is the value of the same indicator before the beginning

of the stage (i.e. after the completion of the preceding stage). In fact, the coefficient of dynamics characterizes the ratio of "input" and "output" of the process. Given the fact that the process consists of several stages, the values of the coefficient of dynamics is a vector $\bar{\mathbf{k}} = \{k_i\}$, $i = \overline{1, N}$, where k_i is the value of the coefficient at the i-th stage of the process, N is the number of stages of the process. Since the considered coefficient of dynamics does not take into account either risks or the influence of the information system, we will call it the basic or risk-free one.

The correlation between the initial value of the performance indicator (before the start of the process) and its final value (after the full completion of the entire process) is also easy to derive:

$$s_N = s_0 \cdot \prod_{i=1}^{N} k_i ,$$

where s_0 and s_N are the initial and final values of the performance indicator, respectively;

k_i is the value of the basic coefficient of dynamics at the i-th stage of the process;

N is the number of stages in the process.

Now, let's make an assumption about process-related general risks (information systems, their impact on the process and their internal risks are not considered yet).

Let's define i-th stage risk realization probability p_i as the probability that this stage of the process will not be executed and, consequently, the expected results will not be achieved. In this case, the probability that the result of the i-th stage will be obtained as $(1 - p_i)$. For example, if the proportion of manufacturing defects in the performance of the i-th technological operation for all possible reasons is 1.3%, then $p_i = 0.013$, and the probability of successful performance of this operation will be $(1 - p_i) = 0.987$.

This allows us to determine the variation of the coefficient of dynamics considered above, by adjusting its values taking into account the risks (we will call this variation the coefficient of dynamics of the first kind). Its connection with the basic (risk-free) coefficient of dynamics is as:

$$k_i^{(1)} = k_i \cdot (1 - p_i),$$

where k_i is the value of the basic (risk-free) coefficient of dynamics at the i-th stage of the process;

k_i is the value of the coefficient of dynamics of the first kind at the i-th stage of the process;

p_i is the probability of risk realization at the i-th stage of the process.

Note that the coefficient of dynamics of the first kind, corresponding to certain stages of the process, is always lower than the values of the base (risk-free) coefficient, relating to the same stages.

Since the implementation of the information system should help to reduce the probability of internal risk realization, we will introduce such an indicator as the influence of the information system on risks. The value of this indicator, relating to the i-th stage of the process (r_i), is in the range [0, 1] and indicates that after the implementation of the information system, the probability of risk realization at the i-th stage will decrease and will be equal to $p_i \cdot (1 - r_i)$. In particular, if $r_i = 1$, then the implementation of the information system should completely eliminate the risk of failure of the i-th stage of the process, and if $r_i = 0$, then it means that the system has no effect on the risk of this stage.

Hence, another variation of the coefficient of dynamics can be introduced, taking into account the influence of the information system on risks (the coefficient of dynamics of the second kind). Its connection with

the basic (risk-free) coefficient of dynamics is as:

$$k_i^{(2)} = k_i \cdot (1 - p_i \cdot (1 - r_i)),$$

where k_i is the value of the basic (risk-free) dynamics coefficient at the *i*-th stage of the process;

$k_i^{(2)}$ is the value of the coefficient of dynamics of the second kind at the *i*-th stage of the process;

p_i is the probability of risk realization at the *i*-th stage of the process;

r_i is the impact of the information system on the *i*-th stage risks of the process.

Finally, let us take into account that the information system itself may fail in its operation. These risks are internal to the system. Their presence means that the real impact of the information system on the *i*-th stage risk of the process may be less than the value of indicator r_i. If we define the probability of internal risk of the information system at the *i*-th stage as q_i, then the impact of the system on the risk of this stage will be $r_i \cdot (1 - q_i)$. It is clear that with zero probability of internal system risk its impact on stage risk will not change, but non-zero values of this parameter will lead to a reduction of this impact, up to zero.

Thus, another variation of the coefficient of dynamics (the coefficient of dynamics of the third kind) will take into account all the considered parameters, including the probabilities of internal risks of the information system. The connection of the coefficient of dynamics of the third kind with the basic (risk-free) coefficient is as:

$$k_i^{(3)} = k_i \cdot (1 - p_i \cdot (1 - r_i \cdot (1 - q_i))),$$

where k_i is the value of the basic (risk-free) dynamics coefficient at the *i*-th stage of the process;

$k_i^{(3)}$ is the value of the coefficient of dynamics of the third kind at the *i*-th stage of the process;

p_i is the probability of risk realization at the *i*-th stage of the process;

r_i is the impact of the information system on the *i*-th stage risks of the process;

q_i is the probability of realization of internal information system risk at the *i*-th stage of the process.

Thus, the assessment of the information system efficiency aimed at reducing the risk of the *i*-th stage of the project can be expressed either as the difference between the values of the coefficients of dynamics of the third kind (after the introduction of the information system, taking into account its impact on the risk of the stage and internal risks of the system itself) and the first kind (before the introduction of the information system), or as the ratio of these values.

The following formulas can be used to estimate the result of the implementation of an ideal (containing no internal risks) and real (with internal risks) information system:

$$\Delta s^{ideal} = s_0 \cdot \left(\prod_{i=1}^{N} k_i^{(2)} - \prod_{i=1}^{N} k_i^{(1)} \right),$$

$$\Delta s^{real} = s_0 \cdot \left(\prod_{i=1}^{N} k_i^{(3)} - \prod_{i=1}^{N} k_i^{(1)} \right),$$

where Δs^{ideal} are Δs^{real} are the efficiency of the ideal and real information system, respectively;

s_0 is the initial value of performance indicator;

$k_i^{(1)}$, $k_i^{(2)}$, $k_i^{(3)}$ are the values of the coefficient of dynamics of the first, second and third kind, respectively, at the *i*-th stage of the process;

N is the number of stages of the process.

The given reasoning can be extended with the case when not one but several per-

formance indicators are considered for the process. In this case instead of a vector of indicator values we will have a matrix $\mathbf{S} = \{s_{ij}\}$, $i = \overline{0, N}$, $j = \overline{1, M}$, where s_{0j} is the initial value of the j-th indicator, s_{ij} is the value of the j-th indicator after the i-th stage of the process, M is the number of indicators, N is the number of process stages.

The values of the coefficients of dynamics will be determined separately for each of the performance indicators and will also represent a matrix. For example, the values of the basic dynamics coefficients will correspond to the matrix $\mathbf{K} = \{k_{ij}\}$, $i = \overline{1, N}$, $j = \overline{1, M}$, where s_{ij} is value of the j-th coefficient at the i-th stage of the process, M is the number of indicators, N is the number of stages of the process. For the coefficient of dynamics of the first, second and third kind there will be similar matrixes.

Conclusion

Increasing the efficiency of company activities can be achieved by implementing information systems, which is associated with appropriate changes in the organizational structure and organization of business processes. However, the risks associated with these actions cause us to think about the feasibility of the project. At the same time, problems from the production and management activities prior to the introduction of an information system also represent a certain threat to the company. Thus, it is possible to assess the IT project by taking into account both positive and negative consequences of the implementation of the system. ■

Acknowledgments

This work was funded by Ministry of Science and Higher Education of the Russian Federation (State assignment "Structure and properties of the polymer materials produced using a system of chemically, thermally and / or mechanically induced both surface and bulk modification techniques," topic number FZRR-2020-0024, code 0699-2020-0024).

References

1. Khanfar A.A., Mavi R.K., Jie F. (2018) Prioritizing critical failure factors of IT projects with fuzzy analytic hierarchy process. Proceedings of the *International Conference on Mathematics, Engineering and Industrial Applications (ICoMEIA), Kuala Lumpur, Malaysia, 24–26 July 2018*, vol. 2013, no 020058. DOI: 10.1063/1.5054257.
2. Sorooshian S., Mun S.Y. (2020) Literature review: Critical risk factors affecting information-technology projects. *Quality – Access to Success*, vol. 21, no 175, pp. 157–161.
3. Isaev E.A., Samodurov V.A., Pervukhin D.V., Filyugina E.K. (2019) The application of convolutional neural networks (CNN) to search for patterns of various nature in data series. *Industrial Automatic Control Systems and Controllers*, no 12, pp. 24–32 (in Russian). DOI: 10.25791/asu.12.2019.1069.
4. Pervukhin D.V., Isaev E.A., Rytikov G.O., Filyugina E.K., Hayrapetyan D.A. (2019) Analysis of the positive effect of the IT solutions implementation based on risk assessment. *Instruments and Systems: Monitoring, Control, and Diagnostics*, no 7, pp. 45–54 (in Russian). DOI: 10.25791/pribor.07.2019.742.
5. Ayuso S., Rodriguez M.A., Garcia-Castro R., Arino M.A. (2014) Maximizing stakeholders' interests: An empirical analysis of the stakeholder approach to corporate governance. *Business & Society*, vol. 53, no 3, pp. 414–439. DOI: 10.1177/0007650311433122.
6. Mishra A., Sinha K.K., Thirumalai S, Van de Ven A. (2020) Sourcing structures and the execution efficiency of information technology projects: A comparative evaluation using stochastic frontier analysis. *Journal of Operations Management*, vol. 66, no 3, pp. 281–309. DOI: 10.1002/joom.1064.

7. Klaus-Rosinska A. (2017) Concept of measuring the performance of IT projects. Proceedings of the *25th International Conference on Systems Engineering (ICSEng), Las Vegas, US, 22–24 August 2017*, pp. 412–417. DOI: 10.1109/ICSEng.2017.63.

8. Rodriguez A., Ortega D., Concepcion R. (2017) An intuitionistic method for the selection of a risk management approach to information technology projects. *Information Sciences*, vol. 375, pp. 202–218. DOI: 10.1016/j.ins.2016.09.053.

9. Shafiee S., Kristjansdottir K., Hvam L., Forz C. (2018) How to scope configuration projects and manage the knowledge they require. *Journal of Knowledge Management*, vol. 22, no 5, pp. 982–1014. DOI: 10.1108/JKM-01-2017-0017.

10. Pervukhin D.V., Isaev E.A., Rytikov G.O., Filyugina E.K., Hayrapetyan D.A. (2020) Theoretical comparative analysis of cascading, iterative, and hybrid approaches to IT project life cycle management. *Business Informatics*, vol. 14, no 1, pp. 32–40. DOI: 10.17323/2587-814X.2020.1.32.40.

11. Boehm B.W. (1991) Software risk management: Principles and practices. *IEEE Software*, vol. 8, no 1, pp. 32–41. DOI: 10.1109/52.62930.

12. Gregory P.H. (2010) *CISA certified information systems auditor all-in-one exam guide*. New York: McGraw-Hill.

13. Ako-Nai A., Singh A.M. (2019) Information technology governance framework for improving organizational performance. *South African Journal of Information Management*, vol. 21, no 1, article no a1010. DOI: 10.4102/sajim.v21i1.1010.

14. Ul Haq S., Gu D., Liang C., Abdullah I. (2019) Project governance mechanisms and the performance of software development projects: Moderating role of requirements risk. *International Journal of Project Management*, vol. 37, no 4, pp. 533–548. DOI: 10.1016/j.ijproman.2019.02.008.

15. Almutairi M., Riddle S. (2018) A framework for managing security risks of outsourced IT projects: An empirical study. Proceedings of the *International Conference on Software Engineering and Information Management (ICSIM 2018), Casablanca, Morocco, 4–6 January 2018*, pp. 40–44. DOI: 10.1145/3178461.3178476.

16. Bokolo A. (Jr.), Pa N.C., Nor R.N.H., Jusoh Y.Y., Aris T.N.M. (2018) Implementation of risk mitigation among IT governance practitioners in Malaysia. *Advanced Science Letters*, vol. 24, no 2, pp. 1344–1347. DOI: 10.1166/asl.2018.10746.

17. Lee J.S., Keil M., Shalev E. (2019) Seeing the trees or the forest? The effect of IT project managers' mental construal on IT project risk management activities. *Information System Research*, vol. 30, no 3, pp. 1051–1072. DOI: 10.1287/isre.2019.0853.

18. Lai S.-T., Leu F.-Y. (2017) A critical quality measurement model for managing and controlling big data project risks. Proceedings of the *12th IEEE International Conference on Broadband Wireless Computing, Communication and Applications (BWCCA), Barcelona, Spain, 8–10 November 2017*, vol. 12, pp. 777–787. DOI: 10.1007/978-3-319-69811-3_69.

19. Pike G. (2006) *Supporting business innovation while reducing technology risk*. White paper. Walldorf, Germany: SAP AG.

20. Wu D.J., Ding M., Hitt L.M. (2013) IT implementation contract design: Analytical and experimental investigation of IT value, learning, and contract structure. *Information Systems Research*, vol. 24, no 3, pp. 787–801. DOI: 10.1287/isre.1120.0448.

21. Ghribi S., Hudon P.A., Mazouz B. (2019) Risk factors in IT public–private partnership projects. *Public Works Management & Policy*, vol. 24, no 4, pp. 321–343. DOI: 10.1177/1087724X18823009.

22. Didraga O., Brandas C., Batagan L., Alecu F. (2019) Characteristics of effective IT project risk management in Romanian IT companies. *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no 4, pp. 176–193. DOI: 10.24818/18423264/53.4.19.11.

23. Pimchangthong D., Boonjing V. (2017) Effects of risk management practices on IT project success. *Management and Production Engineering Review*, vol. 8, no 1, pp. 30–37. DOI: 10.1515/mper-2017-0004.

24. Neumeier A., Radszuwill S., Garizy T.Z. (2018) Modeling project criticality in IT project portfolios. *International Journal of Project Management*, vol. 36, no 6, pp. 833–844. DOI: 10.1016/j.ijproman.2018.04.005.

25. Maruping L.M., Venkatesh V., Thong J.Y.L., Zhang X. (2019) A risk mitigation framework for information technology projects: A cultural contingency perspective. *Journal of Management Information Systems*, vol. 36, no 1, pp. 120–157. DOI: 10.1080/07421222.2018.1550555.

26. Lee J.S., Keil M. (2018) The effects of relative and criticism-based performance appraisals on task-level escalation in an IT project: a laboratory experiment. *European Journal of Information Systems*, vol. 27, no 5, pp. 551–569. DOI: 10.1080/0960085X.2017.1408752.

27. Gloria I. (Jr.), Chaves M.S. (2017) Identification and mitigation of risks in IT projects: a case study during the merger period in the telecommunications industry. *Revista de Gestao e Projetos*, vol. 8, no 3, pp. 1–17. DOI: 10.5585/gep.v8i3.581.

28. ISACA (2012) *COBIT5. A business framework for the governance and management of enterprise IT*. ISACA.

29. Almutairi M., Riddle S. (2018) Managing outsourced IT projects' security risks: A case study. Proceedings of the *10th International Conference on Information Management and Engineering (ICIME 2018), Manchester, England, 22–24 September 2018*, pp. 21–26. DOI: 10.1145/3285957.3285986.

30. Thirasakthana M., Kiattisin S. (2018) Identifying standard testing time for estimation improvement in IT project management. Proceedings of the *3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), Bangkok, Thailand, 12–14 December 2018*, pp. 1–5. DOI: 10.1109/TIMES-iCON.2018.8621787.

About the authors

Eugeni A. Isaev
Cand. Sci. (Tech.);
Senior Researcher, Institute of Mathematical Problems of Biology, Russian Academy of Sciences (Branch of the Keldysh Institute of Applied Mathematics, Russian Academy of Sciences), 1, Professor Vitkevich Street, Pushino, Moscow Region 142290, Russia;
E-mail: is@itaec.ru
ORCID: 0000-0002-3703-447X

Dmitry V. Pervukhin
Senior Lecturer, State University of Management, 99, Ryazansky Prospect, Moscow 109542, Russia;
E-mail: dvperv@gmail.com
ORCID: 0000-0001-6500-035X

Georgy O. Rytikov
Cand. Sci. (Phys.-Math.);
Associate Professor, State University of Management, 99, Ryazansky Prospect, Moscow 109542, Russia;
Associate Professor, Moscow Polytechnic University, 38, Bolshaya Semyonovskaya Street, Moscow 107023, Russia;
E-mail: GR-yandex@yandex.ru
ORCID: 0000-0001-5521-8662

Ekaterina K. Filyugina
Impact Electronics Ltd., 14/19 build. 8, Novoslobodskaya Street, Moscow 127055, Russia;
E-mail: ekaterina.filyugina@mail.ru
ORCID: 0000-0001-6461-7235

Diana A. Hayrapetyan
Economist, Epokha Vozrozdeniya Ltd., 13 build. 2, Rusakovskaya Street, Moscow 107078, Russia;
E-mail: hayrapetyandiana@gmail.com
ORCID: 0000-0001-6646-1748

DOI: 10.17323/2587-814X.2021.1.30.46

Trends in data mining research: A two-decade review using topic analysis

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

Ekaterina A. Anisichkina
E-mail: eaanisichkina@edu.hse.ru

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

Abstract

This work analyzes the intellectual structure of data mining as a scientific discipline. To do this, we use topic analysis (namely, latent Dirichlet allocation, LDA) applied to the proceedings of the International Conference on Data Mining (ICDM) for 2001–2019. Using this technique, we identified the nine most significant research flows. For each topic, we analyze the dynamics of its popularity (number of publications) and influence (number of citations). The central topic, which unites all other direction, is General Learning, which includes machine learning algorithms. About 20% of the research efforts were spent on the development of this direction for the entire time under review, however, its influence has declined most recently. The analysis also showed that attention to topics such as Pattern Mining (detecting associations) and Segmentation (object separation algorithms such as clustering) is decreasing. At the same time, the popularity of research related to Recommender Systems, Network Analysis, and Human Behaviour Analysis is growing, which is most likely due to the increasing availability of data and the practical value of these topics. The research direction related to practical Applications of data mining also shows a tendency to grow. The last two topics, Text Mining and Data Streams have attracted steady interest from researchers. The results presented here shed light on the structure and trends of data mining over the past twenty years and allow us to expand our understanding of this scientific discipline. We can argue that in the last five years a new research agenda has been formed, which is characterized by a shift in interest from algorithms to practical applications that affect all aspects of human activity.

Key words: data mining topics, topic analysis, scientometrics.

Citation: Zelenkov Yu.A., Anisichkina E.A. (2021) Trends in data mining research: A two-decade review using topic analysis. *Business Informatics*, vol. 15, no 1, pp. 30–46. DOI: 10.17323/2587-814X.2021.1.30.46

Introduction

The term "data mining" (DM) appeared in the 1960s to describe the search for correlations without an a priori hypothesis [1]. According to the widely accepted definition that is used in many textbooks now, data mining is the extraction of implicit, previously unknown, and potentially useful information from data [2, 3]. Besides, Rather [4] defines data mining as a combination of three easy concepts:

✦ statistics that include the classical descriptive tools, e.g. degrees of freedom, *F*-ratios, and *p*-values, but exclude inferential conclusions;

✦ big data as an umbrella term for datasets of any size with the accent on big size since a tremendous amount of data impacts almost every aspect of our lives;

✦ machine learning (ML), i.e. tools to build computer programs that sift through databases automatically, seeking regularities or patterns [2].

Statistics and machine learning provide the technical basis of data mining. They are used to extract information from the raw data. Some authors also view DM as part of the process for knowledge discovery from data (KDD). This process may include techniques such as data preprocessing (cleaning and integration), data storage, online analytical processing, data cubes, etc. [3].

As follows from these definitions, data mining is a scientific discipline that combines achievements in several areas of research. The structure of any scientific discipline can be represented as a set of evolving topics, i.e. significant, implicit associations hidden in fragmented knowledge areas. Trends in these topics (for example, a change in the number of publications and their citation) reflect a shift in the interests of the research community. In particular, the study of this dynamic allows us to determine the most relevant areas of research in the present and extrapolate them in the near future. In addition, understanding the fundamental shifts in the interests of researchers helps us to determine the place of the studied discipline in the general body of human knowledge, its interaction with other disciplines and the overall contribution to human progress.

The traditional method of studying the structure of a scientific discipline is survey or review. However, due to the interdisciplinary nature of data mining, there are practically no reviews considering DM as a single discipline (yet it should be noted that surveys of narrower topics are published continuously).

In a review paper published in 2006, Yang and Wu [5] noted that data mining had achieved tremendous success. However, there is still a lack of timely exchange of essential topics in the community as a whole. Authors of [5] ranked the ten most important problems in DM:

✦ developing a unifying theory of data mining;

✦ scaling up for high dimensional data and high-speed data streams;

✦ mining sequential data and time series data;

✦ mining complex knowledge from complex data;

✦ data mining in graph-structured data;

✦ distributed data mining and mining multi-agent data;

✦ data mining for biological and environmental problems;

✦ data mining process-related problems;

✦ security, privacy, and data integrity;

✦ dealing with non-static, unbalanced and cost-sensitive data.

These problems divide the overall DM research flow into smaller, more focused segments. In 2010, Wu provided additional comments on these challenging issues [6], and they were the subject of discussion in a special panel at the 10th International Conference on Data Mining (ICDM).

Yang and Wu [5] view the development of a unified theory of DM as the most critical issue. It should be a theoretical framework that unifies different techniques designed for individual problems, including clustering, classification, association rules, etc., as well as different data mining technologies (such as statistics, machine learning, database systems, etc.). It should help the field and provide a basis for future research.

Most of the identified problems relate to algorithms for working with data types that became relevant in the 2000s (ultra-high dimensional data, high-speed data streams, time series, networks and other complex data). The authors of [5, 6] consider ecological and environmental informatics as the most important area of DM applications.

In addition to the analysis of critical challenges, work [6] presents a list of the most important topics of data mining (*Table 1*). This list was obtained based on expert opinions; hence, it can serve as a reference to the structure of the scientific discipline. However, the expert-based approach does not provide quantitative metrics that measure the relative importance of various topics and their change over time.

Liao et al. [7] presented a review of the literature on data mining techniques and applications from January 2000 to August 2011. They selected 216 articles from 159 academic journals using keywords like 'data mining,' 'decision tree,' 'artificial neural network,' 'clustering,' etc. Based on papers selected, they identified nine categories of DM techniques (systems optimization, knowledge-based systems, modeling, algorithm architecture, neural networks, etc.). In addition, the authors of [7] presented the essential trends in data mining. According to the results presented, the most important trend is the Association Rules (rank 5), followed by Neural Networks (rank 4) and then Classification and Support Vector Machines (both have rank 3). The authors do not describe the method for ranking; however, we can assume that it is based on counting the number of references on each technique in the analyzed corpus of publication.

Table 1.

Top 10 data mining topics [6]

No	Topic
1	Classification (including C4.5, CART, kNN, and Naive Bayes)
2	Statistical learning (SVM and mixture models)
3	Association analysis
4	Link mining (e.g. PageRank algorithm)
5	Clustering
6	Bagging and boosting
7	Sequential patterns
8	Integrated mining (e.g. integrating classification and association rule mining)
9	Rough sets
10	Graph mining

To the best of our knowledge, the publications cited above are the only ones that examine the dynamics of data mining as a single scientific discipline. As already noted, they are based on subjective assessments.

The idea of our work is to apply formal methods of topic analysis to publications in the field of data mining. As an object of analysis, we use the proceedings of the International Conference on Data Mining (ICDM), which has been held annually since 2001.

1. Data

The International Conference on Data Mining (ICDM) is a top conference that, along with the SIGKDD Conference on Knowledge

Discovery and Data Mining (KDD), ACM International Conference on Web Search and Data Mining (WSDM) and a few others, forms a network of major forums in the field of data mining and knowledge discovery from data. The Web of Science (WoS) database contains information on 5120 publications of the main ICDM tracks and related workshops. *Figure 1* represents the time distribution of these publications.

The WoS database contains such data as the authors, title of publication, abstract, and the number of citations that are necessary for our study.

2. Research method

One of the most popular techniques of bibliometric networks analysis is term-level coupling implemented in VOSviewer software [8]. This approach allows us to identify clusters of terms that can be viewed as more or less stable implicit structures shaping the scientific discipline. Authors of a review of literature-based discovery [9] list the main computation techniques that automate the knowledge discovery process. They noted that topic modeling that allows observing how topic-level information is propagated among documents provides more deep insight of document corpora than term-level analysis. However, topic modeling is still relatively rarely used in literature analysis [9, 10].

Mann et al. [11] used a combination of topic modeling and citation analysis to estimate the impact factor of the topic over time and topical diversity of documents in computer science. Dam and Ghose [12] used topic modeling to analyze the content of the proceedings of the International Conference on Principles and Practices of Multi-Agent Systems (PRIMA). Among recent works, Zelenkov [10] applied topic analysis to the knowledge management area. The last paper pays special attention to the topic dynamics, i.e. how a number of publications and citations regarding each topic

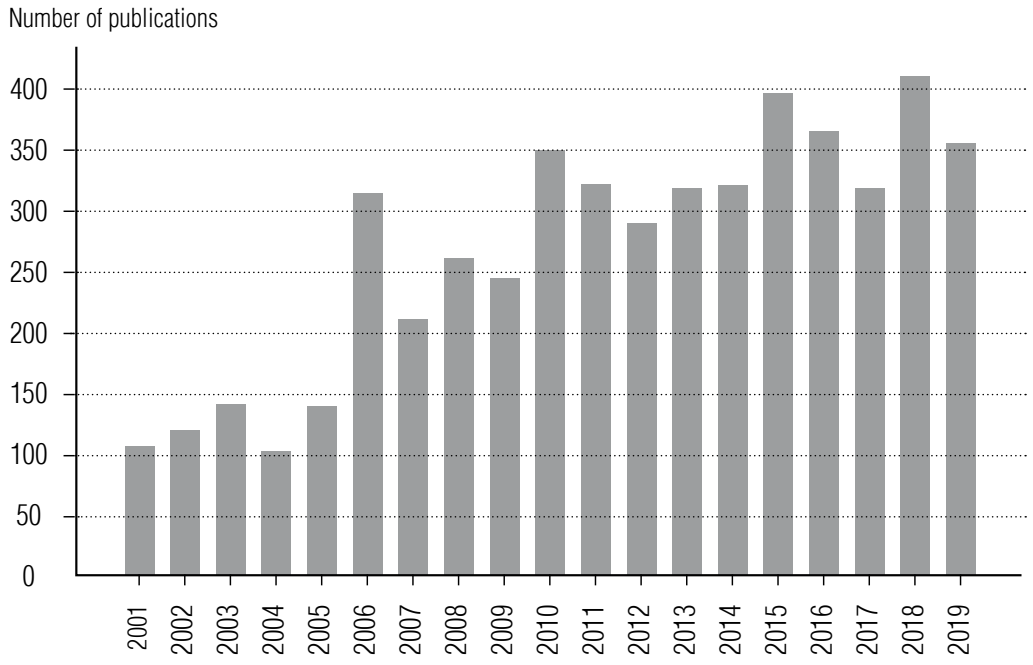


Fig. 1. Distribution of ICDM publications

changes in time. It helps shed light on the shift in research interest and identify critical trends of the present time.

Yet another application of topic analysis is presented in [13], where it is used to quantify the similarity and evolution of scientific disciplines. In [14] the authors propose topic evolution trees generated from the heterogeneous bibliographic network.

A topic is a set of words that often co-occur in texts related to a given subject area. Probabilistic topic modeling is based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over terms.

Let there be a finite set of topics T, which is not known. Each use of the term w in document d is associated with some topic $t \in T$. Thus, a collection of documents is considered as a set of triples (d, w, t) selected randomly and independently from the distribution defined on a finite set $D \times W \times T$. Documents $d \in D$ and the terms $w \in W$ are observable variables. The topics $t \in T$ are latent variables that must be defined.

The topic model automatically detects latent topics by the observed frequencies of words in the documents:

$$p(w|d) = \sum_{t \in T} p(t|d)\, p(w|t).$$

Thus, the input of the algorithm is a matrix $D \times W$, which cells contain counts of the word w in document d.

To prepare matrix $D \times W$, we used abstracts of 5120 papers downloaded from the Web of Science database, as described in the previous section. According to [15], differences between abstract and full-text data are more apparent within small document collections. Therefore, we have selected abstracts as an object of analysis.

According to the general text mining technique, abstracts were tokenized, and the terms obtained were converted to standard form. Next, words that belong to an extended stop word list were deleted. The extended stop-word list includes standard English stop-words and corpus-specific words that appear in less than 5% and more than 60% of documents. We also created bigrams to join terms often co-occurred beside. As a result, we got a sparse matrix $D \times W$ with dimensions of 5120 × 1000, only 1.62% of the cells of which contain values greater than zero.

To compute the topics, we used a latent Dirichlet allocation (LDA) algorithm that is based on the additional assumption that the distribution Θ of documents θ_d and distribution Φ of topics φ_t are spawn by Dirichlet distributions [16]. To build the model, one should define a number of topics $|T|$; the LDA algorithm computes distributions Θ and Φ. As a result, each topic is presented by the weighted list of words; the weight of a word corresponds to its importance in the topic definition. The weighted list of topics presents each document; the weight of the topic corresponds to its significance in the document.

Determining the number of topics is a critical issue in topic analysis; many authors use various kinds of grid search optimizing a specific metric [10]. We used more advanced techniques, namely, Bayesian optimization [17]. Such an approach allows us to optimize simultaneously not only the number of the topics and also parameters of Θ and Φ distributions and other parameters of the algorithm. The optimization target is a perplexity which measures the convergence of a model with a given vocabulary W:

$$P(D) = \exp\left[-\frac{1}{N}\sum_{d \in D}\sum_{w \in d} n_{dw} \ln p(w|d)\right].$$

The perplexity of collection D is a measure of the language quality and is often used in computational linguistics. In our case, language is the distribution of words in documents $p(w|d)$. The less perplexity, the more uneven this distribution.

In general, the approach presented satisfies the guidelines to LDA users presented in [18].

An additional metric that we use to assess the quality of the model is diversity, i.e. the entropy of the distribution of words that characterize the topic:

$$H_t = -\frac{1}{\ln n_w}\sum_i^{n_w} p_t(w_i)\ln p_t(w_i), \qquad (1)$$

where n_w is the number of words describing topics;

$p_t(w_i)$ is the weight of i-th word in the topic t.

Since this metric is normalized by the number of features (words), it's possible values are in the range [0; 1]. The value 0 corresponds to the maximum focus when only one term describes the topic. Value 1 determines the situation when all the features are present in the description of the topic with the same weights, i.e. it is not identified. In a valuable model, the values of this metric should be the small and approximately the same for all topics.

When the optimal number of topics and corresponding topic distribution for each document are found, we can study topic dynamics. Let θ_{dt} is the weight of topic t in document d ($0 \le \theta_{dt} \le 1$). So, the overall popularity of topic across all documents can be defined as [10]:

$$\hat{\theta}_t = \frac{1}{|D|}\sum_{d\in D}\theta_{dt}. \qquad (2)$$

To measure the topic popularity in a particular year y it is enough to set $D = D_y$ in (2), where D_y is the set of all papers in year y.

Let C_d is the number of citations of document d and $C = \sum_{d\in D} C_d$. An impact of the topic can be defined as [10]:

$$\hat{i}_t = \frac{1}{C}\sum_{d\in D}\theta_{dt}C_d \ . \qquad (3)$$

By analogy, to obtain the topic impact in the particular year, one should set $D = D_y$ in (3).

3. Results and discussion

Performing all preprocessing operations described in the previous Section and 100 iterations of Bayesian optimization of the LDA model, we found that the optimal number of topics is 9, and the corresponding value of perplexity is 568.75.

Analyzing the dominant terms (*Figure 2*), we can conclude that each topic represents some coherent area of research. The weights of topics in documents are either large (i.e. the topic is strongly related) or near zero (i.e. the topic is unrelated).

Thus, to assign the labels, we analyzed the term distributions and most representative papers for each topic. To select the most representative papers, we sorted the publications by the topic weight and next by the number of citations, both in descending order. *Figure 2* presents the labels assigned, and *Table 2* lists the topics description.

Table 2 also presents the values of diversity, popularity and impact for each topic in the entire collection D, calculated in accordance with (1), (2) and (3), respectively. Please note that the sum of both popularity and impact is 1, so the values presented can be considered as a share of a particular topic in the total flow of data mining research, i.e. its total weight.

Additional useful information can be obtained from the analysis of the distribution of the topics' weights in the document corpora (*Figure 3*). For a more effective presentation, we excluded from the graph for each topic documents in which the weight of this topic is 0.

As follows from *Table 2*, the topic that has attracted the most attention over the past 20 years is General Learning. More than 20% of efforts in the field of data mining were spent in this direction. The works of this direction cover the widest spectrum of machine learning issues, e.g. the features selection [19] (the weight of dominant topics in this document is θ_{dt} =0.974),

Fig. 2. Visualization of the topic model using word clouds
(each word cloud represents one detected topic where the size
of words indicates the relevance of each word to that particular topic)

multi-label classification using ensembles [20] (θ_{dt} = 0.963), gradient methods [21] (θ_{dt} = 0.925), etc. These example papers were selected since they all have a large number of citations (more than 90, according to WoS). Interestingly, proceedings with the maximum weight of this topic are mainly devoted to kernel methods, e.g. [22] with θ_{dt} = 0.990. Note, that according to *Figure 3*, this topic has a weight close to 1 in the largest number of documents (more than 100). These articles focus solely on machine learning methods and do not overlap with other topics.

We defined the second important topic ($\hat{\theta}_t$ = 0.121) as Human Behavior Analysis because it focuses on the detection and prediction of patterns in the activities of groups of people and systems that these groups influence. In this area, issues are studied, such as the effect of price promotions [23] (θ_{dt} = 0.988), finding suspicious financial transactions [24] (θ_{dt} = 0.985), bitcoin volatility [25] (θ_{dt} = 0.985), and others. Note that a significant part of these works is presented at the workshops accompanying the main conference.

The next topic is Pattern Mining, as it focuses on association (rule) extraction, i.e., on the task of finding correlations between items in a dataset. Researchers study as a practical application of association rules (e.g. market basket data) and general features of patterns found in large databases. On the one hand, it can be the

Table 2.

Topics of Data Mining

Topic	Comments	Diversity	Popularity	Impact
Text Mining	Pattern detection in texts	0.779	0.107	0.110
General Learning	Machine learning algorithms and related methods like feature selection, class labeling, etc.	0.826	0.213	0.211
Segmentation	Methods based on object separation techniques: clustering, outlier detection, etc.	0.777	0.084	0.080
Applications	Practical use of data mining methods	0.826	0.097	0.095
Data Streams	Time–dependent models	0.805	0.097	0.102
Recommender systems	Algorithms that provide useful and explainable recommendations	0.799	0.076	0.079
Pattern Mining	General issues of finding correlations between items in data	0.750	0.110	0.114
Network Analysis	Community and influence flow detection in various networks	0.762	0.093	0.111
Human Behavior Analysis	Detection and prediction of patterns in the people's behavior: customer churn, market segmentation, fraud and security threats, etc.	0.844	0.121	0.096

identification of maximal frequent itemsets, i.e. an itemset that occurs in at least a systematic and realistic set of experiments [26] (θ_{dt} = 0.960). On the other hand, it can be a pattern consisting of infrequent, but highly correlated items rather than ones that occur frequently [27] (θ_{dt} = 0.987). Note that it is the most focused topic (with the lowest value of H_t).

The most representative proceedings of the Text Mining topic consider issues such as the identification and ranking of authors [28] (θ_{dt} = 0.974), topic modeling [29] (θ_{dt} = 0.969), and text clustering using semantics-based models [30] (θ_{dt} = 0.988). This is a research area with clear boundaries, which includes pattern detection in texts only and does not consider other types of unstructured data.

The research flow, which we call Data Streams, concerns time-dependent models. It includes more or less traditional analysis and prediction of time series with concept drift [31] (θ_{dt} = 0.981), and relatively more rare models, e.g. ones based on Granger causality [32] (θ_{dt} = 0.987).

The Applications topic combines works mainly devoted to the practical use of data mining methods, and which do not apply to other directions highlighted above. Examples are the detection of events using the co-locations of mobile users [33] (θ_{dt} = 0.987) and biometric security model for medical Internet of Things [34] (θ_{dt} = 0.983).

The research direction Network Analysis deals with graph models allowing us to restore the spatial structure or topology of the investi-

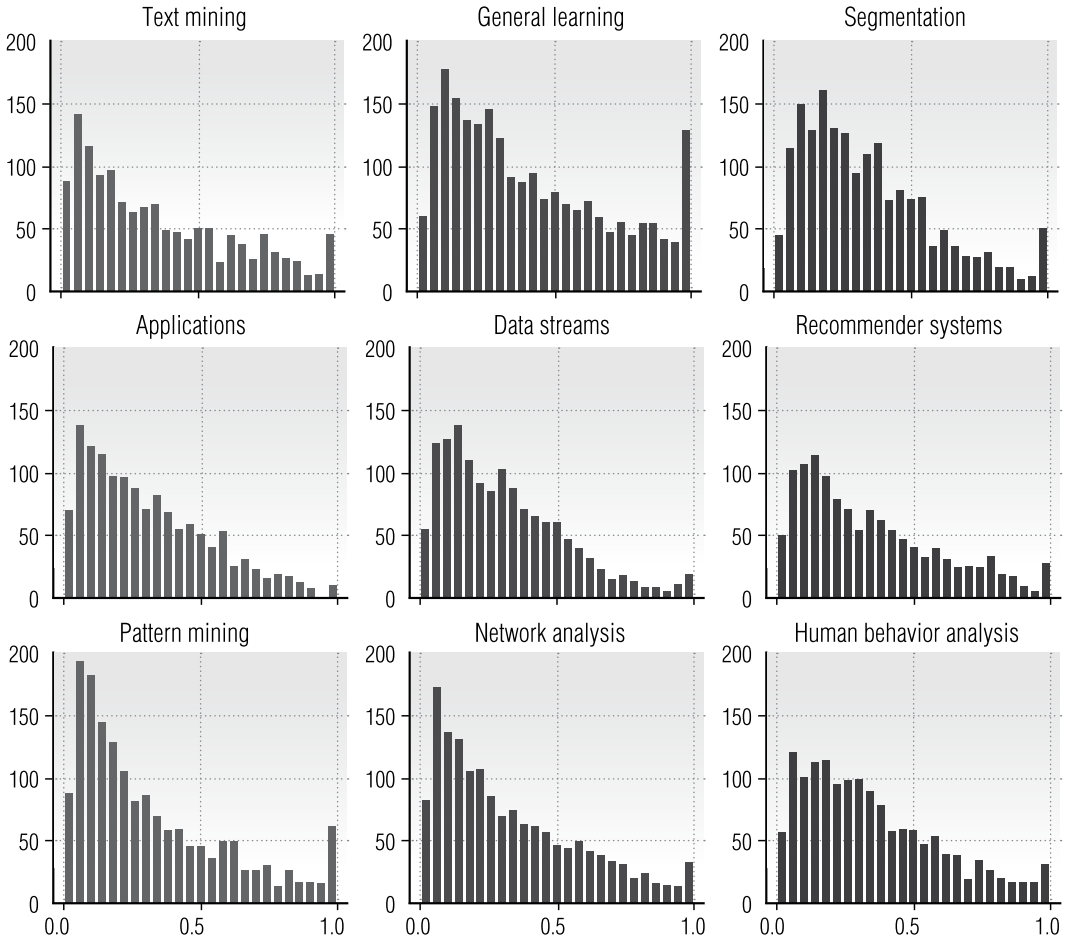


Fig. 3. Distribution of the topics' weights in the document corpora
(the horizontal axis shows the weights, the vertical axis shows the number of documents)

gated object. The most popular topic of such a kind of research is community detection using various methods of network analysis [35] (θ_{dt} = 0.983). The next issue attracting growing attention recently is the analysis of influence flows [36] (θ_{dt} = 0.985), including predictions of the popularity of messages in social networks.

We labeled the next topic as Segmentation since it includes not only a wide spectrum of clustering algorithms [37] (θ_{dt} = 0.977) but applications that are based on object separation techniques, e.g. outlier detection [38] (θ_{dt} = 0.981). Please note that according to our model, this topic is dominated in works that deal with unstructured data (images, video, sound). However, in most cases, the weight of this topic in these kinds of applications does not exceed the weights of other topics significantly.

Finally, the last but not least topic detected by our model is Recommender Systems. This direction does not need additional comments. It is only worth noting that recently researchers have paid special attention to the generation of explainable recommendations, which provides explanations about why an item is recommended [39] (θ_{dt} = 0.986).

According to *Figure 3*, in addition to General Learning, only in three other areas is there a relatively large number of articles with a topic

weight close to 1. These are Segmentation, Pattern Mining and Text Mining. These topics are also subsections of machine learning, so a relatively large number of articles focused exclusively on algorithms is published under each of these areas. On the other hand, a topic such as Applications has virtually no such focused papers. It also has a big value of diversity metric according to (1). This can be explained by the fact that works regarding practical applications, as a rule, also present new modifications of algorithms.

Our model does not distinguish artificial neural networks (ANN) as a separate direction of data mining. This contradicts [7] but is consistent with [5] and [6]. According to our results, publications that use ANN models relate most often to the areas of General Learning and Segmentation.

The next issue that should be considered is topic collaboration, i.e. the topics' co-occurrence. Let θ_{di} and θ_{dj} be the weights of the topics i and j, respectively, in document d. Thus, we can define the topics' co-occurrence in this document as a product $\theta_{di}\,\theta_{dj}$. The maximal possible value of the co-occurrence of two topics in one document is 0.25 when $\theta_{di} = \theta_{dj} = 0.5$. From this, the maximal possible value of the topics' collaboration in the document corpus is $0.25 \cdot |D|$.

Thus, the topic collaboration in the document corpus can be computed as

$$c_{ij} = \sum_{d \in D} \theta_{di} \theta_{dj}.$$

Figure 4 presents these data. General Learning can be viewed as a central topic since it is most closely related to other areas of research. Human

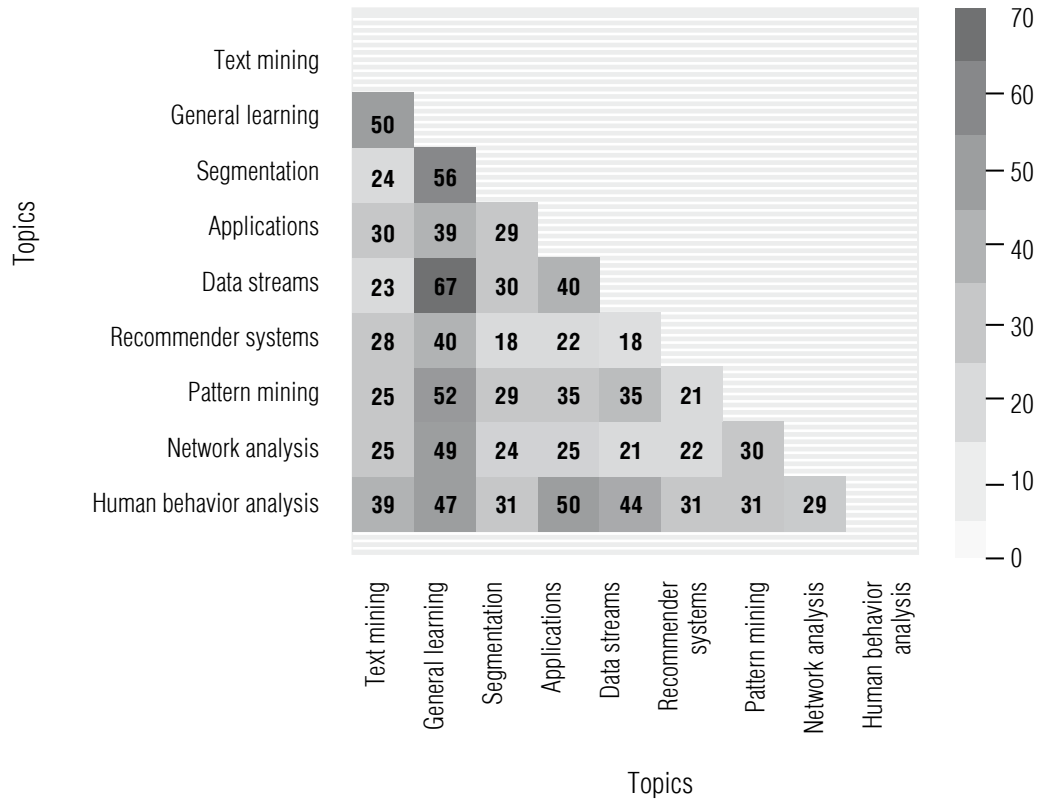


Fig. 4. A topic co–occurrence

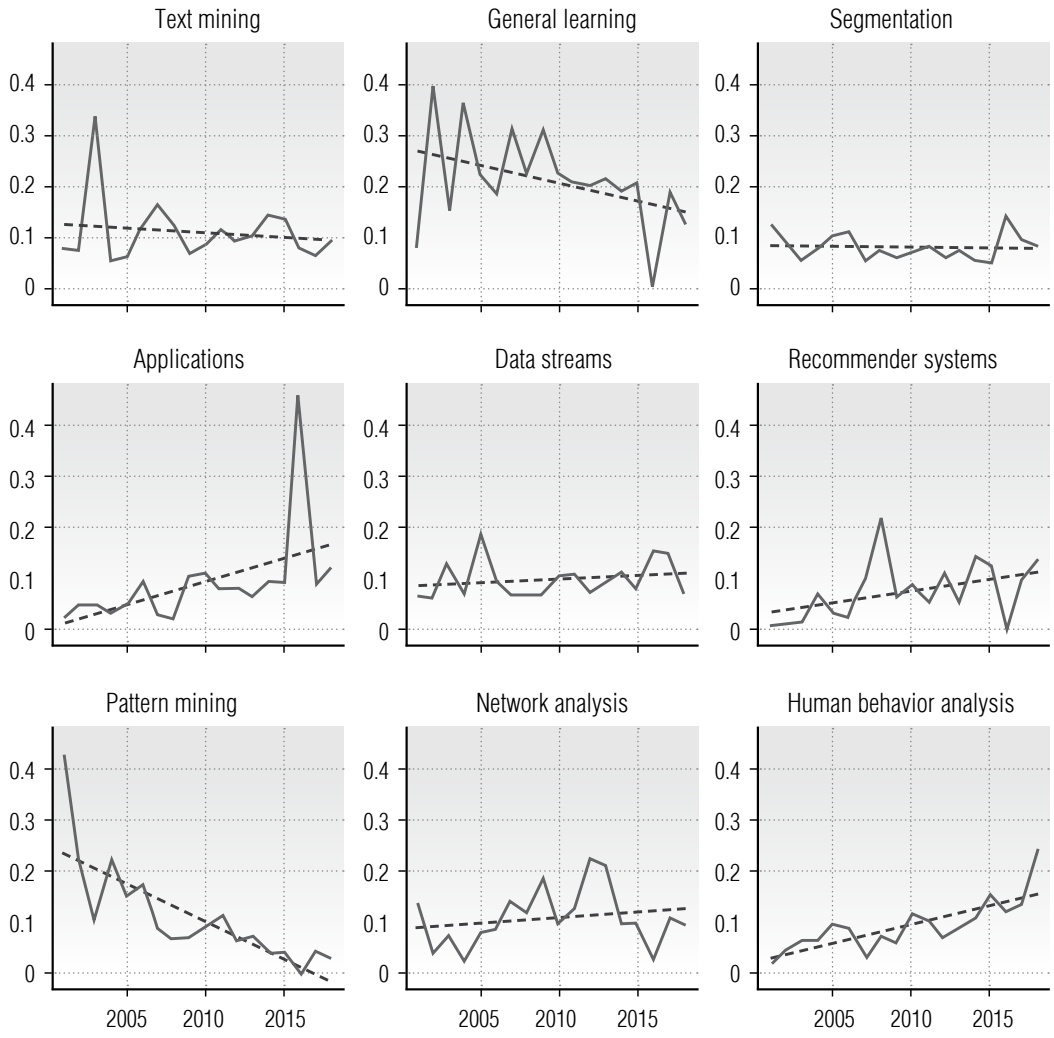


Fig. 5. The dynamics of the topics' popularity (solid line) and the trend (dashed line)

Behavior Analysis also has relatively strong connections with other directions. Recommender Systems, Network Analysis, and Text Mining are more isolated topics in our model because they are based on specialized algorithms.

The next stage of the analysis is a study of the dynamics of popularity and the impact of identified topics. Popularity is the derivative of the number of publications, and the impact is computed using the number of citations. *Figure 5* presents the dynamics of the topics' popularity (solid line), according to (2). The dashed line shows the trend. *Figure 6* presents the same data for the topics' impact, according to (3). These data shed light on the drift of interests of the data mining community regarding each research area.

Please note that the popularity and influence of many topics are subject to significant fluctuations. On the one hand, this can be explained by a short-term shift in the attention of researchers to hot topics. On the other hand, ICDM, although it is one of the most representative forums in the field of Data Mining, may not fully reflect the real dynamics of this discipline. For example, the word-

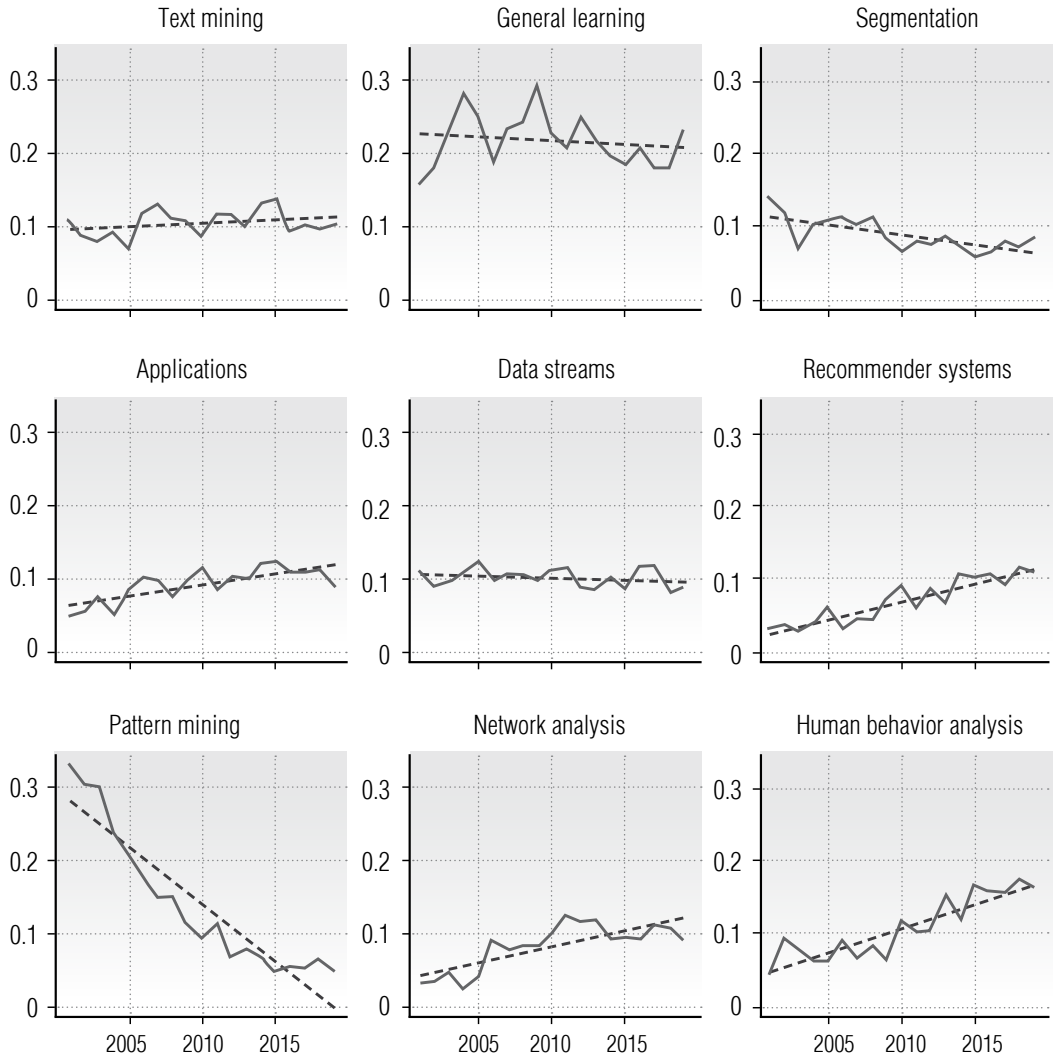


Fig. 6. The dynamics of the topics' impact (solid line) and the trend (dashed line)

ing of call for papers and the conference tracks by a program committee may affect researchers who submit works. However, we believe that an analysis of publications for 2001–2019 allows us to identify global trends, which is the goal of our study.

The data presented show that research attention to Pattern Mining is sharply decreasing both in terms of popularity and impact. The same, but less pronounced tendency is also characteristic of General Learning and Segmentation. These trends require more in-depth study. Firstly, it can be assumed that this is because the achievements in the field of machine learning algorithms and related techniques, including associations mining, are already so outstanding that further advancement requires serious efforts. Today, the most activity is observed in the field of deep learning, though, according to our results and analysis in [5, 6], deep neural networks are not directly related to Data Mining.

At the same time, the popularity of research related to Recommender Systems, Network Analysis, and Human Behavior Analysis is growing, which is most likely due to the increas-

ing availability of data and the practical value of these topics. The research direction related to practical Applications of data mining also is tending to grow. This can also explain the decline in interest in foundational algorithms; a significant part of the research community is focusing on more relevant practical issues.

The last two topics, Text Mining, and Data Streams have attracted steady interest from researchers. The results presented shed light on the structure and dynamics of data mining over the past twenty years and allow us to expand our understanding of this scientific discipline.

The next issue that is of interest from analyzing the content of publications is the diversity of documents. By analogy with (1), we can determine the diversity of a document through the entropy of its topics:

$$H_d = -\sum_i^T \theta_{di} \ln \theta_{di} ,$$

where θ_{di} is the weight of the topics i in document d;

T is the number of topics.

Figure 7 presents the mean diversity of proceedings of ICDM for 2001–2019. We see that the topic diversity of documents has grown steadily since the first conference and peaked in 2015. Over the past four years, there has been a reduction in the number of topics covered in one document.

We believe that this can be explained as follows. In the early 2000s, the main interest of researchers was focused on the knowledge discovery algorithms, which is presented by a list of critical topics highlighted in [5] and confirmed in [6] (*Table 1*). As they matured, these algorithms expanded their applications. Consequently, the set of topics covered in one scientific publication became more and more widespread. This can be considered as a search in the topic space that peaked in 2015. After 2015, a new research agenda was formed. As shown above, General Learning algorithms, as well as related areas such as Pattern Mining and Segmentation, are shifting to the background, although they continue to play an important role. More practical applications related to human behavior analysis, recommender systems, analysis of network communities, etc., come to the fore.

Table 3 presents a comparison of the data mining topics detected in our work and in [5] and [6]. Most of the topics of 2010 concentrate in the direction of General Learning. We

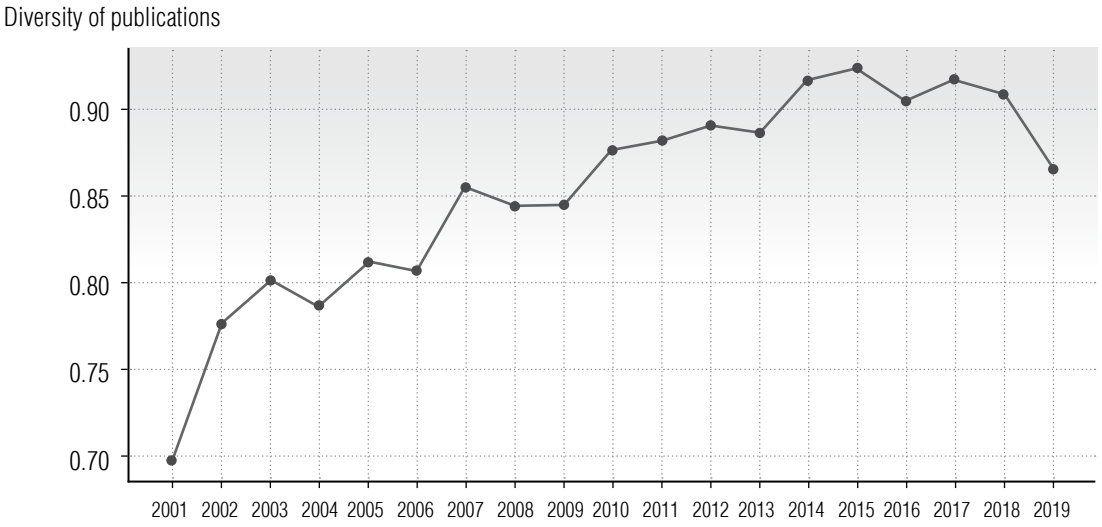


Fig. 7. The diversity of ICDM proceedings for 2001–2019

Table 3.

Mapping Data Mining topics

Data Mining Topics in 2020 (this work)	Data Mining Topics in 2010 [5, 6]
Text Mining	Link mining (e.g. PageRank algorithm)
General Learning	Classification (including C4.5, CART, kNN, and Naïve Bayes)
	Statistical learning (SVM and mixture models)
	Bagging and boosting
	Integrated mining (e.g. integrating classification and association rule mining)
	Rough sets
Segmentation	Clustering
Applications	NA
Data Streams	Sequential patterns
Recommender systems	NA
Pattern Mining	Association analysis
Network Analysis	Graph mining
Human Behavior Analysis	NA

included rough sets also in this category since this theory is applied to the classical knowledge discovery problems, such as discovering patterns in missing data [40].

New topics, the rapid growth of which our model identified, were not considered at all in 2010. We also note that the Text Mining topic detected in our work includes a much larger range of technologies and applications than searching for relations between documents.

Interestingly, in the middle of 2010s, researchers in the field of economics and management recorded an increase in interest in Data-Driven Decision Making (DDD) [41]. That refers to the practice of basing decisions on the analysis of data rather than purely on human knowledge and intuition. Authors of [42] report that the use of DDD in US manufacturing nearly tripled (from 11 percent to 30 percent of plants) between 2005 and 2010. More recent studies confirm the increasing role of DDD as one of the best management practices [43].

The description of DDD which is given in management literature entirely coincides with the definition of Data Mining discussed at the beginning of our work. DDD also includes the engineering and processing of data and the discovery of useful patterns. However, DM considers this activity from a technological point of view. DDD approaches this issue in the context of closely related processes in the organization, including purely human activities. Nevertheless, we can consider the growing interest in DDD as one of the key drivers affecting the shift in DM studies that is presented in *Figure 7*.

Conclusion

We presented a study of the intellectual structure of Data Mining as a scientific discipline carried out using topic analysis. This approach made it possible to identify nine main areas in Data Mining and to study their dynamics.

The main result of our work is that we have discovered a shift in interests from machine learning algorithms to more practical applications. According to our data, this change of focus took shape in the middle of the 2010s. We attribute this shift to a combination of three factors:

Firstly, the basic data mining algorithms have reached a high level of maturity.

Secondly, with the development of social networks, a large amount of data has become available.

Third, there has been a steady demand from the business for data-driven decision-making. ■

References

1. Piatetsky-Shapiro G., Fayyad U. (2012) An introduction to SIGKDD and a reflection on the term 'data mining'. *ACM SIGKDD Explorations Newsletter*, vol. 13, no 2, pp. 102–103. DOI: 10.1145/2207243.2207269.
2. Witten I.H., Frank E., Hall M., Pal C. (2017) *Data mining: Practical machine learning tools and techniques.* Cambridge, MA: Morgan Kaufmann.
3. Han J., Kamber M., Pei J. (2012) *Data mining: Concepts and techniques.* Waltham, MA: Morgan Kaufmann. DOI: 10.1016/C2009-0-61819-5.
4. Rather B. (2011) *Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data*. Sound Parkway, NW: CRC Press.
5. Yang Q., Wu X. (2006) 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, vol. 5, no 4, pp. 597–604. DOI: 10.1142/S0219622006002258.
6. Wu X. (2010) 10 years of data mining research: retrospect and prospect. Proceedings of the *10th IEEE International Conference on Data Mining (ICDM), Sydney, Australia, 13–17 December 2010*, p. 7. DOI: 10.1109/ICDM.2010.172.
7. Liao S.H., Chu P.H., Hsiao P.Y. (2012) Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, vol. 39, no 12, pp. 11303–11311. DOI: 10.1016/j.eswa.2012.02.063.
8. Van Eck N.J., Waltman L. (2009) Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, vol. 84, no 2, pp. 523–538. DOI: 10.1007/s11192-009-0146-3.
9. Thilakaratne M., Falkner K., Atapattu T. (2019) A systematic review on literature-based discovery: general overview, methodology, & statistical analysis. *ACM Computing Surveys*, vol. 52, no 6, article no 129. DOI: 10.1145/3365756.
10. Zelenkov Y. (2019) The topic dynamics in knowledge management research. Proceedings of the *14th International Conference on Knowledge Management in Organizations (KMO 2019), Zamora, Spain, 15–18 July 2019*, pp. 324–335. DOI: 10.1007/978-3-030-21451-7_28.
11. Mann G.S., Mimno D., McCallum A. (2006) Bibliometric impact measures leveraging topic analysis. Proceedings of the *6th ACM/IEE Joint Conference on Digital Libraries (JCDL '06), Chapel Hill, NC, USA, 11–15 June 2006*, pp. 65–74. DOI: 10.1145/1141753.1141765.
12. Dam H.K., Ghose A. (2016) Analyzing topics and trends in the PRIMA literature. Proceedings of the *19th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA* 2016*), Phuket, Thailand, 22–26 August 2016*, pp. 216–229. DOI: 10.1007/978-3-319-44832-9_13.
13. Dias L., Gerlach M., Scharloth J., Altman E.G. (2018) Using text analysis to quantify the similarity and evolution of scientific disciplines. *Royal Society Open Science*, vol. 5, no 1, article no 171545. DOI: 10.1098/rsos.171545.
14. Jensen S., Liu X., Yu Y., Milojevic S. (2016) Generation of topic evolution trees from heterogeneous bibliographic networks. *Journal of Informetrics*, vol. 10, no 2, pp. 606–621. DOI: 10.1016/j.joi.2016.04.002.
15. Syed S., Spruit M. (2017) Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. Proceedings of the *4th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2017), Tokyo, Japan, 19–21 October 2017*, pp. 165–174. DOI: 10.1109/DSAA.2017.61.
16. Blei D.M., Ng A.Y., Jordan M.I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research*, no 3, pp. 993–1022.
17. Mockus J. (2012) *Bayesian approach to global optimization: Theory and applications.* Heidelberg: Springer. DOI: 10.1007/978-94-009-0909-0.
18. Tang J., Meng Z., Nguyen X., Mei Q., Zhang M. (2014) Understanding the limiting factors of topic modeling via posterior contraction analysis. *Proceedings of Machine Learning Research*, vol. 32, no 1, pp. 190–198.

19. Molina L.C., Belanche L., Nebot A. (2002) Feature selection algorithms – A survey and experimental evaluation. Proceedings of the *2nd IEEE International Conference on Data Mining (ICDM* 2002*), Maebashi City, Japan, 9–12 December 2002*, pp. 306–313. DOI: 10.1109/ICDM.2002.1183917.

20. Read J., Pfahringer B., Holmes G. (2008) Multi-label classification using ensembles of pruned sets. Proceedings of the *8th IEEE International Conference on Data Mining (ICDM), Pisa, Italy, 15–19 December 2008*, pp. 995–1000. DOI: 10.1109/ICDM.2008.74.

21. Chen X., Pan W., Kwok J.T., Carbonell J.G. (2009) Accelerated gradient method for multi-task sparse learning problem. Proceedings of the *9th IEEE International Conference on Data Mining (ICDM), Miami Beach, FL, USA, 6–9 December 2009*, pp. 746–751. DOI: 10.1109/ICDM.2009.128.

22. Shin K. (2011) Partitionable kernels for mapping kernels. Proceedings of the *11th IEEE International Conference on Data Mining (ICDM), Vancouver, BC, Canada, 11–14 December 2011*, pp. 645–654. DOI: 10.1109/ICDM.2011.115.

23. Li Z., Yada K. (2015) Why do retailers end price promotions – A study on duration and profit effects of promotion. Proceedings of the *15th IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015*, pp. 328–335. DOI: 10.1109/ICDMW.2015.56.

24. Camino R.D., State R., Montero L., Valtchev P. (2017) Finding suspicious activities in financial transactions and distributed ledgers. Proceedings of the *17th IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 18–21 November 2017*, pp. 787–796. DOI: 10.1109/ICDMW.2017.109.

25. Guo T., Bifet A., Antulov-Fantulin N. (2018) Bitcoin volatility forecasting with a glimpse into buy and sell orders. Proceedings of the *18th IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018*, pp. 989–994. DOI: 10.1109/ICDM.2018.00123.

26. Gouda K., Zaki M.J. (2001) Efficiently mining maximal frequent itemsets. Proceedings of the *2001 IEEE International Conference on Data Mining (ICDM), San Jose, CA, USA, 29 November – 2 December 2001*, pp. 163–170. DOI: 10.1109/ICDM.2001.989514.

27. Ma S., Hellerstein J.L. (2001) Mining mutually dependent patterns. Proceedings of the *2001 IEEE International Conference on Data Mining (ICDM), San Jose, CA, USA, 29 November – 2 December 2001*, pp. 409–416. DOI: 10.1109/ICDM.2001.989546.

28. Zhou D., Orshanskiy S.A., Zha H., Gees C.L. (2007) Co-ranking authors and documents in a heterogeneous network. Proceedings of the *7th IEEE International Conference on Data Mining (ICDM), Omaha, NE, USA, 28–31 October 2007*, pp. 739-744. DOI: 10.1109/ICDM.2007.57.

29. Tang J., Jin R., Zang J. (2008) A topic modeling approach and its integration into the random walk framework for academic search. Proceedings of the *8th IEEE International Conference on Data Mining (ICDM), Pisa, Italy, 15–19 December 2008*, pp. 1055–1060. DOI: 10.1109/ICDM.2008.71.

30. Shehata S., Karray F., Kamel M. (2006) Enhancing text clustering using concept-based mining model. Proceedings of the *6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China, 18–22 December 2006*, pp. 1043–1048. DOI: 10.1109/ICDM.2006.64.

31. Yang D., Li B., Rettig L., Cudre-Mauroux P. (2017) HistoSketch: Fast similarity-preserving sketching of streaming histograms with concept drift. Proceedings of the *17th IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017*, pp. 545–554. DOI: 10.1109/ICDM.2017.64.

32. Dhurandhar A. (2010) Learning maximum lag for grouped graphical Granger models. Proceedings of the *10th IEEE International Conference on Data Mining Workshops (ICDMW), Sydney, Australia, 13 December 2010*, pp. 217–224. DOI: 10.1109/ICDMW.2010.9.

33. Wang H., Li Z., Lee W.-C. (2014) PGT: Measuring mobility relationship using personal, global and temporal factors. Proceedings of the *14th IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14–17 December 2014*, pp. 570–579. DOI: 10.1109/ICDM.2014.111.

34. Pirbhulal S., Wu W., Li G. (2018) A biometric security model for wearable healthcare. Proceedings of the *18th IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, 1 7–20 November 2018*, pp. 136–143. DOI: 10.1109/ICDMW.2018.00026.

35. Yang J., Leskovec J. (2012) Defining and evaluating network communities based on ground-truth. Proceedings of the *12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium, 10–13 December 2012*, pp. 745–754. DOI: 10.1109/ICDM.2012.138.

36. Shi L., Tong H., Tang J., Lin C. (2014) Flow-based influence graph visual summarization. Proceedings of the *14th IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14–17 December 2014*, pp. 983–988. DOI: 10.1109/ICDM.2014.128.

37. Hung M.-C., Yang D.-L. (2001) An efficient fuzzy c-means clustering algorithm. Proceedings of the *2001 IEEE International Conference on Data Mining (ICDM), San Jose, CA, USA, 29 November – 2 December 2001*, pp. 225–232. DOI: 10.1109/ICDM.2001.989523.

38. Pei Y., Zaiane O.R., Gao Y. (2006) An efficient reference-based approach to outlier detection in large datasets. Proceedings of the *6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China, 18–22 December 2006*, pp. 478–487. DOI: 10.1109/ICDM.2006.17.

39. Wang X., Chen Y., Yang J., Wu L., Wu Z., Xie X. (2018) A reinforcement learning framework for explainable recommendation. Proceedings of the *18th IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018*, pp. 587–596. DOI: 10.1109/ICDM.2018.00074.

40. Wang H., Wang S. (2009) Discovering patterns of missing data in survey databases: an application of rough sets. *Expert Systems with Applications*, vol. 36, no 3, part 2, pp. 6256–6260. DOI: 10.1016/j.eswa.2008.07.010.

41. Provost F., Fawcett T. (2013) Data science and its relationship to big data and data-driven decision making. *Big Data*, vol. 1, no 1, pp. 51–59. DOI: 10.1089/big.2013.1508.

42. Brynjolfsson E., McElheran K. (2016) The rapid adoption of data-driven decision-making. *American Economic Review*, vol. 106, no 5, pp. 133–139. DOI: 10.1257/aer.p20161016.

43. Song P., Zheng C., Zhang C., Yu X. (2018). Data analytics and firm performance: An empirical study in an online B2C platform. *Information & Management*, vol. 55, no 5, pp. 633–642. DOI: 10.1016/j.im.2018.01.004.

About the authors

Yury A. Zelenkov

Dr. Sci. (Tech.);

Professor, Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: yzelenkov@hse.ru

ORCID: 0000-0002-2248-1023

Ekaterina A. Anisichkina

Student, BSc Program "Business Informatics", Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: eaanisichkina@edu.hse.ru

DOI: 10.17323/2587-814X.2021.1.47.58

Statistical sustainability of a digital organization

Vladimir I. Ananyin [a]
E-mail: v.ananiin@gmail.com

Konstantin V. Zimin [b]
E-mail: konst.zimin@gmail.com

Mikhail I. Lugachev [e]
E-mail: mlugachev@gmail.com

Rinat D. Gimranov [d]
E-mail: gimranov_rd@mail.ru

[a] Russian Presidential Academy of National Economy and Public Administration
Address: 82 build. 1, Prospect Vernadskogo, Moscow 119571, Russia

[b] The Russian Union of CIO
Address: 34, Seleznevskaya Street, Moscow 123056, Russia

[c] Lomonosov Moscow State University
Address: 1 build. 46, GSP-1, Leninskie Gory, Moscow 119991, Russia

[d] PJSC Surgutneftegaz
Address: 1 block 1, Grigoriya Kukuevitskogo Street, Surgut 628415, Russia

Abstract

An important feature of a digital organization is its ability to change rapidly. For an organization to remain capable of rapid change, it must be on the brink of resilience, since a resilient organization always resists change. The article examines the borderline state of the organization, which is on the verge of its stability and instability. In this state, the organization begins to lose predictability in the details of behavior, but still retains predictability in general. The authors called this borderline state the statistical sustainability of the organization. The phenomenon of statistical sustainability of an organization is very similar to the property of stability of the frequency of mass events and average

values described in mathematical statistics by a similar term. To analyze the nature of the statistical sustainability of the organization, the authors used the ideas of strange attractors and modes with sharpening from the theory of complex systems. A strange attractor is an area of the organization's behavior that, outside this area, is an area of stability for the organization, and inside it is an area of complete unpredictability. The theory of complex systems has shown that it is in the regions of strange attractors that the conditions for the variability of systems are created, and the theory of modes with aggravation shows the conditions under which this variability can lead to self-organization, that is, the spontaneous emergence of new structures. This article shows that systematic digitalization objectively leads to the formation of the statistical sustainability of the organization and creates the preconditions for maintaining the organization's ability to make rapid changes. In traditional management, the statistical sustainability of an organization is viewed as a threat and a source of risk. Therefore, in the context of systematic digitalization, traditional approaches to management should be significantly refined.

Key words: digital organization; variability; stability/instability of the organization; exponential organization; self-organization; strange attractor; diversity; scattering; digital transformation; statistical sustainability.

Citation: Ananyin V.I., Zimin K.V., Lugachev M.I., Gimranov R.D. (2021) Statistical sustainability of a digital organization. *Business Informatics*, vol. 15, no 1, pp. 47–58.
DOI: 10.17323/2587-814X.2021.1.47.58

Introduction

In the first half of the 2000s, successful startups built on new information technologies appeared in the United States. The business models of these organizations were radically different from the existing ones, as they were built on the sharing of data. Such organizations are called digital. Many digital organizations had an amazing property: their businesses had explosive growth and fit well on an exponential curve. These rapidly growing digital organizations are called exponential organizations [1]. Their phenomenal success in the second half of the 2000s triggered a wave of new digital organizations. The inclusion of global business leaders in this wave has launched a powerful trend in the digital transformation of the largest companies themselves. The wave of the formation of digital organizations acquired a grandiose scale when, in the mid-2010s, many states took a course towards building a digital economy and created powerful incentives for digital transformation in their national enterprises [2–5].

By 2020, a large number of digital organizations and organizations are in the process of digital transformation. The study of the nature of exponential organizations began in the early 2010s, and already in 2014, a monumental work by a team of authors from the University of Singularity on the analysis of the patterns of formation of exponential organizations was published in the United States [1]. In this work, the authors showed that all exponential organizations managed to discover not only new forms of organizational, human and informational capital, but also new forms of their synthesis [6]. The authors also showed that in these organizations, the time for both decision-making and their implementation has been drastically reduced. In

fact, exponential organizations began to transform towards a real-time enterprise (RTE) [7], in which decisions are made and implemented at the rate of development of emerging management situations. In this case, the organization's business gets powerful competitive advantages in the market due to the fact that it finds a solution faster than others and adapts to new changed conditions.

Leadership is important not only to achieve, but also to be able to keep it. The organization will have a competitive advantage in the market until its competitors or partners adopt similar decision-making practices. Then the speed of changes in market situations will be determined not by the speed of some economic market factors, but by the speed of decision-making and changes in competitors or even partners. If suddenly it turns out that our competitors or partners are changing faster than ourselves, then at one point our business loses leadership: for competitors we become an easy target, and for partners – a burden.

Digitalization creates support for making not only short-term and medium-term decisions, but also strategic long-term decisions, when it is necessary to recognize a potential "black swan" [8] and have time to prepare for a meeting with it. This problem is of great concern to all the top executives of companies – world leaders. Peter Diamandis, in the preface to the book [1], managed to formulate it succinctly: "Today you are forced to compete not only and not so much with established multinational corporations. Your competitor can be any guy from Silicon Valley or Bandra County in Mumbai who sits in his garage and uses the latest online tools to develop and distribute his innovative products" [1]. There are "black swans" that come without any precursors at all, for example, the global pandemic of the coronavirus COVID-19 in 2020. In this case, the ability to quickly change and adapt to new conditions becomes for business just a matter of life or death.

If quick adaptability, that is, the ability to quickly change and adapt to new conditions, is so important for a business, why are there so few companies capable of doing this? What new quality is lacking in business for information technology to become a driver of its real transformation? In our opinion, this new quality is the instability of the organization. This article is devoted to the analysis of instability of the organization.

1. The instability of the organization as a threat and an opportunity

Instability is strongly associated with the crisis of the organization and therefore in management traditionally enjoys a bad reputation. Instability threatens to lose the predictability of the organization's behavior and the loss of its manageability. Management feels threatened by it and fights it with all its might. For risk management and crisis management, instability is a marginal state of the organization, which must either be avoided or quickly eliminate the consequences of the crisis that has broken out in the organization [9–11].

However, volatility can be useful when management wants to make rapid changes in their organization. The more resilient an organization is, the more resistant it will be to change, and the less room there will be for rapid and large-scale change. And this is not someone's malicious intent, but an objective manifestation of the same stability. In order to make such changes, the organization must be brought to the brink of sustainability, when it has not yet completely lost the predictability of its behavior. Those who have mastered this art gain a powerful competitive advantage – a high degree of adaptability and innovation of their business.

As we can see, instability brings not only threats, but also opportunities. This can be well observed in both technical and natural systems. In engineering practice, it is well known that the

high level of maneuverability of a modern fighter is achieved by the fact that its design is initially based on aerodynamic instability. The predictability of fighter control is provided by the operation of the onboard computer control system, which, based on a mathematical model of the dynamics of the fighter flight in real time, provides the pilot with the predictability of motion control. The fighter's aerodynamic instability gives it a great competitive advantage in maneuverability and air superiority, but it is this instability that also creates great risks for it: the failure of the onboard electronics makes its behavior chaotic and unmanageable. If instability is an opportunity for a fighter, then for a transport or passenger aircraft it is a threat that engineers and pilots are constantly fighting against.

In medical practice, it is known that when transplanting an organ, it is necessary to weaken the patient's immunity. The human immune system is the most important mechanism for ensuring the stability of the body, ensuring its integrity and individuality by recognizing and removing foreign substances and cells. Strong immunity will lead to the rejection of someone else's organ, so the immunity must first be weakened, then allow it to carefully, purposefully accept someone else's organ as its own, and only then-to restore immunity again. During the period of weakening of the immune system, it is very important to protect the body from foreign infections. The engraftment of a foreign organ is a complex process of self-organization of the entire body. Here, the human body itself decides how it will act in such a situation, and doctors can only understand it in time and help it.

In contrast to technical and natural systems, an organization is a social system that consists of people who have their own ideas and history of relationships, their own intentions and expectations, motives and norms of behavior. People in an organization build various kinds of mutual relations, so the organization cannot be reduced to a simple set of people, but represents a certain new quality that manifests itself through the objective properties of the organization as a whole. One of these properties is its stability/instability.

2. Statistical sustainability of the organization

The concept of stability / instability of the organization is closest to the field of knowledge about management. In the theory of management [12], the object and the subject of management are traditionally distinguished. In our case, the object of management is an operating organization, and the subject is the head (team of managers) of this organization. The management team manages the organization, that is, influences the organization in a way that allows them to achieve the expected results. Management is always carried out in conditions of disturbances, that is, unexpected (unplanned) deviations from the normal operating conditions of the organization and the management itself. Disturbances can have both an external origin (for example, a change in the organization's operating conditions) and an internal origin (for example, the appearance of internal problems, conflicts, or initiatives). Outrages require special management actions from the management team.

An organization will be considered sustainable if any situation caused by outrage is predictable for the management team, that is, the team has sufficient knowledge and experience both in terms of the development of the situation and in ways to resolve it.

The organization will be considered unsustainable if the development of the situation associated with the outrage is unpredictable for the management team.

Let's consider the stability of an organization on the example of a separate business process of a real industrial company.

In a large industrial company, there is a process of agreement of contracts. Each contract

has a supervisor who is responsible for developing, approving and monitoring the execution of the contract with a supplier. At the stage of joint development of the contract with the supplier and during the approval process with the company's services, many changes are made to the contract, which often require repeated approval with the supplier and other participants. Despite the fact that the company has developed clear regulations and standards for approval, the movement of most contracts in the process is poorly predictable. This is due to the high level of market volatility, changes in the organization itself, as well as a lot of contradictions between the company's divisions, which give rise to a low level of mutual trust, which results in distrust of the contract supervisor. As a result, the approval of each contract for the curator turns from a "race with clear rules on pre-laid paths" to a "cross-country race with changing rules and landscape."

All participants in the process are generally interested in getting contracts agreed as quickly as possible, but also do not want to take on the additional responsibility associated with the uncertainty of the consequences of the resulting changes. Each time, all the participants somehow manage to agree on the speed of development of the situation. It is interesting that the average time for the approval of contracts for the process has remained stable for several years, but the movement of each contract is difficult to predict.

Naturally, the process of agreeing on contracts is subject to disturbances, both external and internal. As a result, we see signs of both stability and instability.

At the beginning of a contract company, when there are still few contracts and standard contracts prevail, the process becomes stable: the movement of each contract becomes predictable and largely complies with the rules of the process.

Last year, by order of the Director General, another control service was introduced to restore order in the process. As a result, the coordination of the participants in the process was greatly slowed down, the process became unstable and unexpected failures of coordination began. Senior managers were surprised to find that they themselves now had to intervene in the "resolution" of coordination conflicts. After all the participants got used to each other again, the process returned to the old mode and its average indicators again became the same as before.

In the given example, we can observe the evolution of the mode of operation of the process from classical stability, that is, complete predictability, to classical instability, that is, complete unpredictability of the process. It is interesting that between these extremes there is still a regime where the movement of each treaty turns out to be a process that has remained very predictable for several years. This mode of operation of the process will be called "statistical sustainability." The phenomenon of statistical sustainability can often be observed both at the level of individual management situations, projects of operational processes, and at the level of the organization as a whole.

3. Statistical sustainability and self-organization

Effects similar to statistical sustainability are also widespread in technical and natural systems and are well described in the theory of complex systems [13–17]. The theory of complex systems describes the behavior of technical and natural systems in the form of dynamic mathematical models, where the behavior of the system is represented as the trajectory of its movement in phase space. The theory identifies special states of a complex system, which are peculiar areas of attraction for all the trajectories of its mathematical model. Such a region of attraction of trajectories in the mathematical model is called an attractor [13, 16]. In the attractor region, the behavior of the system becomes absolutely predictable. An exam-

ple of such an attractor is the center of a funnel formed by a vortex of water going into the hole of the sink. The attractor is a property of the system itself and reflects the modes of its stable behavior, regardless of the disturbances that occurred in it.

The theory of complex systems also suggests that under certain conditions, the system may have attractors, which represent a limited set of states of the system (a closed region of the phase space of the system), which has special properties. From the outside to this set of states (the phase domain) trajectories describing the behavior of the system are also attracted, but within this region, the behavior of the system is fundamentally unpredictable. In mathematics, such areas are called "strange attractors" [16]. It is interesting that inside such a strange attractor, the behavior of the system at any given time is unpredictable, but, nevertheless, the behavior of the system on average turns out to be quite predictable.

Strange attractors only appear in systems that are open, that is, through which the flow of energy and information passes, and which are in a state of increased dynamics. Strange attractors have an amazing property: when the conditions of the system's activity change, they can disintegrate, giving rise to chaotic behavior of the entire system, but they can also transform into other strange attractors, and even generate stable structures that are classical attractors. Such generation of stable structures, which occurs not by someone's will, but by virtue of the natural laws of behavior of a complex system, is called self-organization [13–16].

The process of contract negotiation discussed above, which is in a state of statistical sustainability, is surprisingly similar to a strange attractor. The constantly reproducing set of practices for the negotiation of contracts clearly manifests itself as an attractor. And this attractor is strange, because when implementing the agreement of each contract, the participants each time find their own specific solution. In addition to the accepted practices, they include interpersonal relationships and "on a live thread" each time they find unique solutions. At the same time, the process indicators for each contract can be considered random variables, but the average indicators for the process as a whole are quite predictable. In this case, the coordination activity is no longer like a race on pre-laid tracks (process), but rather like cross-country orienteering, and even with changing terrain.

In the theory of complex systems, two fundamental processes are distinguished during the formation of structures [15, 18] – the mode that generates an increase in diversity, and the mode that reduces (blurs, disperses) diversity in the system.

The mode that generates the growth of diversity in the system is called the "mode with aggravation" (S-mode). This is the mode of explosive growth of any process parameters, for example, the burning process, which turns into an explosion. The regime with aggravation generates a rapid increase in the variety of structures formed in the system. An analog of a regime with aggravation for an organization can be an avalanche-like increase in internal initiatives or conflicts.

The mode that generates a decrease in diversity in the system is called dispersion or dissipation. This is the mode of erosion, dispersion of structures that have arisen in the system, for example, the extinction of flame due to fading fuel burn-up and the dissipation of the released energy. An analogous mode of reducing diversity in an organization can be the coordination of participants, during which they communicate with each other, share information and knowledge, come to agreements and find common solutions.

According to the theory of complex systems, the dynamics of self-organization is determined by the ratio of these two processes.

The dominance of diversity growth over dispersion. This relationship is called the LS-

mode. The dispersion is less intense (L – lower) than the diversity growth process (S-mode). An analog of LS-mode in an organization is a situation when the participants are hit by a stream of initiatives and conflicts, which for them becomes less predictable and/or they do not have time to respond in a coordinated manner. Under these conditions, the emerging new structures are extremely unstable: they form quickly and disintegrate quickly.

The dominance of dispersion over the growth of diversity. This relationship is called the HS-mode. The dispersion is more intense (H – higher) than the diversity growth process (S-mode). An analogue of the HS-mode in an organization is a situation where any initiative among its participants is drowned in agreements, generating resistance to changes in established practices. In HS-mode, the organization exhibits the properties of classical stability.

The parity of dispersion and the growth of diversity is the condition for the formation of strange attractors and the phenomenon that in complex systems is called self-organization. Self-organization is the process of spontaneous ordering, the emergence of spatial, temporal, spatiotemporal, or functional structures, occurring in open nonlinear systems [15, 18, 19]. Self-organization is a concept that expresses the ability of complex systems to organize their internal structure not by someone's will, but by virtue of the objective laws of the evolution of complex systems. In complex systems, the shift of parity towards the regime of diversity growth leads to the formation of many new embryonic structures. The subsequent shift of parity towards dispersion triggers the process of natural selection of the most viable structures from the newly formed embryos.

What are these new structures that are emerging in our process of negotiating contracts, which is in the mode of statistical sustainability? New structures are practices that begin to form as informal individual or collective experiences of participants in the reconciliation process. It begins as intuitive partial solutions, without claims to generalization. Further, if the experience survives (that is, if it is reused), then it can be realized as general knowledge. If this general knowledge proves its value, it can be transformed into rational organizational, methodological, personnel and technical solutions.

In our contract agreeing process, its participants manage to find a solution and agree on the changing terms of agreement. Changes in the process come not only from the outside, but are also generated internally by the unique decisions of the participants themselves. The process of agreeing on each contract becomes difficult to predict, but at the same time, parity is achieved between the processes of generating change (increasing diversity) and coordinating the actions of participants (scattering). Under the conditions of statistical sustainability of the process, the shift of parity towards the LS-mode leads to the appearance of a variety of practices, and the shift of parity towards the HS-mode triggers the natural selection of the most viable practices.

In the statistical sustainability mode, the organization requires a lot of attention and respect from the manager. To the one who begins to understand it, it will "prompt" the possible directions of its development. It is pointless to impose your rational ideas and will on the organization. In this case, intuition begins to play a large role in management, but at the same time, intuitive search must go hand in hand with constant rational analysis. If in a stable state of the organization the manager dictates his will to it, then in crisis management he seeks an alliance with it. For many managers, this can be a worldview shock. "Her Majesty" statistical sustainability never forgives disrespect, inattention and slowness. But for understanding and tact, it can reward "royally": the organization receives a powerful generator of new practices. It is in these situations

that the effect of self-organization appears, which is the basis of innovations and breakthrough solutions.

4. Statistical sustainability and digitalization

In traditional organizations, the basis of the value of the created product was the physical properties of the product itself (for example, the level of consumer quality) or the organization of business activities (for example, just-in-time delivery). In digital organizations, the product becomes digital; its value is based on data. At the same time, the more data and program code associated with a digital product, the more opportunities there are to extract value from it [6]. In this case, the business models of interaction with other organizations become digital, that is, they are already fundamentally based on data. The value chains in which an organization is included along with other digital organizations are also becoming digital. Data becomes the "air" that digital organizations "breathe", and new technologies (for example, big data, the Internet of things, machine learning and artificial intelligence, virtual and augmented reality, blockchain) help them to produce, assimilate and use it to build new forms of business organization.

The systematic use of information and new technologies has a strong impact on the parity of the processes of generating change (increasing diversity) and coordinating the actions of participants (dispersion) (*Table 1*).

The analysis of the impact of digitalization on the balance of processes of generating changes and coordinating the actions of participants (Table 1) shows that digitalization contributes to the growth of variability and diversity of the organization. At the same time, one of the limiting factors for the growth of variability is the development of coordination processes, that is, the ability of an organization to coordinate its actions in the increasing flow of changes. By increasing variability and diversity, digitalization shifts parity towards the LS-mode and makes newly formed practices unviable. They disintegrate, and the organization regains parity in the processes of generating change and coordination. This parity manifests itself as the statistical sustainability of the organization. Let's consider an example from real practice.

In many manufacturing companies, the classic implementation of the logistics supply chain of material and technical resources (MTP) includes the stages of planning the production need for MTP, preparing and executing purchases, transportation, storage, release to production, and use. The computerization of the process is implemented by classical ERP systems based on the MRP II model. Due to the fact that there is a gap between the real state of affairs (delivery, volume of MTR in warehouses, sales) and information in the system, as well as due to delays in the receipt and processing of information in the ERP system, stocks are formed in the logistics chain that dampen both areas of uncertainty and the costs of opportunistic behavior. It is with the help of these dampers that the organization is stable in terms of meeting the production needs for MTR. For each MTR application, key indicators are recorded, which are monitored by management when making decisions. And for the process as a whole, some parameters are fixed – the average duration of the purchase, the permissible amount of insurance stock, etc., which determine the organizational and administrative decisions on the process.

Digital transformation is changing this process both in terms of ensuring the reliability of information and in terms of reducing the costs of opportunistic behavior. In the process of transformation, 100% marking and automatic (without manual input) tracking of MTR (barcodes or RFID tags, readers) is carried out, the transition to fully robotic warehouses is carried out, electronic document management is provided, inforobots are launched to perform all routine and standard operations.

Table 1.

The impact of digitalization on the parity of the processes of generating change (increasing diversity) and coordinating the actions of participants (scattering)

Changes related to digitalization	Offset of parity to the side of LS-mode	Offset of parity to the side of HS-mode
Increasing the depth and scale of analytics	Intelligent analytics on big data allows you to see what you could not see before: new threats and opportunities are opening up more often, which require an increasingly complex response. In fact, digitalization opens up a large number of new potential external and internal disturbances and increases the speed of their development	The same intelligent analytics allows you to quickly find a solution that is acceptable to all participants in a difficult situation. It creates a unified picture that accelerates the coordination of participants
The growth of the scale and depth of integration	The Information space goes beyond the boundaries of the digital organization itself and opens up (within the framework of access rights) to other participants in the value chains. This increases not only the scale, but also the depth of information integration. Information links in the chains become stronger, which leads to the fact that changes in one organization become common to organizations of the entire chain. Digital organization becomes more sensitive to external changes	It is the same integration of the information space that is the basis for the rapid synchronization of changes in value chains (value chain orchestration). Stronger information links in the chains are able to transmit not only changes, but also options for synchronizing the efforts of the chain participants
Increasing the number of participants in coordination	Value chain management dramatically expands the number of participants in coordination, which means that there is a sharp increase in the diversity and variability of their interests, expectations and intentions	Modern information technologies allow us to increase the efficiency of coordination: quickly find the necessary participants and form a working group, quickly create a working information space and organize group work, attract and reuse accumulated knowledge, make decisions and monitor their implementation. It should be borne in mind that effective coordination requires constant professional development and the development of expert and leadership motivation of the participant
The growing complexity of models outstripping the growth of intelligence	Working with big data leads to the emergence of a wide variety of operating models, as well as to the complication of cause–and–effect relationships in these models. Using these models to make decisions will require a lot of intellectual resources. If the growth of intelligence will constantly lag behind the growth of the complexity of models, then decision–making will be constantly accompanied by errors, their corrections and additional changes	The constant increase in the complexity and diversity of activity models will stimulate the development of our intellectual abilities, as it has always been in history.
Increasing the speed of processes and their individualization	Digitalization allows you to radically increase the productivity of operational processes, reaching the level of value chains that go beyond the boundaries of the organization. At the same time, the implementation of each order in the process becomes more and more individual. This leads to the fact that the speed of implementing changes increases and the implementation of changes itself becomes more complex and diverse	Some of the decisions at the lower level, especially in the conditions of increasing speed and individualization of processes, will be taken over by artificial intelligence systems. Decision–making at the higher levels in the context of increasing speed and diversity will remain up to the individual. New thinking “in complexity” can reduce the impact of these changes

At each moment of time, complete and reliable information is provided about the current status of the MTR supply request and their movement, which serves as feedback for the process control loops. Based on this information, management decision-making systems work, many of which are implemented using artificial intelligence.

Thus, from the point of view of the execution of each specific application for MTR, the system behaves in a completely unpredictable way, making operational decisions at any stage. For example, if a delivery delay is detected under a framework contract, the system can automatically generate purchase orders from a retail network. When the previous production stage required less MTP than planned, changes are made throughout the supply chain. The smartest systems can take other factors into account. For example, a decrease in ambient air temperature will lead to an increase in electricity consumption for boiler houses and additional material consumption for maintenance and repair of equipment that is serviced on time. Thus, the statistical sustainability mode is implemented, when the instability of each instance of the process makes it possible to implement a stable provision for the production of MTR. Controlling this process is more like the stability/instability of a fighter plane than the stability of a transport plane.

We see that the planning and execution of MTP applications is becoming more individual and less predictable, that is, there is an increase in diversity in the flow of MTP applications (LS-mode). The planning system manages not only to rapidly execute a new application and, if necessary, quickly adjust plans to already running applications, that is, dispels a variety of flow applications (HS-mode). This MTP planning system ensures the parity of growth and dispersion of the diversity of the application flow and makes it statistically stable. Interestingly, it is in this mode of statistical sustainability that elements of self-organization begin to manifest themselves: the system begins to "suggest" new possible planning methods that are specific to this business and are very different from the widely used MRP II models.

The existence of a regime of statistical sustainability of a digital organization is not a problem, but an objective property that it is pointless to fight. Statistical sustainability has to be learned to live with, and it looks like it could become the new norm for digital organizations.

The study of the nature of exponential organizations [1] clearly shows that their explosive growth is always accompanied by an increase in variability, which constantly brings them to the edge of the stability of their business. Exponential organizations have learned to live in a statistical sustainability mode that allows them to balance on the edge of sustainability and maintain a high level of innovation in their business. It is in the mode of statistical sustainability, through trial and error, that these companies have managed to find the sources of their explosive growth. Interestingly, exponential organizations consciously maintain a statistical sustainability regime even as they transform from startups to large companies. Even when exponential organizations stabilize their individual operational business processes, they still maintain a high readiness to transition to statistical sustainability and readiness for change.

Conclusion

Systematic digitalization puts an organization in a state of statistical sustainability, when the predictability of the actions of its individual participants begins to disappear, but the predictability as a whole still persists. From the point of view of traditional management, focused on long – term predictability and classical stability of the forms of organization, statistical sustainability is a chaos and a source of risks that must be dealt with. From the point of view of the management of digital organi-

zations, focused on a high level of variability and rapid growth, statistical sustainability is a generator of new practices and a source of new opportunities. In this mode, the organization can balance on the edge of sustainability and maintain a high level of variability and innovation.

Digitalization is often perceived as a traditional reengineering of business processes, but only on the basis of new technologies. In this case, statistical sustainability is traditionally considered as a source of risks and they begin to fight it. In fact, statistical sustainability is a necessary condition for both the transformation into a digital organization and its further life activity, so it must be stimulated and supported.

Digitalization encourages the development of the instability of the organization, and it pushes it to the edge of a cliff. But it is here, on the edge of the cliff, that its new nature is revealed: it turns out that the organization is not a machine, but an organism to which management will have to adapt. If nature is for us, who is against us? Woe to those who struggle with nature. ■

References

1. Ismail S., Malone M.S., van Geest Y. (2014) *Exponential organizations: Why new organizations are ten times better, faster, and cheaper than yours (and what to do about it)*. Diversion Books.
2. Schwab K. (2017) *The fourth industrial revolution*. London: Penguin Books.
3. Agamirzyan I.R., et. al (2016) *Challenge 2035*. Moscow: Olymp Business (in Russian).
4. PwC (2016) *Industry 4.0: Building the digital enterprise. 2016 Global Industry 4.0 Survey*. Available at: https://www.pwc.com/gx/en/industries/industries-4.0/landing-page/industry-4.0-building-your-digital-enterprise-april-2016.pdf (accessed 15 October 2020).
5. The Government of the Russian Federation (2017) *The program "Digital economy of the Russian Federation."* Approved by Order of the Government of the Russian Federation No 1632-r of 28 July 2017. Available at: http://static.government.ru/media/files/9gFM4FHj4PsB79I5v7yLVuPgu4bvR7M0.pdf (accessed 15 December 2020) (in Russian).
6. Ananyin V.I., Zimin K.V., Lugachev M.I, Gimranov R.D., Skriprin K.G. (2018) Digital organization: Transformation into the new reality. *Business Informatics*, no. 2, pp. 45–54. DOI: 10.17323/1998-0663.2018.2.45.54.
7. Ananyin V.I., Zimin K.V., Gimranov R.D., Lugachev M.I, Skriprin K.G. (2019) Real time enterprise management in the digitalization era. *Business Informatics*, vol. 13, no 1, pp. 7–17. DOI: 10.17323/1998-0663.2019.1.7.17.
8. Taleb N.N. (2007) *The black swan: The impact of the highly improbable*. New York: Random House.
9. Korotkov E.M. (2003) *Anti-crisis management*. Moscow: INFRA-M (in Russian).
10. Sheremet A.D., Negashev E.V. (2016) *Methodology of financial analysis of commercial organizations' activities*. Moscow: INFRA-M (in Russian).
11. The Russian Federation (1994) *Federal Law of the Russian Federation No 68-FZ of 21 December 1994 "On the protection of the population and territories from natural and man-made emergencies."* Available at: http://www.consultant.ru/document/cons_doc_LAW_5295/ (accessed 15 December 2020) (in Russian).
12. Gaponenko A.L., Savelieva M.V. (2020) *Theory of management*. Moscow: Urait (in Russian).
13. Prigogine I., Stengers I. (1984) *Order out of chaos: Man's new dialogue with nature*. Bantam New Age Books.
14. Haken H. (2006) *Information and self-organization: A macroscopic approach to complex systems*. Berlin, Heidelberg: Springer.

15. Kurdyumov S.P., Malinetsky G.G., Potapov A.B. (1989) *Synergetics – new directions.* Moscow: Znanie (in Russian).

16. Malinetsky G.G. (2017) *Mathematical foundations of synergetics: Chaos, structures, computational experiment.* Moscow: URSS (in Russian).

17. Maturana H.R., Varela F.J. (1980) *Autopoiesis and cognition: The realization of the living.* Dordrecht: D. Reidel Publishing.

18. Vasilkova V.V. (1999) *Order and chaos in the development of social systems: Synergetics and theory of social self-organization.* St. Petersburg: Lan' (in Russian).

19. Kershengolts B.M., Chernobrovkina T.V., Shein A.A., Khlebhy E.S., Anshakova V.V. (2009) *Nonlinear dynamics (synergetics) in chemical, biological, and biotechnological systems.* Yakutsk: Ammosov Yakut State University (in Russian).

About the authors

Vladimir I. Ananyin
Senior Lecturer, Department on Business Processes Management, Russian Presidential Academy of National Economy and Public Administration, 82 build.1, Prospect Vernadskogo, Moscow 119571, Russia;
E-mail: v.ananiin@gmail.com

Konstantin V. Zimin
Editor-in-Chief, Information Management Journal;
Member of the Board, The Russian Union of CIO, 34, Seleznevskaya Street, Moscow 123056, Russia;
E-mail: konst.zimin@gmail.com

Mikhail I. Lugachev
Dr. Sci. (Econ.), Professor;
Academic Supervisor of Department of Economic Informatics, Lomonosov Moscow State University, 1 build. 46, GSP-1, Leninskie Gory, Moscow 119991, Russia;
E-mail: mlugachev@gmail.com
ORCID: 0000-0002-6871-3328

Rinat D. Gimranov
Head of IT Department, PJSC Surgutneftegaz, 1 block 1, Grigoriya Kukuevitskogo Street, Surgut 628415, Russia;
E-mail: gimranov_rd@mail.ru

DOI: 10.17323/2587-814X.2021.1.59.77

Simulation of development of individual heavy industry sectors*

Evgeniy V. Kislitsyn
E-mail: kev@usue.ru

Victor V. Gorodnichev
E-mail: helltoaster@yandex.ru

Ural State University of Economics
Address: 62, 8 Marta Street, Yekaterinburg 620144, Russia

Abstract

Nowadays, in the context of the coronavirus crisis, the issue of ensuring the sustainable development of heavy industries is acute. However, theoretical and analytical researches alone are not sufficient for this, and economic science needs to develop fundamentally new approaches to the study of the development of industrial sectors. This article is devoted to the creation and testing of a simulation model for the development of individual sectors of the economy. The object of research is the metallurgical industry, as well as related ore mining, mechanical engineering and production of finished metal products. The theoretical basis of the research is a systematic approach that combines the theory of industry markets, economic growth, industrial economics, system dynamics and mathematical economics. The main research methods used are system analysis, statistical analysis to identify trends in changes in the main economic indicators, econometric modeling to build production functions, as well as mathematical modeling of macroeconomic systems. As a result, a simulation model developed in system dynamics notation is proposed, which makes it possible to evaluate the development of individual industries taking into account various changes. This model is built on the basis of the three-sector model of the national economy, where separate adjacent industries connected by dynamic feedback loops are identified as structural elements. The paper details the structure of the simulation model based on first-order dynamic equations, balance equations and nonlinear production functions. The simulation model allowed us to predict a number of scenarios for the development of metallurgical industries, taking into account changes in the labor force and investment in fixed assets. The results of the work can be used for forming proposals on industrial policy, monitoring the condition and efficiency of individual industries.

Key words: simulation; system dynamics; industry; metallurgy; three-sector model of the economy; production function; mathematical model.

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

Citation: Kislitsyn E.V., Gorodnichev V.V. (2021) Simulation of development of individual heavy industry sectors. *Business Informatics*, vol. 15, no 1, pp. 59–77. DOI: 10.17323/2587-814X.2021.1.59.77

Introduction

Sustainable development of heavy industries remains a top priority for the national economy. Mechanisms for ensuring such development form the basis of the industrial policy of the Russian Federation, individual national projects and programs [1]. Therefore, the task of finding new ways, methods and tools to ensure and support the sustainable development of heavy industry sectors is most relevant. In modern Russian and foreign studies, both classical economic [2, 3] and new interdisciplinary approaches [4, 5] are used to deal with these tasks. However, it is almost impossible to develop such tools without using modern mathematical tools and computing systems.

For the most part, analytical studies in this area boil down to the analysis of the current situation in individual industries [6, 7], regions [8–10] and the economy as a whole [11], as well as to the construction of econometric models and interpretation of the dependencies found for further adjustments to industrial policy. However, such models practically do not take into account, firstly, the nonlinear behavior of economic systems and, secondly, the presence of causal relationships between individual indicators of these systems. The use of mathematical and simulation modeling in the research of industries is encountered in modern studies more and more often, however, the models proposed by the authors are not always theoretically substantiated and empirically tested.

In this study, an attempt is made to use the ideas of mathematical economics and, in particular, the modeling of macroeconomic dynamic systems in combination with simulation modeling to study the development of sectoral economic systems. This synthesis is possible through the use of causal and flow simulation notations. A feature of the proposed approach is the use of methods of system dynamics first proposed in the works of J. Forrester [12, 13] and D. Meadows [14, 15], and subsequently developed in the works of V.N. Sidorenko [16], A.S. Akopov [17–19] and other researchers. Dynamic models of macroeconomic systems, as a rule, are based on differential equations of the first order, which can be represented in the notation of system dynamics using flow stratification.

The aim of this study is to develop a simulation model for the development of individual industries using the example of metallurgy and related industries (mining of metal ores, mechanical engineering, production of finished metal products), using the ideas of a three-sector model of the economy and system dynamics. In accordance with the purpose of the study, the following tasks were set and solved:

1) to analyze the available research in the field of simulation modeling of economic systems;

2) to modify the three-sector mathematical model of the economy in relation to individual industries;

3) to create the structure of a simulation model for the development of individual industries, to determine the main storage devices, flows, variables and parameters of the model;

4) to test the simulation model so created and simulate six different scenarios for the development of the metallurgical and related industries.

1. Theoretical background for study of mathematical and simulation models of industries

For studying long-term trends and growth factors, as well as for evaluating the consequences of various macroeconomic decisions, non-linear small-sector models are used. In such models, the structure of the economy is presented in the form of sectors, each of which produces some aggregated product. The basis here is the one-sector Solow model [20], in which the economic system is considered as a kind of unstructured apparatus that produces a single universal product. This model reflects the process of reproduction and allows us to analyze the relationship between consumption and accumulation.

Further development of the one-sector Solow model led to the two-sector [21] and three-sector models [22]. The appearance of such models allows us not only to analyze the growth of the economy as a whole, but also to focus attention on its individual sectors. However, the development of these ideas mainly lies in the study of macroeconomic systems. At the same time, no less important is the problem of the development of individual sectors, first of all, industry. The ideas presented in the three-sectoral model [22, 23], according to the author, can be used to study economic growth not of the sectors of the economy as a whole, but of its individual branches, most closely related to each other. The only assumption that must be observed when modifying this model is the need to represent industries from three sectors of the economy – material, fund-creating, and consumer.

The mathematical tools presented in the three-sector model of the economy are a set of differential equations, balance equations, and nonlinear production functions. The analytical solution of the three-sector model of the economy for balanced economic growth is described in detail in [23, 24]. Within the framework of this study, it is proposed to use a fundamentally different approach – simulation modeling.

Simulation modeling is an experimental method for studying a real or projected system based on its simulation model, which combines the features of an experimental approach with specific conditions for using computer technology [25]. According to the accepted classification, there are three approaches to creating simulation models: discrete-event modeling, system dynamics, and agent-based modeling [26].

Discrete-event simulation is mainly used to analyze business processes and production systems in enterprises, where logic can be described as a sequence of specific actions. For example, in [27], simulation is used to assess the risks of industrial enterprises. This simulation model is a process (discrete-event) model of the production process, on which statistical tests were carried out and the variation in the value of the profitability indicator was investigated depending on various input factors.

When describing meso- and macroeconomic systems, it is rational to use two other approaches – system dynamics and agent-based modeling. At the same time, agent-based modeling is used in the case when in the system, in an explicit form, separate subjects are distinguished, with such characteristics as possessing discreteness, sociality, activity, autonomy and flexible behavior [28]. In particular, a stochastic model of economic growth with four types of agents: production (firm), consumer, foreign markets and the state is presented in the work of A.V. Vorontsovsky [29]. This study examines a single-product model with domestic and foreign producers which uses the classic Cobb-Douglas production function and considers production conditions, the consumer's problem and relationships for the state as the main governing elements. The work of D.N. Shultz and I.N. Yakupova [30] is devoted to solving the classical problem of agent-based modeling – assessing how the behavior of each of the agents affects the development of

the system as a whole. In particular, this work examines the influence of microstructures on the properties of the economy as a whole.

System dynamics models make it possible to identify cause-and-effect relationships in the economic system. They usually consist of accumulators (levels), which are accumulations in feedback circuits, flows that regulate the rate of change of accumulators, auxiliary variables and cycles (feedback loops).

System dynamics models are based on differential equations, which justifies their use when trying to breathe new life into classical dynamic (continuous) economic models. N.V. Yandybaeva and V.A. Kushnikov [31] developed a complex system-dynamic model for predicting the indicators of the economic security of the Russian Federation. This model contains 11 systemic levels, including the growth rate of consumer prices, the unemployment rate, etc., as well as four functional dependencies and several dozen auxiliary variables.

E.I. Piskun [32] used system dynamics to build a model of the mutual influence of variables for assessing the effectiveness of the strategy of innovative development of production and economic systems, using several scenarios for the development of the system for the analysis. The model calculates indicators such as the average chain growth rate, as well as the average growth rate of levels of innovation intensity, efficiency of investment activities, innovation potential and development of production and economic systems.

Particular attention is paid to studies of the interaction between economic growth and resource consumption, which can also be expressed using differential mathematical models. Thus, in [33], a model in the notation of system dynamics was developed in which economic growth and resources are considered from a systemic point of view in order to identify the dependence of economic growth on delayed feedback effects from resource depletion [34]. The model consists of two subsystems – renewable natural resources and economies, where the level of resources and GDP are used as accumulators.

G.L. Beklaryan [35] developed a simulation model of the economic development of the Far Eastern Federal District which uses the methods of system dynamics and the agent-based approach. The main idea of the model is to study the influence of a number of factors on the economy of the federal district, such as the growth rate of investment in fixed assets, average wages, subsidies from the federal budget and prices for natural resources. Macroeconomic indicators in the model are described in the notation of system dynamics, and the regions of the federal district are presented as individual agents in one of three states.

Modeling the processes of economic dynamics by industry is presented in [36], where the main modules of the model are public administration, production, consumption, employment, demography, education and the financial system. As the main methods in this study are used agent-based modeling, system dynamics, intersectoral balance models, elements of artificial intelligence and cognitive technologies. The model is based on the determination of the need for various resources (goods, personnel, etc.), gross fixed capital formation and investment demand. In this simulation model, households and legal entities of various forms are represented as agents [37], whose behavior is modeled using special algorithms described in the notation of UML diagrams. System dynamics is used to model the budgets of the state, legal entities and households, as well as to reproduce the dynamics of the population [38].

Another attempt to use system dynamics in the analysis of industries is presented in [39], where a simulation model of the region's supply chain system was developed. The model consists of four main interconnected blocks: population, agriculture, manufacturing and transport. The model made it possible to identify the cyclicality and interdependence of

the efficiency of the main industries and their influence on each other [40].

Thus, a large number of economic studies use the tools of simulation modeling and, in particular, system dynamics in the analysis of macroeconomic systems. Nevertheless, such models consider either the national economy as a whole or its individual enlarged sectors. Within the framework of this study, an attempt was made to apply the tools of system dynamics to study the model of sustainable development of a separate group of industries, based on a three-sector model of the economy.

2. Mathematical and algorithmic basis for building a simulation model of the industries' development

Within the framework of this study, the objects of research are the branches of metallurgy, production of finished metal products, ore mining and mechanical engineering. The choice of the object of research was not made by chance, since it is metallurgy that is one of the leading industries in the Ural macroregion. In accordance with the thesis on the distribution of industries in three sectors, the following are highlighted (in accordance with the classification according to OKVED 2.0):

- material sector: 07 – Mining of metal ores; 24 – Manufacturing of basic metals;
- fund-creating sector: 28 – Manufacturing of machinery and equipment not included in other groupings (including: 28.4 – Manufacturing of metal forming machinery and machine tools; 28.91 – Manufacturing of machinery for metallurgy);
- consumer sector: 25 – Manufacturing of fabricated metal products, except machinery and equipment.

Other industries are not directly considered in this model, but are counted as a kind of "other" industry. The material sector is represented here by the production of objects of labor (iron ores, ores of non-ferrous metals, cast iron, steel, ferroalloys, steel pipes, precious and non-ferrous metals, casting), labor instruments (machinery and equipment) are considered as a fund-creating sector, and consumer goods are consumption (finished metal products). It should be noted here that consumers can be both other sectors of the economy and households. Within the framework of this model, only the metallurgy sector and the industries closest to it are considered.

The technological structure in this model is considered constant, and the output in the industries is set by neoclassical production functions:

$$X_i = A_i \cdot K_i^{\alpha_i} \cdot L_i^{\beta_i}, \quad (1)$$

where X_i – output volume in the i-th industry;

A_i – coefficient of the neutral technical process in the i-th industry;

K_i – fixed production assets (hereinafter – FA) of the i-th industry;

L_i – number of people employed in the i-th industry;

α – fund elasticity coefficient;

β – labor elasticity coefficient;

i – industry number ($i = 7, 24, 25, 28$).

The model distinguishes the total number of employees, which changes with a constant growth rate:

$$\frac{dL}{dt} = \vartheta L, \quad (2)$$

where L_i – number of people employed in the i-th industry;

ϑ – employment growth rate.

The distribution of workers in industry is presented as a balance ratio:

$$L = L_7 + L_{24} + L_{25} + L_{28} + L_{oth}, \quad (3)$$

where L_i – number of people employed in the i-th industry;

L_{oth} – number of people employed in other industries.

The model makes assumptions about the absence of a lag of capital investments and the constancy of the wear rates of fixed assets. Hence, the change in the FA of the *i*-th industry consists of depreciation and growth due to gross investment:

$$\frac{dK_i}{dt} = -\mu_i K_i + I_i, \quad (4)$$

where K_i – FA of *i*-th industry;

μ – wear factor of FA in the *i*-th industry;

I_i – gross investment in the *i*-th industry.

Capital investments in FA of all industries are the purchase of machinery, equipment, buildings, structures, etc. Thus, the output of sectors 28 and 41 together constitutes the total volume of investments that can be invested in different sectors:

$$X_{28} + X_{41} = I_7 + I_{24} + I_{25} + I_{28} + I_{oth}, \quad (5)$$

where X_i – the volume of output in the *i*-th industry;

I_i – gross investment in the *i*-th industry;

I_{oth} – gross investment in other industries.

At this stage of the study, sector 41 "Construction of buildings" is not included in the model; therefore, it is considered constant.

Another balance ratio shows the distribution of products of the branches of the material sector (in particular, metallurgy):

$$X_{24} = a_0 X_{24} + a_{25} X_{25} + a_{28} X_{28} + a_{oth} X_{oth}, \quad (6)$$

where X_i – output volume in the *i*-th industry;

a_i – coefficients of direct material costs.

Thus, the mathematical model consists of five first-order dynamic elements, three static distribution elements and four non-linear static elements. Endogenous variables are FA and industry outputs. Exogenous variables are the growth rate of the number of employees, the wear rates of the FA of industries, the coefficients of direct material costs, the initial value of the number of employees, the initial distribution of those employed by industries, the initial values of the OPF of industries, and the parameters of production functions. Accordingly, the management in the model is carried out by allocating labor and investment resources.

Relying on equations (1)–(6) and the basic postulates of system dynamics, the author has designed and developed a simulation model for the development of metallurgical and related industries in the AnyLogic software environment. This model is designed in accordance with the following principles:

Dynamic elements are presented in the form of flow diagrams and accumulators, each dynamic element represents a bin, and the flow sets the rate of change of these accumulators.

Non-linear static elements are specified in the form of functional variables, the values of which are calculated based on the values of other variables, parameters and accumulators.

Static distribution elements are specified as parameters, the values of which are adjusted by the user of the model.

Thus, in the developed simulation model, there are five flow diagrams, four of which are almost identical in appearance (*Figure 1*).

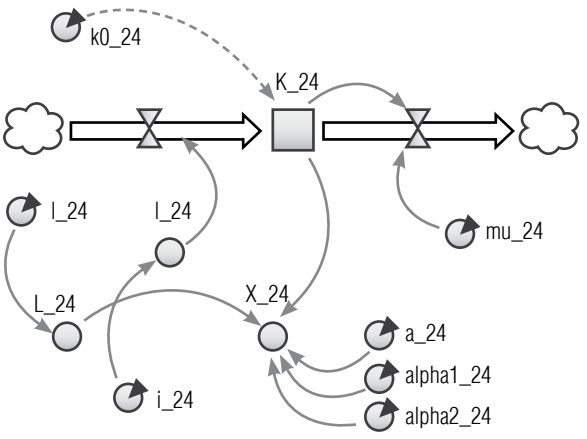


Fig. 1. Flowchart of the metallurgical industry

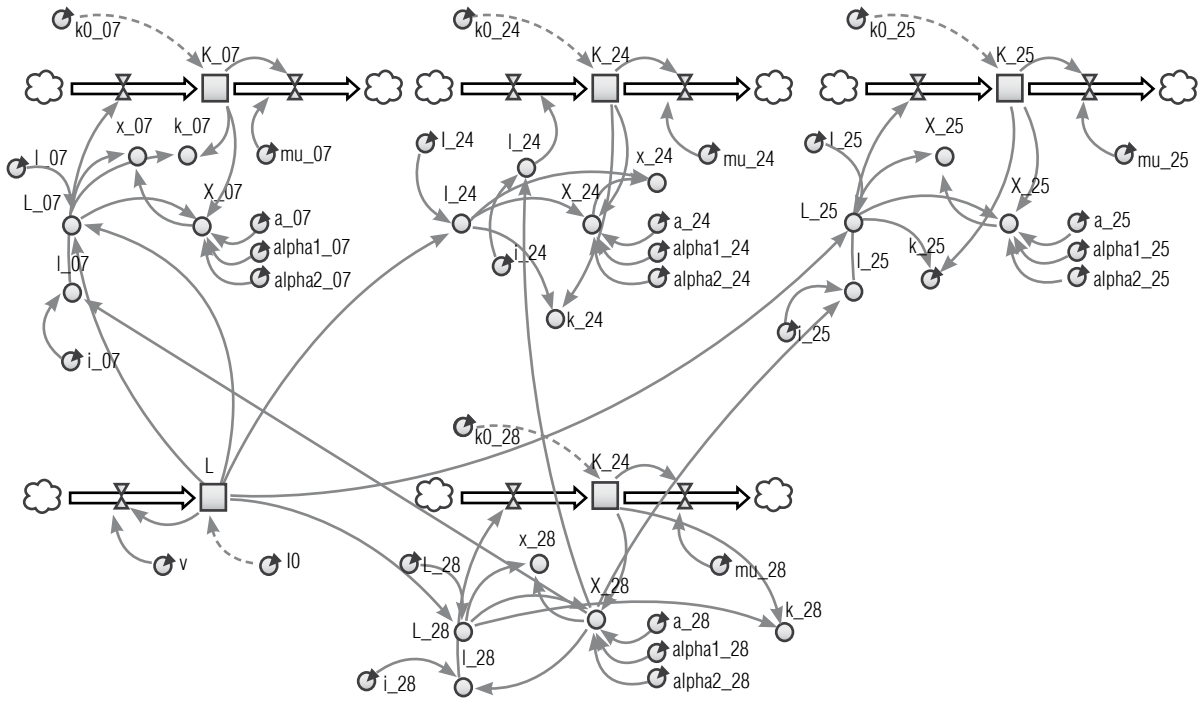


Fig. 2. General view of the simulation model

In general, the simulation model is a set of flow and cause-effect diagrams, where each element is calculated dynamically, in continuous time. The general view of the simulation model is shown in *Figure 2*.

As accumulators in the model, the volume of fixed assets (capital) of each of the studied industries and the number of those employed in industry are presented, calculated according to the given differential equations (2) and (4) using special elements – flows. Threads set the rate of change of drives per unit of time in a dynamic way, i.e. continuously. With the help of dynamic variables, the output of the industry, as well as the number of employees and the volume of investments, are described (*Table 1*).

Parameters in the simulation model denote values that are set and changed by the user (not functionally). These are data obtained statistically and empirically, calculated using mathematical and econometric methods, and loaded into the model. The static parameters are mainly the initial values of the storage devices and the coefficients of the production functions (*Table 2*).

Thus, the structure of a simulation model for the development of several industries is described, including metallurgy, mining of metal ores, mechanical engineering and the production of finished metal products.

3. The results of simulation modeling of the development of the metallurgical industry

To test the model, it is necessary to calculate the values of the exogenous variables (parameters) of the model and set equations (1)–(6) with the corresponding variables. In order to calculate the coefficients A_i, α_i and β_i, production functions are constructed for each industry. To build production functions, the official

Table 1.

Endogenous variables of the simulation model

Variable name	Variable type	Interpretation of the variable
K_i	Stock	Capital volume of the *i*–th industry
L	Stock	Number of people employed in industry
flowInvest_i	Flow	Increase in investments in fixed assets of the *i*–th industry
flowVyb_i	Flow	Retired fixed assets of the *i*–th industry
flow_PrirLabor	Flow	Increase in those employed in industry
X_i	Dynamic variable (function)	Gross output of the *i*–th industry
L_i	Dynamic variable	The number of people employed in the *i*–th industry
I_i	Dynamic variable	The volume of investments in the *i*–th industry
x_i	Dynamic variable	Labor productivity of the *i*–th industry
k_i	Dynamic variable	Capital–labor ratio of the *i*–th industry

Table 2.

Exogenous model variables

Variable name	Variable type	Interpretation of the variable	Calculation method
k_i	Parameter	The initial value of the volume of fixed assets of the *i*–th industry	Taken from statistical sources for the period preceding the forecast
mu_i	Parameter	The rate of retirement of fixed assets of the *i*–th industry	Calculated according to statistical data
l_i	Parameter	Share of people employed in the *i*–th industry in the total number of people employed in industry	Calculated according to statistical data
i_i	Parameter	The share of investments in fixed assets of the *i*–th industry in the total output of mechanical engineering products	Calculated according to statistical data
a_i	Parameter	Ratio of scientific and technological progress of the *i*–th industry	Calculated by the production function of the *i*–th industry
*alpha*1_*i*	Parameter	Capital elasticity coefficient of the *i*–th industry	Calculated by the production function of the *i*–th industry
*alpha*2_*i*	Parameter	Labor elasticity coefficient of the *i*–th industry	Calculated by the production function of the *i*–th industry
*l_*0	Parameter	The initial value of the number of persons employed in industry	Taken from statistical sources for the period preceding the forecast
v	Parameter	Growth rate of those employed in industry	Calculated according to statistical data

data of the Federal State Statistics Service for the period from 2002 to 2018 were used. Unfortunately, it was not possible to use the earlier data due to their incomparability. However, a period of 17 years is quite possible to build a regression model with two explanatory variables at a high level of significance. Values have been recalculated to 2002 comparable prices.

As indicators X_i, were indicators of the volume of shipped goods of own production for each of the studied industries: mining of metal ores, metallurgical production, production of finished metal products (except for machinery and equipment), production of machinery and equipment not included in other groups. The K_i indicator is calculated as the value of the indicator of fixed assets of commercial organizations for each of the industries. The L_i indicator is calculated as the value of the indicator of the average annual number of employees of organizations (*Table 3*)[1].

As a result of calculations using the R programming language and the RStudio package, the following production functions of the studied industries were obtained:

$$\begin{aligned} X_{07} &= K_{07}^{0.89} \cdot L_{07}^{0.28} \\ X_{24} &= 0.68 \cdot K_{24}^{0.74} \cdot L_{24}^{0.55} \\ X_{25} &= 5.9 \cdot K_{25}^{0.38} \cdot L_{25}^{0.49} \\ X_{28} &= K_{28}^{0.97} \cdot L_{28}^{0.22}. \end{aligned} \quad (7)$$

Table 3.

Descriptive statistics of the initial data for the construction of production functions

Index[2]	Mean	Standard error	Median	Minimum	Maximum	Dispersion	Standard deviation
X_07	407056.8	46944.1	988000.0	343633.8	484446.1	2034227172	45102.4
X_24	1398471.5	220461.5	2768500	1061710	1837724.9	4.4865E+10	211812.6
X_25	597651.1	91035.0	1186500	455018.6	768581.6	7649879809	87463.6
X_28	529757.8	80704.7	1125420	376389	646794.0	6012228031	77538.6
K_07	368006.7	85864.8	697009	274479	548720.3	6805624929	82496.2
K_24	708306.6	139333.0	1320351.5	570974.3	984168.4	1.792E+10	133866.8
K_25	150132.3	29339.2	280074.6	121115.8	210779.5	794574579	28188.2
K_28	171378.7	12066.6	354314	153625.6	191204.4	134402506	11593.2
L_07	245.9	85.7	1090000	167	357.5	6781.9	82.4
L_24	557.0	108.3	1076803	440.7	731.8	10822.8	104.0
L_25	393.7	74.4	513309	292.9	487.8	5114.5	71.5
L_28	742.0	313.2	13.1	299.4	1205.0	90548.1	300.9

[1] To calculate the indicators, the following statistical collections were used: "Industrial production in Russia – 2019", "Industrial production in Russia – 2016", "Industrial production in Russia – 2010", "Industrial production in Russia – 2002"

[2] Indicators X_i and K_i are calculated in millions of rubles, indicators L_i – in thousands of people

It should be noted here that the sum of the coefficients of elasticity for funds and labor is greater than one, which indicates the presence of increasing returns. This thesis is based on the results of the research by S.V. Orekhova and E.V. Kislitsyn [41] devoted to the analysis of the aggregate productivity of factors, which is understood as the unexplained remainder of the growth of the final product. In particular, the study proved the presence of the aggregate productivity of factors in the heavy industry sectors, including metallurgy, which suggests the presence of increasing returns.

All constructed functions and their coefficients are statistically significant at the level of 5%, the coefficients of determination range from 0.4 for metallurgy to 0.99 for mechanical engineering. The values of all exogenous variables (parameters) described in *Table 2* are calculated and presented in *Table 4*.[3]

The main idea of simulation models is to simulate various situations that arise in the system, i.e. simulate scenarios. Within the framework of this work, six different scenarios for the development of metallurgy and related industries were investigated. The first scenario is the development of the economic system without significant changes in accordance with the given production functions. In this scenario, it is assumed that all the parameters of the model are constants (i.e., all the growth rates of the main indicators are preserved). This scenario reflects the existing picture and allows us to predict the development of the economies of various industries. Labor productivity, capital–labor ratio (*Figure 3*) and gross output (*Figure 4*) are identified as the analyzed variables.

Table 4.

Model parameter values as of 2018

Variable name / Industry	07 – Mining of metal ores	24 – Manufacture of basic metals	25 – Manufacture of fabricated metal products, except machinery and equipment	28 – Manufacture of machinery and equipment not included in other groupings
k_i	1709200	3255000	699600	509900
mu_i	0.013	0.009	0.007	0.011
l_i	0.032127	0.043398	0.04877	0.039027
i_i	0.207959	0.250129	0.06384	0.033877
a_i	1	0.68	5.9	1
*alpha*1_*i*	0.89	0.74	0.38	0.97
*alpha*2_*i*	0.28	0.55	0.49	0.22
*l*_0	9887.1			
v	–0.00036			

[3] To calculate the indicators, the following statistical collections were used: "Industrial production in Russia – 2019", "Industrial production in Russia – 2016" and the database "SPARK–Interface"

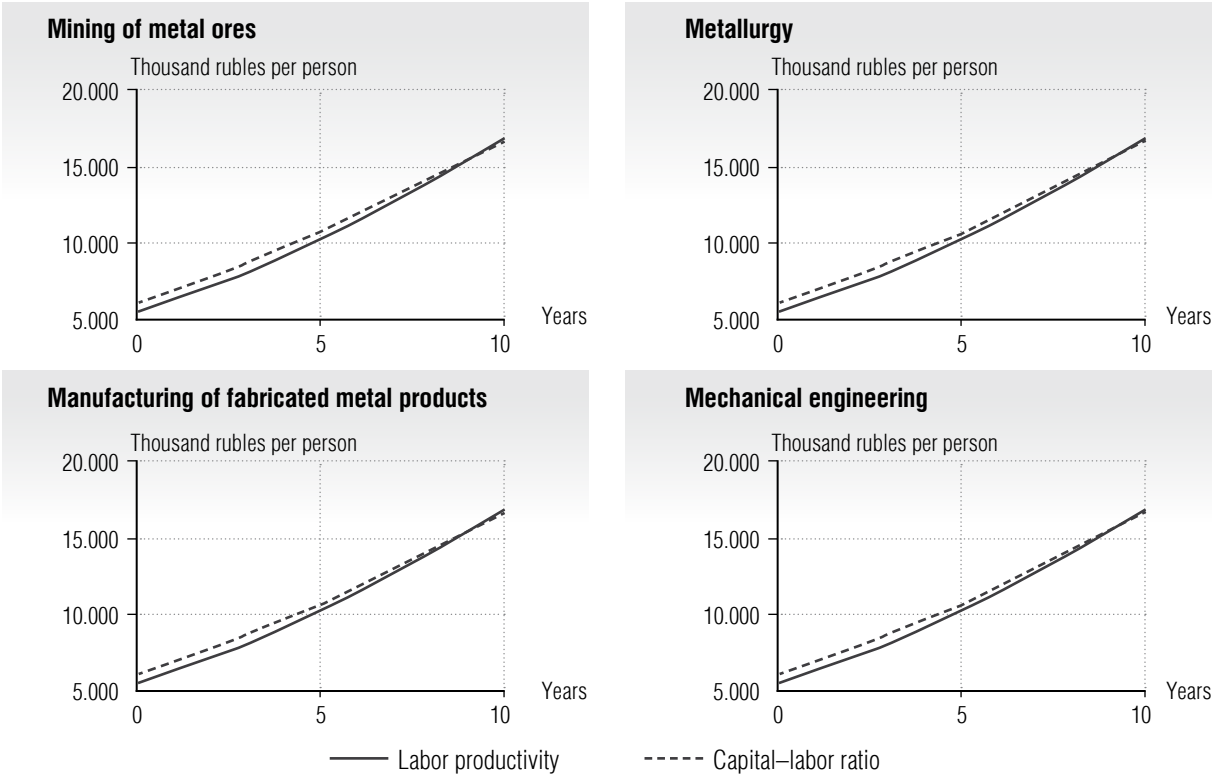


Fig. 3. Dynamics of changes in labor productivity and capital–labor ratio according to the scenario no 1

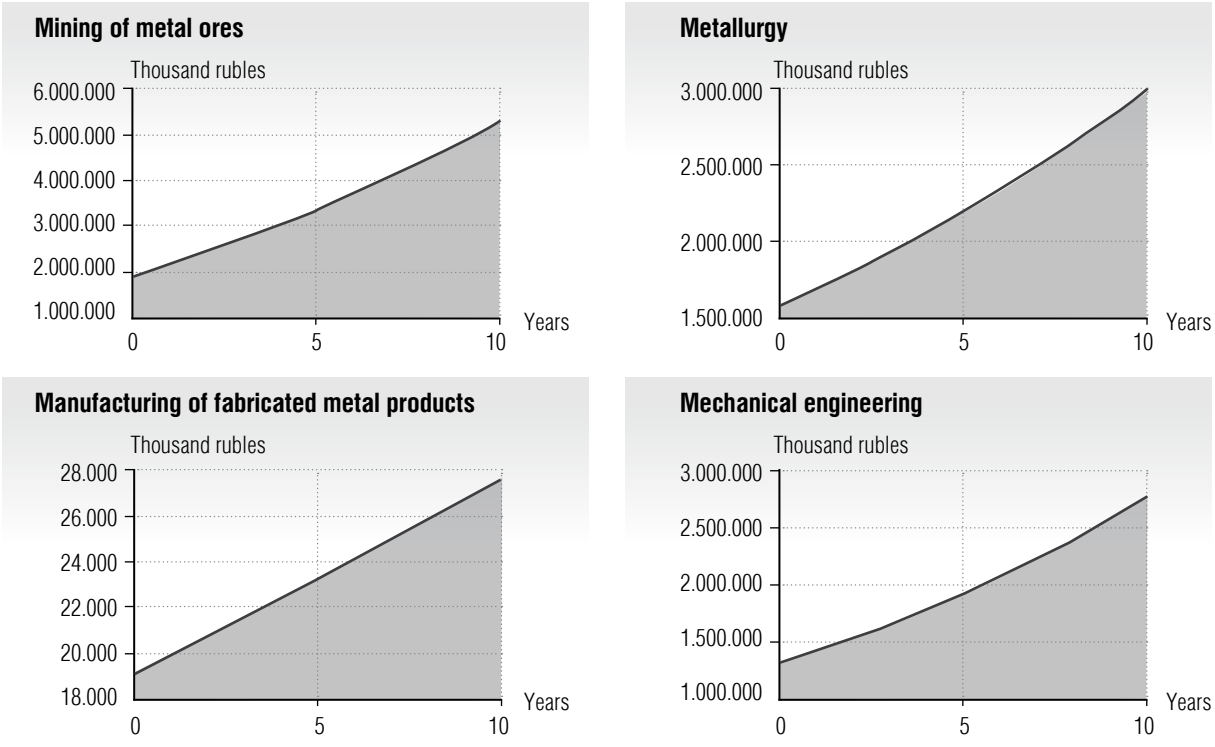


Fig. 4. Dynamics of changes in gross output according to scenario no 1

It should be noted right away that the total volumes of output in some industries may not reflect the actual values at the beginning of the analyzed period, since, firstly, production functions were based on prices in 2002 and, secondly, the initial value of output was not specified in this model. The main goal here is to track the growth rates of individual industries.

So, with the existing growth rate of the number of workers in industry (and as of 2018 this rate was –0.36% per year) and the growth rate of capital investment in fixed assets, the capital–labor ratio of all studied industries gradually reaches the level of labor productivity in 10 years. With constant indicators, the growth in output in the mining industry of metal ores will grow 2.78 times, metallurgy – 1.9 times, production of finished metal products – 1.44 times, and mechanical engineering – 1.95 times over 10 years.

Next, three pessimistic and two optimistic scenarios are analyzed. The second scenario is a reflection of the already existing trend towards a reduction in the number of workers in industry. The simulation model specifies a reduction in the growth rate of the number of workers in industry by 1% per year (*Figure 5*).

First of all, the reduction in the inflow of new labor into industry as a whole will affect the industries in which labor productivity affects output to a greater extent than the capital–labor ratio. According to the calculated values, output in the metallurgy industry will gradually slow down and practically stop in the ninth year. In the production of finished metal products, the situation is even more negative – already in the sixth year, production volumes will begin to decline. The metal ore mining and mechanical engineering industries will not suffer so much, though their production growth rates will also decline.

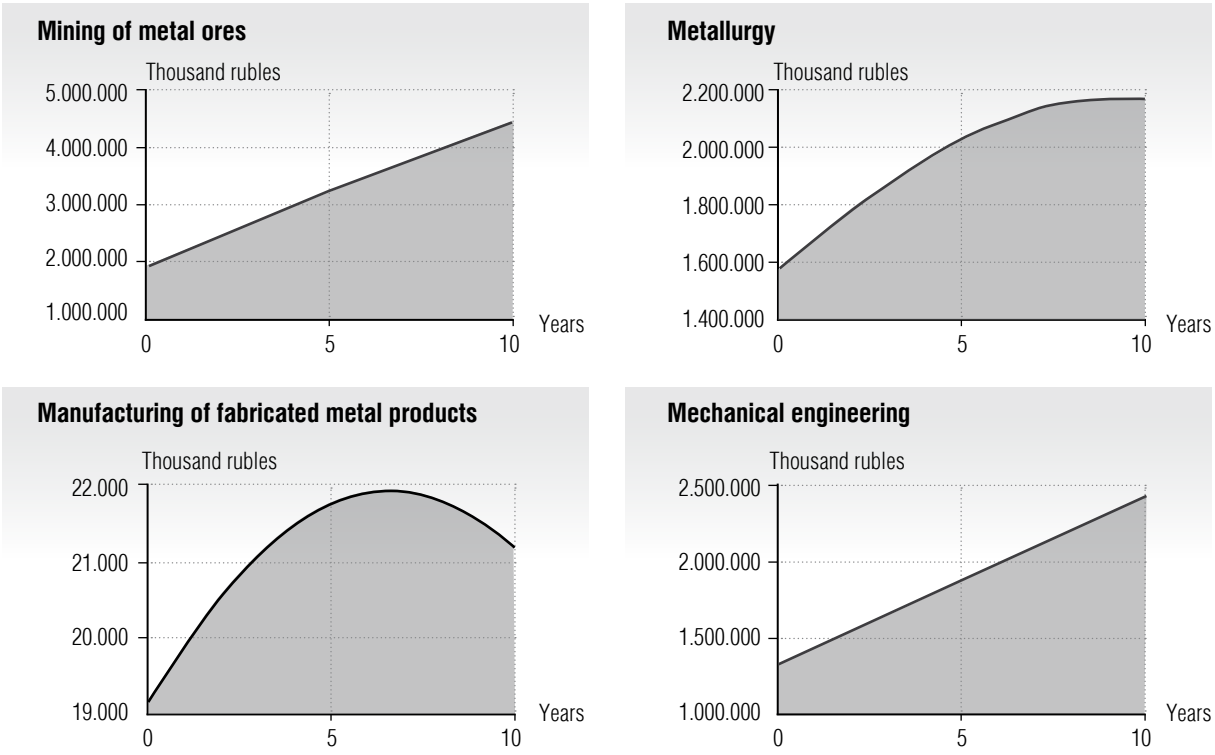


Fig. 5. Dynamics of changes in gross output according to scenario no 2

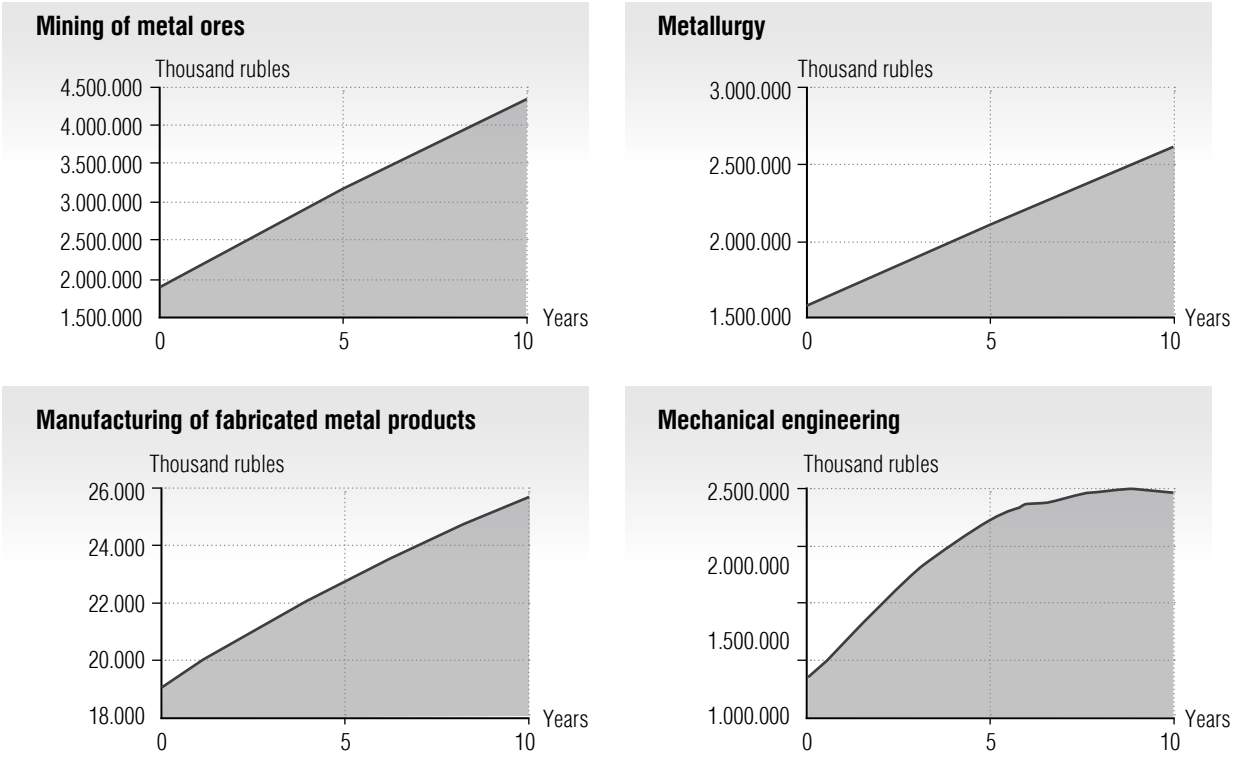


Fig. 6. Dynamics of changes in gross output according to scenario no 3

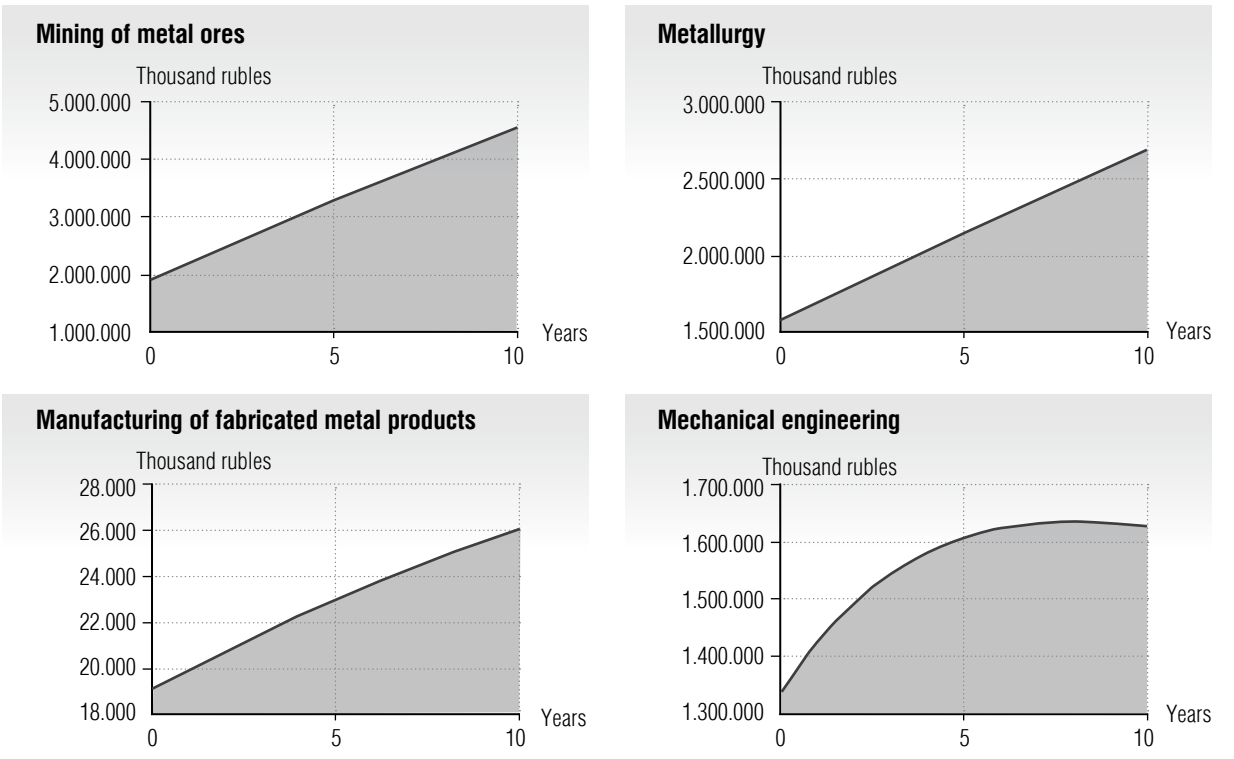


Fig. 7. Dynamics of changes in gross output according to scenario no 4

The third scenario is a reduction in the share of workers in the engineering industry in the total number of industrial workers by 25% per year. This scenario is an attempt to trace the causal relationship between the output of industries and the production of fixed assets (*Figure 6*).

It is easy to see that the rate of output of the machine-building industry has been slowing down quite clearly since the third year, while with a general decrease in the number of employees in all industries, such a trend was not observed. In addition, similar indicators in other industries are also declining. Thus, the growth rate of output in the metal ore mining industry will be only 2.28, metallurgy – 1.27, and in the production of finished metal products – 1.34 over 10 years, which is significantly lower than in scenario no 1.

The next scenario is a reduction in the share of investments in the mechanical engineering industry by 25% per year (*Figure 7*) in the total volume of investments in industrial production. As expected, there should be a trend similar to the third scenario. Nevertheless, while maintaining the general trends, there is still a quantitative difference with the previous scenario.

The output of the machine-building industry suffers more: its rates begin to decline already in the eighth year of the forecast. However, the output rates of other industries are falling more slowly. Here, the balance equation (5) affects, i.e. investments in other sectors remain the same. Even when the pace of production in the mechanical engineering industry begins to fall, other industries feel it less, which is very unusual. It turns out that the industries under study compensate the inflow of investments from other industries (for example, construction) to a greater extent than the outflow of workers from the engineering industry.

The fifth scenario is optimistic – an increase in the growth in the number of workers in industry while maintaining their distribution by industry by 1% per year (*Figure 8*).

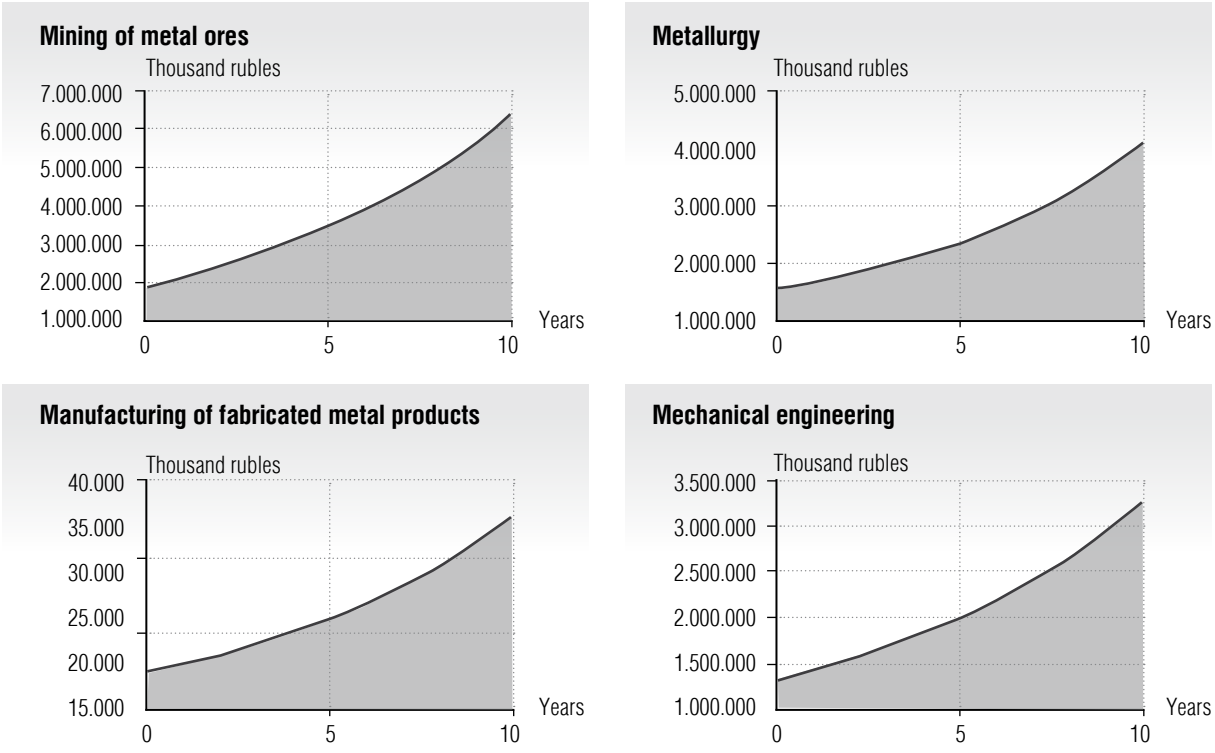


Fig. 8. Dynamics of changes in gross output according to scenario no 5

The increase in the number of employees has a positive effect on the growth rate of products in all sectors. Moreover, a feedback loop is actively starting to work, which links the output of fund-forming industries (in particular, mechanical engineering) with capital gains in other industries. Thus, the further we build our forecast, the greater the share in the increase in production output is taken by the increase in capital investments in fixed assets. This effect is called a self-replicating feedback loop.

The last scenario in this study is devoted to the analysis of the development of industries with an increase in the share of investments in the mechanical engineering industry by 10% per year (*Figure 9*).

Naturally, in this scenario, there is also an increase in the growth rates of output in all sectors in comparison with the standard scenario no 1. However, it should be noted that, first of all, the output in the machine-building industry is greatly increasing – by 3.7 times over 10 years. However, in the metallurgical industries the acceleration of growth rates is slower in comparison with the scenario №5. It should be noted here that, firstly, the increase in labor resources as a whole increases the growth to a greater extent linearly and, secondly, the increase in investment in one industry according to the balance equation (5) implies their outflow (albeit small) from other industries. However, the exponential growth of the mechanical engineering industry in the longer term will entail similar growth in the metallurgical industries, which will subsequently exceed the growth rate in scenario 5. However, the developed simulation model cannot build sufficiently plausible forecasts for a period exceeding 10 years.

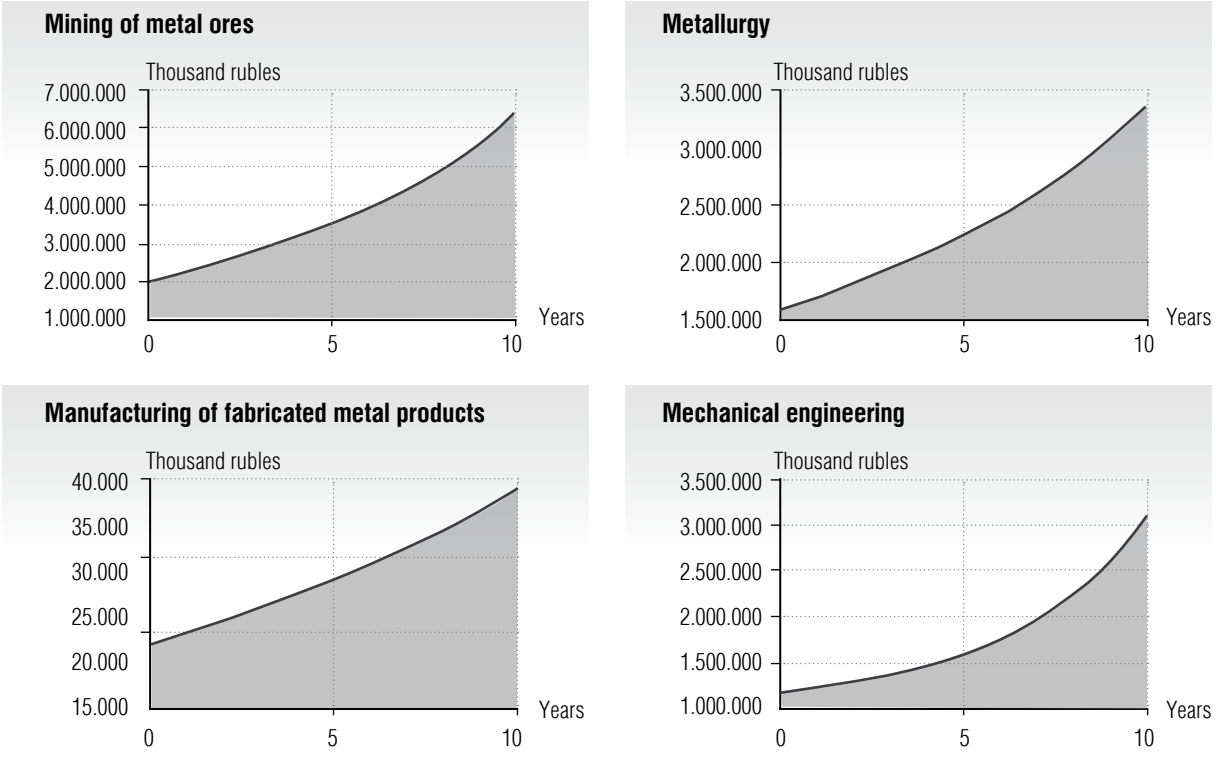


Fig. 9. Dynamics of changes in gross output according to scenario no 6

Thus, the possibility of using the developed simulation model for analyzing and forecasting the development of metallurgy and related industries has been demonstrated. Naturally, this set of scenarios is conditional. Nevertheless, the simulation model allows us to add new scenarios in accordance with the industrial policy of the Russian Federation.

Conclusion

In the course of this study, the following results were obtained.

Analysis of scientific works devoted to the construction of simulation models of economic systems showed that research in this area is extremely relevant, and their number is increasing every year. However, the use of system dynamics is mainly limited to studies of either the national economic system as a whole, or individual regions. Moreover, the development of individual industries also needs to be investigated with modern tools.

The three-sector model of the economy of V.A. Kolemaev, where the sectors are not the global sectors of the national economy but individual industries: within the framework of this work, these are metallurgy, iron ore mining, mechanical engineering and the production of finished metal products is modified. Five dynamic equations of the first order, three balance equations and four nonlinear functions are created, which together allow us to analyze, predict and optimize the sustainable development of individual industries.

On the basis of the analytical model of metallurgical industries, a simulation model of the development of individual industries in the notation of system dynamics has been developed. The accumulators (fixed production assets of each industry and the number of workers in the industry), changing their flows, as well as dynamic variables and model parameters are determined. The simulation model so constructed is a complex of flow causal diagrams that reflect not only the structure of the economic system under study, but also its behavior over several periods of time.

The constructed simulation model was tested taking into account the available data for 2002–2018. Production functions are constructed and all exogenous variables of the model are empirically calculated. To demonstrate the work of the simulation model, six scenarios for the development of the economic system under study were proposed - standard, three pessimistic and two optimistic. Scenario analysis showed the main points of management of the model and, as a consequence, the entire economic system being studied. The results of this study can be applied, first of all, by public authorities to adjust industrial policy. ■

References

1. Orekhova S.V. (2017) Resources and sustainable growth of an industrial metallurgical enterprise: an empirical assessment. *Modern Competition*, vol. 11, no 3, pp. 65–76 (in Russian).
2. Tretyakova E.A. (2013) Evolution of research and evaluation methodology of sustainable development of social and economic systems. *World Applied Sciences Journal*, vol. 25, no 5, pp. 756–759. DOI: 10.5829/idosi.wasj.2013.25.05.13335.
3. Orekhova S.V. (2017) Assessment of the economic growth stability of the metallurgical complex. *Vestnik NGUEU*, no 2, pp. 204–220 (in Russian).
4. Baranenko S.P., Dudin M.N., Lyasnikov N.V., Busygin K.D. (2014) Use of environmental approach to innovation-oriented development of industrial enterprises. *American Journal of Applied Sciences*, vol. 11, no 2, pp. 189–194. DOI: 10.3844/ajassp.2014.189.194.

5. Rudenko L.G., Zaitseva N.A., Larionova A.A., Chudnovsky A.D., Vinogradova M.V. (2015) Socio-economic role of service-sector small business in sustainable development of the Russian economy. *European Research Studies Journal*, vol. 18, no 3, pp. 219–234. DOI: 10.35808/ersj/468.

6. Dubrovsky V., Yaroshevich N., Kuzmin E. (2016) Transactional approach in assessment of operational performance of companies in transport infrastructure. *Journal of Industrial Engineering and Management*, vol. 9, no 2, pp. 389–412.

7. Orekhova S.V., Kislitsyn E.V. (2019) Small business and structural changes in industry. *Terra Economicus*, vol. 17, no 4, pp. 129–147 (in Russian). DOI: 10.23683/2073-6606-2019-17-4-129-147.

8. Demidova O.A., Ivanov D.S. (2016) Models of economic growth with heterogeneous spatial effects (on the example of Russian regions). *HSE Economic Journal*, vol. 20, no 1, pp. 52–75.

9. Surnina N.M., Ilyuhin A.A., Ilyuhina S.V. (2017) Demographic landscape of the region: factors, dynamics, trends, forecasts. *Journal of the Ural State University of Economics*, no 4, pp. 32–44 (in Russian). DOI: 10.29141/2073-1019-2017-16-4-3.

10. Barkhatov V.I., Benz D.S. (2019) Industrial markets of the Ural region: Economic growth under "new normal". *Upravlenets – The Manager*, vol. 10, no 3, pp. 83–93 (in Russian). DOI: 10.29141/2218-5003-2019-10-3-8.

11. Kulkov V.M., Kaimanakov S.V., Tenyakov I.M. (2014) The economic growth in Russia: national model, quality and security. *National Interests: Priorities and Security*, vol. 10, no 38, pp. 9–19 (in Russian).

12. Forrester J.W. (1969) *Urban dynamics.* Waltham, MA: Pegasus Communications.

13. Forrester J.W. (1958) Industrial dynamics: A major breakthrough for decision makers. *Harvard Business Review*, vol. 36, no 4, pp. 37–66.

14. Meadows D.H. (1972) *Limits to growth: A report for the Club of Rome's project on the predicament of mankind.* N.Y.: Universe Books.

15. Meadows D.H., Randers J., Meadows D.L. (2005) *Limits to growth: The 30 year update.* London: Earthscan.

16. Sidorenko V.N. (1998) *System dynamics.* Moscow: TEIS (in Russian).

17. Akopov A.S. (2006) System-dynamic approach to managing the investment activity of an oil company. *Audit and Financial Analysis*, no 2, pp. 165–200 (in Russian).

18. Akopov A.S. (2012) Designing of integrated system-dynamics models for an oil company. *International Journal of Computer Applications in Technology*, vol. 45, no 4, pp. 220–230. DOI: 10.1504/IJCAT.2012.051122.

19. Akopov A.S., Khachatryan N.K. (2014) *System dynamics.* Moscow: CEMI (in Russian).

20. Solow R.M. (1956) Contribution to the theory of economic growth. *Quarterly Journal of Economics*, vol. 70, no 1, pp. 65–94. DOI: 10.2307/1884513.

21. Intriligator M. (1971) *Mathematical optimization and economic theory.* N.Y.: Prentice-Hall.

22. Kolemaev V.A. (2005) *Mathematical economics.* Moscow: UNITY-DANA (in Russian).

23. Kolemaev V.A. (2008) Optimal balanced growth of the open three-sector economy. *Applied Econometrics*, vol. 11, no 3, pp. 15–42 (in Russian).

24. Trofimova V.Sh., Ozerova K.A. (2017) Economic and mathematical modeling of macroeconomic dynamics. *Mathematics and Computer Modeling in Economics, Insurance and Risk Management*, no 2, pp. 85–89 (in Russian).

25. Kislitsyn E.V. (2017) Principles of building a simulation model of a market with limited competition (for example, the market of mobile operators in Yekaterinburg). *Transbaikal State University Journal (Bulletin of ZabGU)*, vol. 23, no 10, pp. 101–110 (in Russian). DOI: 10.21209/2227-9245-2017-23-10-101-110.

26. Borshchev A., Karpov Y., Kharitonov V. (2002) Distributed simulation of hybrid systems with AnyLogic and HLA. *Future Generation Computer Systems*, vol. 18, no 6, pp. 829–839. DOI: 10.1016/S0167-739X(02)00055-9.

27. Bocharov E.P., Aleksenceva O.N., Ermoshin D.V. (2008) Assessment of risks of industrial enterprises on the basis of simulation modeling. *Applied Informatics*, no 1, pp. 15–24 (in Russian).

28. Kislitsyn E.V. (2018) Simulation of the process of crediting individuals using a credit rating. *Proceedings of Voronezh State University. Series: Economics and Management*, no 3, pp. 112–118 (in Russian).

29. Vorontsovskiy A.V. (2010) Modern approaches to modeling of the economic growth. *St. Petersburg University Journal of Economic Studies*, no 3, pp. 105–119 (in Russian).

30. Shults D.N., Yakupova I.N. (2016) Agent-based modeling of the influence of microstructure on economic properties. *Russian Journal of Economic Theory*, no 1, pp. 70–81 (in Russian).

31. Yandybaeva N.V., Kushnikov V.A. (2014) Mathematical model for forecasting of indicators of economic safety of the Russian Federation. *Vestnik of Astrakhan State Technical University. Series: Management, Computer Sciences and Informatics*, no 3, pp. 93–101 (in Russian).

32. Piskun E.I. (2014) The evaluation of effectiveness of the implementation of the strategy of innovative development of industrial and economic systems. *Science Journal of VolSU. Global Economic System*, no 5, pp. 55–68 (in Russian). DOI: 10.15688/jvolsu3.2014.5.6.

33. Mathur M., Agarwal S. (2015) *Sustainability dynamics of resource use and economic growth. A discussion on sustaining the dynamic linkages between renewable natural resources and the economic system. Discussion paper.* Available at: http://www.teriin.org/policybrief/files/aug15/files/downloads/Discussion_paper_Sustainability_Aug2015.pdf (accessed 15 December 2020).

34. Mathur M., Agarwal S. (2015) Dynamics of sustainable interaction between resource consumption and economic growth. *Karelian Scientific Journal*, no 4, pp. 48–59 (in Russian).

35. Beklaryan G.L. (2018) Decision support system for sustainable economic development of the Far Eastern Federal District. *Business Informatics*, no 1, pp. 66–75. DOI: 10.17323/1998-0663.2018.4.66.75.

36. Mashkova A.L., Savina O.A., Mamatov A.V., Novikova E.V. (2018) Computer modeling of sectoral economic dynamics. *Proceedings of the Southwest State University*, vol. 22, no 5, pp. 96–108 (in Russian). DOI: 10.21869/2223-1560-2018-22-5-96-108.

37. Mashkova A.L. (2016) Forecasting long-term development of macroeconomic systems based on agent modeling. *Public Administration. E-Journal*, no 57, pp. 49–68 (in Russian).

38. Novikova E.V., Mashkova A.L. (2018) Creation of the initial generation of agents in the computer model of industrial development of the Russian economy. Proceedings of the *VII International Scientific and Technical Conference "Information Technologies in Science, Education and Manufacturing", Belgorod, Russia, 17–19 October 2018*, pp. 313–318 (in Russian).

39. Molodetskaya E.Yu. (2017) Simulation of supply chain systems at the macroeconomic level as a tool for managing the economic development of the region. *Economics: Yesterday, Today and Tomorrow*, vol. 7, no 9A, pp. 180–191 (in Russian).

40. Lychkina N., Molodetskaya E., Morozova Yu. (2017) The simulation model of supply chains on the macroeconomic level is the tool to control the economic development of the region. *Strategic Innovative Marketing*. Springer, pp. 357–362. DOI: 10.1007/978-3-319-56288-9_47.

41. Orekhova S.V., Kislitsyn E.V. (2019) Total factor productivity in the Russian industry: Small vs large enterprises. *Journal of New Economy*, vol. 20, no 2, pp. 127–144. DOI: 10.29141/2073-1019-2019-20-2-8.

About the authors

Evgeniy V. Kislitsyn
Cand. Sci. (Econ.);
Head of Information Technology and Statistics Department, Ural State University of Economics, 62, 8 Marta Street, Yekaterinburg 620144, Russia;
E-mail: kev@usue.ru
ORCID: 0000-0003-1518-0043

Victor V. Gorodnichev
Senior Lecturer, Information Technology and Statistics Department, Ural State University of Economics, 62, 8 Marta Street, Yekaterinburg 620144, Russia;
E-mail: helltoaster@yandex.ru

DOI: 10.17323/2587-814X.2021.1.78.96

Big data analysis of IoT-based supply chain management considering FMCG industries

Hamed Nozari [a]
E-mail: Ham.nozari.eng@iauctb.ac.ir

Mohammad Fallah [a]
E-mail: Mohammad.fallah43@yahoo.com

Hamed Kazemipoor [a]
E-mail: Hkazemipoor@yahoo.com

Seyed Esmaeil Najafi [b]
E-mail: E.najafi@srbiau.ac.ir

[a] Islamic Azad University, Central Tehran Branch
Address: Hamila Blvd., Poonak Sqr., Tehran 1469669191, Iran

[b] Islamic Azad University, Tehran Branch, Science and Research
Address: Daneshgah Blvd., Simon Bulivar Blvd., Tehran 1477893855, Iran

Abstract

Supply chain is one of the main pillars of manufacturing and industrial companies whose smartness can help business to be intelligent. To this end, the use of innovative technologies to make it smart is always a concern. The smart supply chain utilizes innovative tools to enhance quality, improve performance and facilitate the decision-making process. Internet of things (IoT) is one of the key components of the IT infrastructure for the development of smart supply chains that have high potential for creating sustainability in systems. Furthermore, IoT is one of the most important sources of big data generation. Big data and strategies for data analysis as a deep and powerful solution for optimizing decisions and increasing productivity are growing rapidly. For this reason, this paper attempts to examine informative supply chain development strategies by investigating the supply chain in FMCG industries as a special case and to provide a complete analytical framework for building a sustainable smart supply chain using IoT-based big data analytics. The proposed framework is based on the IoT implementation methodology, with emphasis on the use of input big data and expert reviews. Given the nature of the FMCG industry, this can lead to better production decisions.

Key words: big data; internet of things (IoT); IoT-based supply chain management; FMCG supply chain.

Citation: Nozari H., Fallah M., Kazemipoor H., Najafi S.E. (2021) Big data analysis of IoT-based supply chain management considering FMCG industries. *Business Informatics*, vol. 15, no 1, pp. 78–96.
DOI: 10.17323/2587-814X.2021.1.78.96

Introduction

An intelligent supply chain is an innovative supply chain that utilizes information technology and other technology tools to improve efficiency, improve processes and increase service levels. In today's world, the modern business environment is dealing with data and this has created many challenges and opportunities. The volume of date produced in various sectors is enormous, and their analysis requires specific capabilities and technologies. These technologies include information technology, robotics, internet technologies, commercial automation, Augmented reality (AR) and Virtual Reality (VR) technologies among others. These developments are known as the digital economy or industry 4.0. This provides a great deal of demand for supply chain management by increasing consumers' expectation of service level and delivery time [1]. The advent of IoT and ICT has changed many of the concepts so that the "smart supply chain" can be one of them [2]. Therefore, these supply chains (which actually make business smarter) use information technology to be intelligent and efficient in using resources to maximize the quality of businesses. IoT is one of the key components of the information technology infrastructure in smart supply chain management due to its high potential for increasing the sustainability of the business environment [3]. IoT is associated with big data analytics, which is clearly leveraging many areas of the business to optimize energy efficiency and reduce effects damaging to the environment [4]. Data-driven applications and IoT have a huge impact on facilitating and improving the sustainable development process of the environment [5]. In general, the development of IoT, as a computational paradigm and analytical process of big data, promotes sustainable smart city initiatives and programs in the environmental and technological fields of developed countries [6]. In the context of a sustainable smart supply chain, the volume of data produced is beyond the imagination that is produced using different technologies. IoT technologies include a variety of sensors, data processing systems, wireless communication networks, and system activators in the physical environment [5]. Despite increasing research on IoT and big data related applications, most research on IoT has focused more on urban development, and the applications of these technologies in businesses are often less considered. Therefore, the main question of the present study is how and to what extent can the information perspective of a sustainable smart supply chain be enhanced by the use of IoT and big data processing?

Of course, in recent years, with the development of the concept of IoT and its close relationship with data analytics, there has been a great deal of research in this field and this research is growing. To this end, some researchers have tried to show that using IoT data can improve port-based intermodal supply chain performance [7]. Other research also addressed the role and impact of using IoT and blockchain in supply chain management in industries such as agriculture. Research has shown that the use of these technologies can fundamentally change the economics of these industries [8]. Although the challenges

and opportunities for using IoT-derived big data in the manufacturing industry have also been addressed in research [9], but so far, there has been no framework for using the big data coming from connected devices. At the moment, there is a lack of innovative solutions based on big data and IoT. Therefore, in this study, we try to explore supply chain informative improvement solutions using IoT. The supply chain of the FMCG (fast-moving consumer goods) industry was considered as a special case in this research. These industries are of great importance due to the nature of production as well as the distribution of products. So, in addition to reviewing the literature and previous research, the opinions of experts active in FMCG industries were also used for this research. In this regard, this paper presents an analytical framework and describes the ways of generating big data in the field of sustainable intelligent supply chain (in the FMCG supply chain). This framework is based on the 4-step process of IoT implementation in intelligent business. This framework illustrates how the use of Big Data Input based on IoT devices can be used to make decisions in the supply chain in the FMCG industries. This framework illustrates the direct relationship between data entry (derived from IoT) and final decision-making in the FMCG supply chain network. This work provides a basis for supply chain researchers to develop analytical frameworks for future research. The framework introduced here can be developed, tested and evaluated in empirical research and will lead to deeper studies of the smart supply chain.

The rest of the paper is organized as follows. Section 1 presents a review of the literature in terms of big data and the internet of things. Section 2 presents smart businesses. The smart supply chain is illustrated in Section 3 considering supply chain and IoT and big data in the FMCG supply chain. In Section 5 a sustainable smart supply chain framework is provided, and lastly, the conclusions are presented.

1. Literature review

In order to illustrate the effects of big data and IoT concepts on supply chain management, a literature review of big data and IoT technologies is provided in this section.

1.1. Big data

The term 'big data' was first proposed by Cox and Ellsworth in 1997 as an interesting challenge for computer systems, when data sets do not fit in main memory or when they do not fit even on local disk or remote disk [10]. In recent decades, 'big data' concept refers to greatly increased amounts of data that are constantly generated in various fields. Generally, big data could be defined as the datasets that could not be captured, stored, managed and analyzed by IT systems of an organization within a particular time frame [11]. It is assumed that as technology advances over time, the size of datasets that qualify as big data will also increase and it will vary by sectors, in many of which big data can range from a few dozen terabytes to multiple petabytes [12]. Some scholars use the notation of 'V' for some characteristics to describe 'big data'. Some of them [13–16] defined big data in terms of 3 Vs, Velocity, Variety and Volume, in which 'velocity' refers to the speed of data generation and/or frequency of data delivery, 'variety' represents a large variety of sources and formats from which data are generated, and 'volume' denotes the large amount of data [14]. In some definitions of big data, another V has been added (known by 4 Vs) as Value referring to the significance of extracting economic benefits from the big data [17, 18]. Furthermore, White [19] proposed an additional characteristic as Veracity, in order to stress the importance of sufficient quality of the data and the level of trust in different data sources. In addition, three main characteristics of 'big data' were identified by IDC [20] as the data itself, its analytics and the presentation of the results obtained by analytics. Boyd

and Crawford [21] suggested a more holistic description of 'big data' that includes a cultural, technological, and scholarly phenomenon that rests on the interplay of technology, analysis, and mythology. In this definition, the technology aspect refers to storage and computation power to process and analyze datasets, the analysis aspect is related to patterns identification for economic, social, technical, and legal claims or the type of analysis implemented on datasets, and the mythology aspect includes the widespread belief that the big data offers a higher form of intelligence. Therefore, 'big data' can be defined as an approach to manage and analyze the V's characteristics to establish competitive advantages as well as creating sustained value delivery and measuring performance [18]. Typically, the 5 Vs model or its derivations is known as one of the most common definitions of the 'big data' [9].

Big data analytic (BDA) is defined as one of the key foundation technologies including analytics research, alongside Text Analytics, Web Analytics, Network Analytics and Mobile Analytics, and is applied to describe data mining and statistical analysis using business intelligence and analytics technologies [22]. Beyond data analytics decisions in customer marketing and customer research, Big Data Analytics (BDA) has increasingly changed the business value propositions of product businesses and services by increasing the efficiency of physical products and providing personalized services [23]. In this case, security is certainly one of the most important issues in the production and use of critical big data. But research shows that data security concerns are not key factors in the use of big data analytics [24]. For this reason, a growing number of firms are accelerating the deployment of big data analytics plans aimed at developing critical insights that can ultimately provide them with competitive advantage [25].

Some scholars described BDA as the "fourth paradigm of science" [26] or even "the next frontier for innovation, competition, and productivity" [12]. In fact, BDA can enable data miners and researchers to analyze a large volume of data that may not be tackled by applying traditional tools [27]. BDA can be employed in various tools such as social media, portable devices like laptops and smartphones, automatic identification systems enabling the IoT and cloud-enabled platforms in order to support all organizational business processes [28].

The extensive applications of IoT have made BDA challenging due to the processing and collecting of data from different sensors in the IoT environment. So, in the IoT big data analytics perspective, a variety of IoT data are examined to reveal trends, unseen patterns, hidden correlations, and new information [27, 29]. As recognized by statistics, the number of sensors will be increased by 1 trillion in 2030 [30] that can be led to the growth of big data and subsequently huge resources will be required. To effectively communicate among various deployed applications, IoT services can provide appropriate resources and intensive applications of the platforms. It is found that the integration of IoT and big data can help address issues on storage, processing, data analytics, and visualization tools [31]. In accordance with the requirements of IoT applications, different analytic types are used including real-time, off-line, memory-level, business intelligence, and massive level analytics categories [32]. Real-time analytics is usually implemented on data collected from sensors, according to parallel processing clusters using traditional relational databases and memory-based computing [31, 33]. If there is no need for a quick response, off-line analytics can be applied and when the size of data is smaller than the cluster's memory, memory-level analytics is employed [32]. Business intelligence (BI) analytics is utilized when the size of data is larger than the memory, so data can be imported to the BI analysis environment [34]. Additionally, if the size of data is much

greater than the whole capacity of the BI analysis product in addition to the traditional databases, massive analytics is used [35]. Big data analytics together with the IoT concept are usually employed to improve decision making. So, increasing the amount of data in IoT applications can lead to development of big data analytics. Moreover, employing big data technologies in IoT can facilitate future research advances and business models of IoT [31].

The IoT-based data in a supply chain are definitely considered as big data, satisfying the sufficient conditions in terms of the V's characteristics. In general, IoT data is constantly generated in real time within the supply chain processes in addition to providing a variety of data formats [36]. Moreover, when a large number of tags and sensors are connecting through the internet, an unprecedented number of transactions and amounts of data are generated [37, 38]. In this regard, BDA by representing a critical source of meaningful information can help supply chain stockholders to improve their insights for competitive advantage [39] in addition to reducing their exposure to various risks [40]. Furthermore, it was reported that BDA can lead to increasing the efficiency and profitability of supply chains by maximizing speed and visibility, improving supply chain stakeholders' relationships, and enhancing supply chain agility [41]. In addition, BDA results in faster time to market and the potential for superior revenue recognition [28]. It is found from a survey within 720 firms that, although 64% of respondents were planning to invest in BDA projects, less than 8% of them had actually deployed a solution [42].

Recent research in recent years has addressed the issue of big data and its impact on the supply chain from multiple angles. Some research has specifically provided frameworks for evaluating supply chain performance using big data [43] and some research has addressed the effects of agility using big data [44]. But in operational terms, most academic papers in this field focus on describing the term and its key factors as well as making predictions about the level of impact it will have in certain sectors in the future [45–47]. There are a limited but growing number of publications concentrated on industrial cases, education, sports, public sector, mining and logistic [48–52]. However, there is still a lack of case studies of big data analysis in IoT supply chain management especially in FMCG industries.

1.2. Internet of things

The term 'internet of things' (IoT) was coined by Kevin Ashton in the late 1990s in order to examine work concerning Radio-Frequency Identification (RFID) infrastructure [53, 54]. Subsequently, the new concept of sensors and actuators through a wireless sensor network came to sense and monitor objects [55]. In recent decades, the modern concept of IoT includes GPS devices, smartphones, social networks, cloud computing and data analytics [56]. In general, not only IoT tracks the objects and collects the new data but also combine them to generate a greater level of information [54]. The IoT concept has developed to describe a global network of infrastructure where things, wireless networks, and computing abilities combine to form a network of information [57]. So that, the recent adoption of IoT has witnessed smart cities with regard to developing intelligent systems, such as smart energy, smart office, smart agriculture, smart water, smart retail, smart transportation, smart healthcare, etc. [58]. The rapidly growing number of interlinked "things" in the so called industrial internet of things (IIoT) offers various options for connecting the physical and digital spheres together [59]. In Europe, IoT is one of the founding technologies of Industry 4.0 as the integration of information and communications technologies with industrial technology [60]. The 'things' in IIoT can include smart machines, smart products, and smart services. IoT can be applied in an incredibly

wide range of application scenarios [61]. Some instances of physical items that IoT connects to the digital worlds include actuators, sensors, electronic toll devices in vehicles, washing machines, lighting systems, front door locks, thermometers, air conditioning units and much more [54, 57, 62–64].

Recent data IoT protocols are specifically designed for IoT devices that use low-power wide-area networks to connect at a low bit rate a wide variety of devices with low cost and minimum energy consumption [65]. Lee and Lee [66] presented that IoT technologies, as essential parts in the deployment of successful IoT-based products and services, are classified into five categories as radio-frequency identification, wireless sensor networks, middleware, cloud computing, IoT applications. On the other hand, it is recognized that IoT are faced with several challenges such as scalability, self-Organizing, data volumes, data interpretation, interoperability, automatic discovery, software complexity, security and privacy, fault tolerance, power supply, and wireless communications [67, 68]. However, it is important for new designs to support the development of smart and connected products in order to expand opportunities for new functionalities and increase capabilities that cut across and transcend traditional product boundaries [69]. In *Figure 1*, the architecture of the internet of things is shown. As can be seen in the figure, this architecture has four stages. In summary, these four steps include data acquisition, data analysis, data processing and computation, and ultimately the use of data to improve performance. This model can be the basis for presenting industrial frameworks in different industries or different organizational units.

In light of the future of IoT that can enable everyone to access and contribute rich information about things and locations, several developments towards the IoT have concentrated on the combination of Auto-ID and networked infrastructures in logistics and product life cycle applications [70]. Sun [71] claimed that enterprises with IoT can monitor their every product in real time. Therefore, due to analyzing the information generated from every procedure as well as more accurate forecasting, the abilities of enterprises for responding to the market can be improved [71]. Wortmann and Fluchter [64] discussed that existing business models may have to be adapted or redefined based on a new positioning of products in the IoT and even entire industry boundaries may need to be re-assessed in today's competitive marketplace. Moreover, it was argued that the IoT promises future new technologies when related to big data, cloud and distributed com-

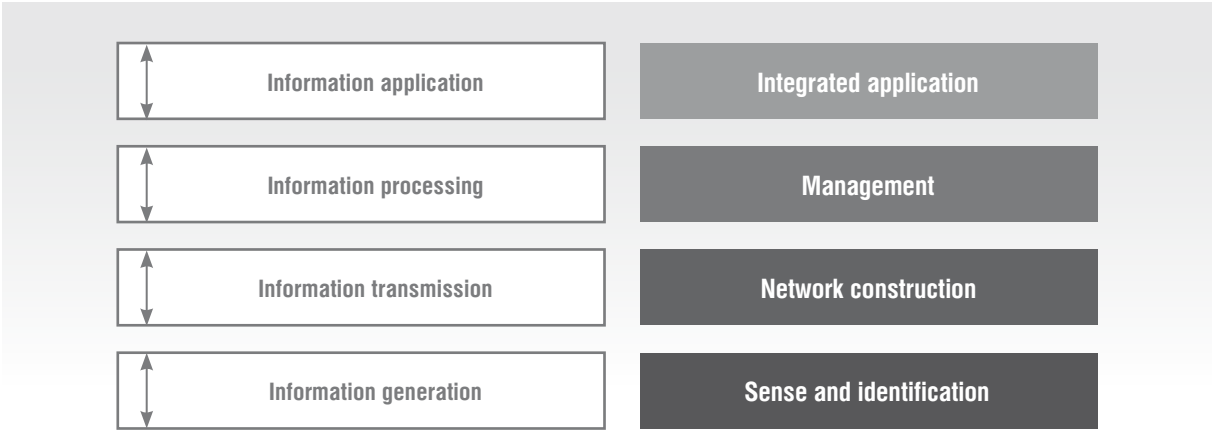


Fig. 1. Four–layer architecture of internet of things

puting in order to develop smarter applications as soon [72]. It can be seen that this concept is expanding day by day and provides high value for all activities in life and business. In recent years this concept has been more broadly referred to as the internet of everything (IoE). The IoE connects people, organizations, and smart things, and also promises to fundamentally change the way we live, work and interact, and may redefine a wide range of industry segments [73].

The application of IoT in supply chain management consists of several areas such as cost-saving, inventory accuracy and product tracking [56]. In order to monitor the good, it should be tracked in both indoors by developing RFID system, and outdoors by mainly using GPS [74]. Tao et al. [75] designed an IoT-based framework to achieve intelligent perception and access of various manufacturing resources. Gnimpieba et al. [76] proposed the architecture of a platform based on advanced technologies related to IoT for a better collaboration and interoperability enhancement in supply chain. Verdouw et al. [77] provided a reference architecture for IoT-based supply chain information systems in the agricultural industry. Decker et al. [78] described several benefits of IoT connected to sourcing and examined the impact of the cost of sensors and alerts on the unit purchase cost. It was found that smart manufacturing can lead to smarter decisions and more efficient operations in the factory and supply chain visibility based on real-time information [56]. Xu [79] believed that successful supply chain quality management depends on more sophisticated systems. Chen [80] proposed the intelligent IoT-enabled system in green supply chain to simulate complex system by linked physical and digital objects with relationships. Yan et al. [81] illustrated an intelligent supply chain integration and management system to provide flexible and agile approaches to facilitate the resource sharing and participant collaboration in the whole life cycle of supply chain. Parry et al. [82] demonstrated how the IoT may be operationalized in the domestic setting to capture data on a consumer's use of products and the implications for reverse supply chains. Kang et al. [36] proposed a sensor-integrated RFID data repository-implementation model that can integrate and store a large amount of IoT-generated data collected from supply chain processes. Thoben et al. [83] indicated that smart manufacturing can include different technologies such as IoT, robotics/automation, big data analytics and Cloud computing.

Therefore, due to lack of academic publications that address big data analysis of a smart supply chain in an IoT environment, it is important to investigate the applications and interactions of the concepts of BDA and IoT in supply chain management.

2. Smart businesses

Organizations are pursuing new technologies in their business for a variety of reasons, including facilitating activities, reducing human errors, decreasing costs and speeding up services and product deliveries [2]. One of the latest computer-related technologies, also referred to as information technology developments, is IoT, which creates value for both businesses and customers [84]. IoT enables the fabrication of digital and physical structures and provides a whole new level of complete applications and services that must be used with regard to environmental sustainability. In the context of sustainable smart businesses, increasing the volume of data production is beyond imagination and the sheer volume of current available information from various business areas are so valuable to be used by IT planners and IT professionals in order to promote environmental sustainability. This phenomenon has created a huge revolution in industry [85]. IoT is one of the major sources of big data generation. As mentioned

in the previous section, IoT technologies include a variety of sensors, data processing systems, wireless communication networks, and system activators in the physical environment [5]. Moreover, IoT can improve the performance of value chain management and smart commerce by implementing functional technologies and transforming product and other artifacts into smart items in all parts of the value chain (suppliers, manufacturers and retailers), which in turn leads to flexibility, high reliability, and proper information monitoring throughout the supply chain [86]. So a smart business can be introduced as a large, organic system that integrates many subsystems and components. In other words, business intelligence combines the enhancing and effective digital telecommunications networks (nerve), embedded information (brain), sensors and labels (sensory organs) and software (knowledge and cognitive competence). *Figure 2* shows an example of an IoT-based smart business [2]. As can be seen in *Figure 2*, the relationship between the market, suppliers as well as manufacturers is based on local as well as wireless communication in smart business. In this case, the integration and refinement and optimization of data from these sources ensure system intelligence. A smart business can have many benefits, some of which are [87–90]:

- ✦ decreasing the use of vital resources and water;
- ✦ reduction of the air pollution;
- ✦ proper use of the existing fundamental constraints and then increasing personal satisfaction and reducing the need for conventional venture capitalists;
- ✦ how to deal with many abuses;
- ✦ increasing business investment by disseminating continuous information about activity in different geographic areas.

The IoT-based supply chain is then discussed as one of the most important pillars of smart businesses.

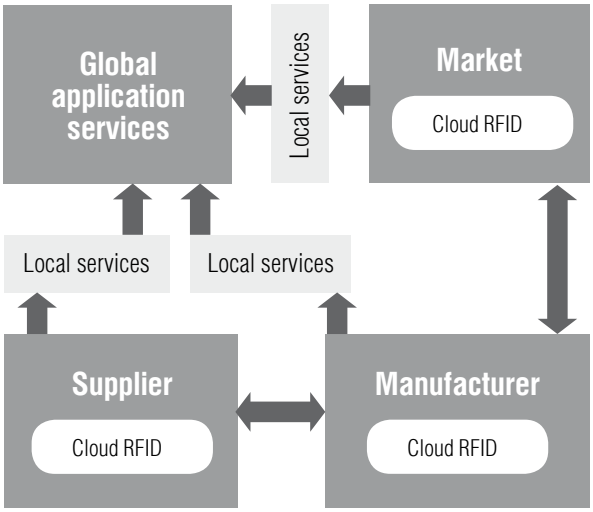


Fig. 2. Smart business based on IoT

3. Smart supply chain

3.1. Supply chain and IoT

Supply chain management integrates supply and demand processes throughout the organization. Supply chain management claims to offer solutions to help industry professionals better manage the entire supply chain from suppliers to end customers. Supply chain systems have been empowered using the internet and this has led to increased efficiency and productivity in the supply chain. In traditional supply chain management systems, there exist several problems such as overstocking, delivery delays and stock out. These problems return to several factors such as complexity and uncertainty as exist in real supply chains [91]. IoT technology can help the industrial system to manage these uncertainties and improve supply chain efficiency [92]. The benefits of managing an electronic supply chain can include operational and strategic improvement in communications, collaboration and coordination to meet organizational barriers [93]. Thus, IoT smart devices enable supply chain companies to reduce the costs of the knowledge acquisition process. Using IoT in supply chain man-

agement makes it smarter and has the following features [91]:

- ✦ the entire system must connect through the use of intelligent objects and IT systems;
- ✦ their performance is optimized by large-scale decision making;
- ✦ all processes must be automated and therefore less efficient resources will be lost;
- ✦ all stages of the supply chain are integrated;
- ✦ new values evolve through solutions to meet new needs.

Firms will put resources into the IoT to build deceivability of materials stream as well as decreasing the loss of materials, and lower circulation costs. This is delineated in *Figure 3*.

In the current paper, we selected the FMCG (fast-moving consumer goods) supply chain as a special case. So in the following, the term 'supply chain' refers to the supply chain of FMCG industries.

Due to the nature of products in the FMCG industry, supply chain management has a special place in this industry. Usually the products in these industries can be perishable and therefore the distribution and delivery system

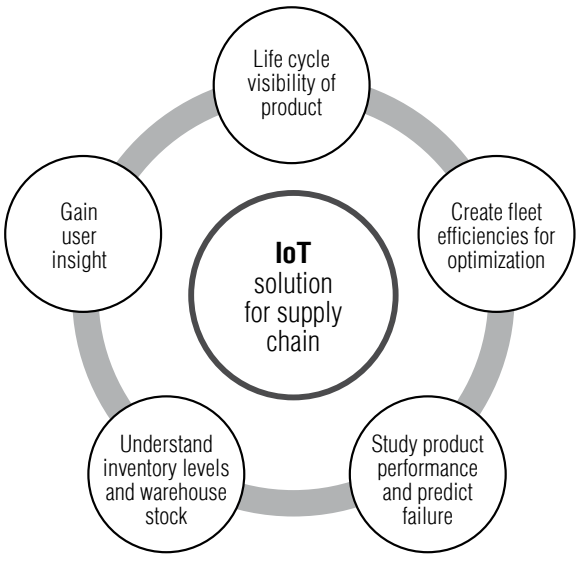


Fig. 3. The applications of IoT in the supply chain [94]

in a given time is of great importance. Freight tracking links and interconnections with suppliers and end users are other features of the supply chain in these industries. As the FMCG products meet many people's daily needs, so their demands are high, and meeting the demands at the right time is another important component of the supply chain in these industries. Therefore, IoT, as one of the most important solutions for the production, maintenance and tracking of data, can have a huge impact on the supply chain of these companies and can have many effects. A simple diagram of FMCG supply chain management is presented in *Figure 4*. Supply chain improvement is essential for FMCG companies. Supply chain innovation is defined as the change in a supply chain network, supply chain technology or supply chain process or a combination of these that can be incorporated into the functions of a company, an entire company, industry or supply chain to enhance new value for stakeholders [95]. Network information of warehouse, production, distribution and other communications are gradually generated by the system of sensors, RFID tags, meters, actuators, GPS and various frameworks. Companies will put resources into the IoT to build deceivability of materials stream in addition to decreasing the loss of materials, and lower circulation costs. As IoT comes to the core business forms, a large number of corporate resources are converted into green store networks, and the use of IoT will become a matter of great excitement for bosses [96]. But security risks analysis on the IoT should always be considered.

One of the most important things about the IoT is that it is one of the largest data producers for an intelligent system created from internet-related tools. So gadgets are very helpful in using them. Although gadgets and systems are physically available, IoT applications reinforce gadget communication together as well as gadget and human communication in a robust way. IoT applications in the context of gadg-

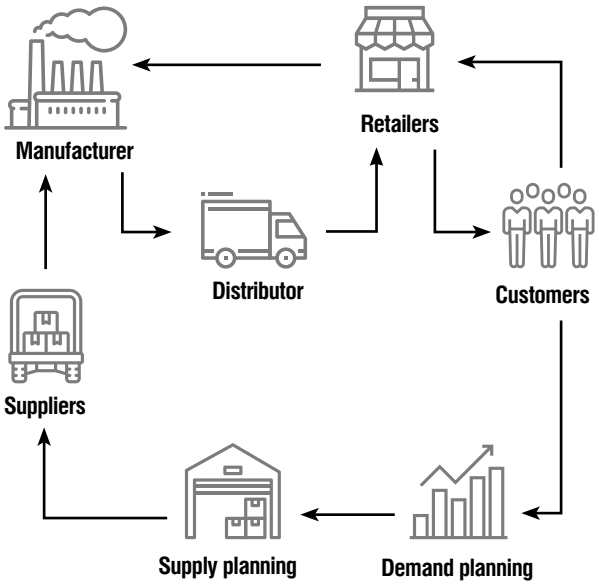


Fig. 4. FMCG supply chain

ets should ensure that information is accessed and tracked appropriately. For example, transportation and coordination programs that show the status of goods shipped (for example organic products, meat and dairy or even medicine). For proper transportation, the status of raw materials as well as products such as temperature, humidity, etc. will be checked and appropriate solutions to reduce the amount of waste as well as reducing air pollution are also adopted. IoT devices can enhance supply chain capability in terms of sustainability as well as greenness (such as green transportation and cold and green storage) [97] and there are, of course, many obstacles in this way for the supply chain [98]. Innovative developments, communications and IoT capabilities are on the rise and are increasingly open and flexible, so this makes the use of IoT be a welcome addition. IoT supply chain gadgets generate huge amounts of data that need to be collected and phased in to provide information about the status, area, usefulness and status of items and management. Therefore, intelligent information regulation enables powerful sensors and gadgets to use specific data and can be used to collect and decode data. These data are very big and therefore have their own way of analysis. In the following, big data is described as one of the capabilities and outputs of IoT.

3.2. Big data in the FMCG supply chain

Big data generation in industrial IoT (IIoT) is evident due to the widespread deployment of IoT sensors and devices. However, processing big data is challenging due to the limited computing, networking and storage resources at the end of the IoT device [99]. The FMCG industry supply chain is a changing organizational system that encompasses individuals, information, activities and resources in the manufacturing and production, processing and distribution of products from raw material supplier to consumer [100]. Big data is used in the production, packaging, sales and consumption of products. *Figure 5* shows the big data rotation network in the FMCG supply chain. In the supply chain (as presented by *Figure 4*) manufacturers track large volumes of useful data to gain a better understanding of customers' desired values, then process that data and enter them in new product design cycles to optimize production. Retailers can use this big data to categorize and differentiate customer types and use this information for accurate sales plan, timely marketing, and customer loyalty development. By the generated data, customers can also make more precise decisions to buy products. It should be noted that every year hundreds of billions value are generated by big data in the FMCG supply chain and that illustrates the importance of big data in this industry.

Logistical data is provided by distribution activities and people involved in processes as well as government information. But one of the most important benefits of using big data for the FMCG industry is that it can be used to enhance the ability of distributors and the timely delivery of goods and products to cus-

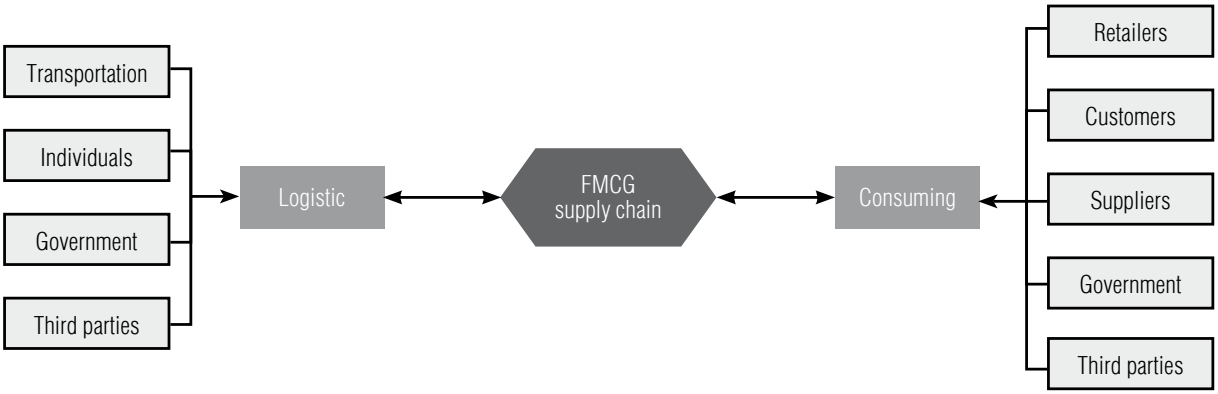


Fig. 5. Big data in FMCG network

tomers. In addition, companies can optimize many of their activities using this data and consequently they gain a better understanding of the decision-making in the organization and make decisions more powerfully and transparently. In general, the following can be cited as the reasons for the importance of using and utilizing big data in the organization's supply chain [101]:

✦ Performance optimization and improvement: If big data extracted from different parts of a supply chain is properly processed and analyzed, it can be one of the most important components of performance improvement in organizations.

✦ Transparency: Continuously receiving data from different parts of the supply chain can increase the transparency of the system and thus can reduce the possibility of errors in different parts.

✦ Increasing decision-making power: Using powerful algorithms to refine big data in the supply chain enables better understanding of the situation and can also increase decision-making power in the organization. By analyzing customer behavior using big data we can offer different marketing strategies for delivering products. In addition, business model structure can be changed to gain more market share.

4. Sustainable smart supply chain framework

In this section, we will develop a sustainable intelligent supply chain framework based on IoT and big data analytics. By studying the literature, it can be seen that frameworks for building smart cities based on big data and even internet of things have been provided [102], but the specific framework for the smart supply chain as one of the most important organizational units or in a particular community has not yet been presented. The framework presented is based on findings from the literature review and thematic analysis along with conceptual and theoretical backgrounds. In this framework, a layered approach is used which, according to the scientific literature, is usually applied in system architectures and infrastructures in the field of intelligent systems. The purpose of this framework is to develop a smart supply chain information landscape with IoT applications and big data processing. In this context, big data Analytics is aimed at optimizing and supporting smart decisions related to control, optimization, automation, management and planning of supply chain systems as business operating processes. Supply chain systems must be managed using IoT and big data analytics (as a set of advanced technologies, and new applications). Supply chain sys-

tems provide large amounts of data as inputs for applications. The data is received at different sensors at different scales and speeds and is automatically stored in the data warehouse for large scale use. Therefore, these components include various sources of data in different types and sizes that must be stored, processed, analyzed, and shared in different operations, functions, and states. Big data applications include a variety of program enabled by IoT in relation to environmental sustainability in different parts of the supply chain. A program usually contains several solutions to the different subdivisions of each domain that depend on the type of environmental sustainability issue that each domain faces. *Figure 6* shows the deployment of big data processing using IoT technologies implemented on the cloud in the field of sustainable supply chain systems. These technologies include sensors, data warehouses, data processing platform and cloud computing model. Sensor data for various parts of the supply chain (in FMCG industries), which have been collected, integrated and pre-processed, are used for automation, support and decision making in operations using data mining and machine learning techniques for modeling, pattern recognition and correlation creation.

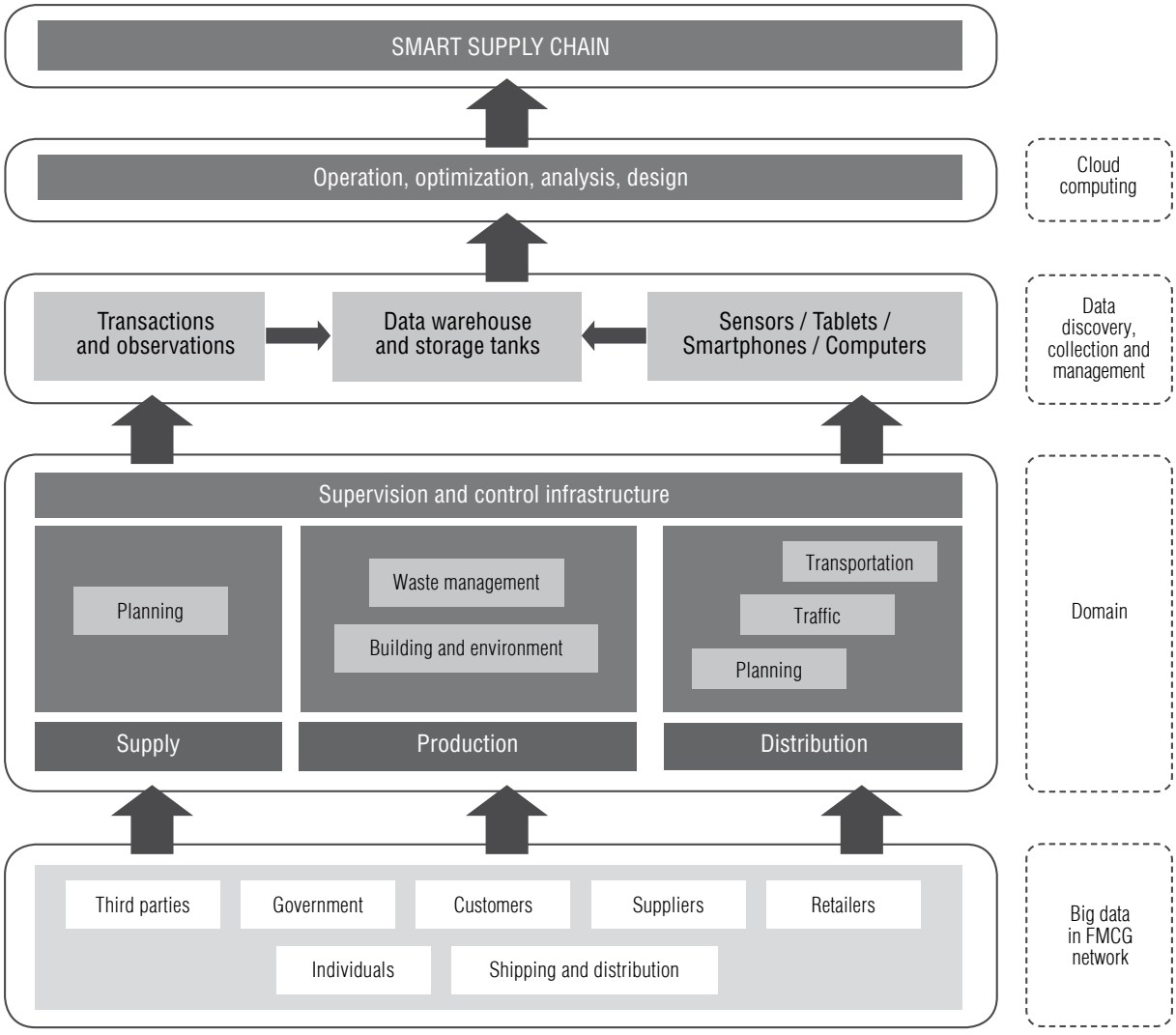


Fig. 6. Smart supply chain framework based on IoT and big data analytics

This framework includes comprehensive computing, data processing systems and wireless network infrastructure, cloud computing, data processing, data analysis, and data modeling to discover useful knowledge. This helps supply chain management by utilizing the big data related to monitoring, control, automation, optimization and management of infrastructure, resources, facilities, services and networks. Some of the concepts used in the framework are described as follows.

Collecting, storing and processing of big data: Sensors are one of the key features of sustainable smart systems that rely on common approaches to computing including the IoT. Sensor data will be available in a variety of formats, including time and space labels, along with a variety of data mining techniques and data visualization techniques for data processing and displaying interconnected patterns. But the end goal is to create a comprehensive system that supports data extraction, data integration, data processing, data evaluation, pattern exploration, pattern simulation and modeling, and deploying the results from the processes.

Cloud computing for big data processing: Cloud computing is a new processing method in which scalable and often virtualized resources are delivered as a processing service through communication networks such as local area networks and the internet. The focus of this model is on-demand service to the user, without the user having specific equipment to process or being aware of the place of processing. Cloud computing offers many solutions to the problems of sustainable smart systems by facilitating big data storage and providing the capabilities to process, manage and extract useful knowledge from them.

Intelligent logistics and transportation systems: Dynamic transportation is one of the key applications of IoT related to environmental sustainability. IoT plays an important role in improving the dynamics of all types of transportation in sustainable smart systems. Use of IoT includes automated tracking of distribution systems and vehicles, monitoring of road and traffic conditions for timely delivery and integrated safety mechanisms, and monitoring when distributing valuable goods. In general, the uses of IoT, as a form of pervasive computing are strongly affected by the dynamics of transportation in order to reduce energy consumption, reduce pollution, and eliminate inefficiencies.

Smart environment: IoT employs sensors to monitor and control mechanical and electronic systems used in industrial buildings.

Infrastructure monitoring: IoT devices can be used to improve disaster management, improve emergency response coordination, improve service quality, and reduce operating costs in all infrastructure-related areas. As an ancillary product, IoT infrastructure enables efficient maintenance of planning activities by coordinating tasks between service providers and users of these infrastructures.

With the growing demand for IoT as well as the analysis of big data in smart systems, the use of these technologies also faces challenges. Some of these challenges include:

- ✦ providing security for information;
- ✦ lack of understanding of many features and information;
- ✦ technical challenges;
- ✦ the cost of many technologies;
- ✦ quality assurance and data access;

So in addition to the positive and undeniable effects of using IoT and big data analytics, these challenges must also be addressed.

After the design, the framework was provided and approved by 24 supply chain specialists in FMCG companies. This study was tested using a five-point Likert scale questionnaire and the results showed the acceptability of this framework. It should be noted, however, that the selection of these experts was also challenging. Because these professionals had to be selected among those who had a thorough knowledge of IoT technologies and had a track record of using some of these technologies in their background.

Conclusion

IoT is a new form of pervasive computing and applications for big data that is increasingly being taken into account with operational performance and planning for sustainable development. Therefore, the use of IoT in the supply chain can greatly enhance its intelligence and thus improve its performance. On the other hand, analyzing big data and using it to realize key features of a smart and sustainable supply chain (such as operation and service efficiency, resource optimization, and intelligent infrastructure and facility management) has a huge impact. For this reason, the purpose of this paper was to review and integrate the related literature in order to identify and discuss IoT-based big data applications for supply chain sustainability. In this paper, the feasibility of developing smart supply chains using big data from IoT is explored to achieve the level of required intelligence. In this paper, the supply chain of FMCG companies was selected. Products in these industries have a special nature and as a result their supply chain has different characteristics. Timely distribution and delivery system is one of the main features of this supply chain. The most important data-driven applications in these supply chains that are enabled through IoT include transportation, dynamics, energy, environment, infrastructure monitoring and management and supply chain planning. Therefore, expanding the intelligence landscape of the smart supply chain using sensor-based big data has great potential to promote environmental sustainability. To this end, a framework has been outlined that provides a lot of information on the smart supply chain.

This framework defines the scope of decision-making in the core parts of the supply chain in the FMCG industry according to their nature. With this framework, the ways in which big data can be accessed through the internet of things (which can itself be a path to sustainability and in some cases being green) are identified in this industry. By placing appropriate analytical tools in the computational parts of the framework (such as statistical analysis and optimization), one can gain a good understanding of the audience's beliefs. Finally, analyzing this data can help to make the organization more intelligent. One of the most important advantages of using this framework can be the production of optimal products based on customer preferences.

The implementation of the framework will always have challenges. One of the problems is always the use of huge volumes of data and therefore the high complexity of analysis. Since data is produced using the IoT technology without any restrictions, maintenance and computation on this data is very time consuming and complex and therefore there is a need for high knowledge in this field. In addition, IoT devices are not always available and can be costly to obtain. Therefore, all of these must always be considered when implementing the framework.

This framework provides the basis for developing research opportunities to integrate this type of supply chain as well as providing analytical insights into future research. This framework can be expanded and operational solutions in analytics and computing can be customized to suit the different industries. Also, providing computational solutions and incorporating them precisely into the framework can provide a clear pathway for solving new challenges. ■

References

1. Evtodieva T.E., Chernova D.V., Ivanova N.V., Wirth J. (2020) The internet of things: Possibilities of application in intelligent supply chain management. *Digital transformation of the economy: Challenges, trends and new opportunities.* Series: Advances in Intelligent Systems and Computing, vol. 908 (S. Ashmarina, A. Mesquita, M. Vochozka, eds.). Cham: Springer, pp. 395–403. DOI: 10.1007/978-3-030-11367-4_38.
2. Tajfar A.H., Gheysari M. (2016) Analysis the effects of internet of things technology in managing supply chain. *International Journal of Information & Communication Technology Research*, vol. 8, no 3, pp. 15–25.

3. Grossmann I.E. (2018) Optimization and management in manufacturing engineering: resource collaborative optimization and management through the internet of things. *Optimization Methods and Software*, vol. 34, no 1, pp. 220–223. DOI: 10.1080/10556788.2018.1527332.
4. Wazid M., Das A.K., Hussain R., Succi G., Rodrigues J.J.P.C. (2019) Authentication in cloud-driven IoT-based big data environment: Survey and outlook. *Journal of Systems Architecture*, vol. 97, pp. 185–196. DOI: 10.1016/j.sysarc.2018.12.005.
5. Bibri S.E. (2015) *The shaping of ambient intelligence and the internet of things: Historicoepistemic, socio-cultural, politico-institutional and eco-environmental dimensions.* Atlantis Press. DOI: 10.2991/978-94-6239-142-0.
6. Bibri S.E., Krogstie J. (2016) On the social shaping dimensions of smart sustainable cities: A study in science, technology, and society. *Sustainable Cities and Society*, vol. 29, p. 219–246. DOI: 10.1016/j.scs.2016.11.004.
7. Muñuzuri J., Onieva L., Cortés P., Guadix J. (2020) Using IoT data and applications to improve port-based intermodal supply chains. *Computers & Industrial Engineering*, vol. 139, article no 105668. DOI: 10.1016/J.CIE.2019.01.042.
8. Borah M.D., Naik V.B., Patgiri R., Bhargav A., Phukan B., Basani S.G.M. (2019) Supply chain management in agriculture using blockchain and IoT. A*dvanced applications of blockchain technology*, Studies in big data, vol. 60 (S. Kim, G. Deka, eds). Singapore: Springer, pp. 227–241. DOI: 10.1007/978-981-13-8775-3_11.
9. Dai H.-N., Wang H., Hu G., Wan J., Imran M. (2020) Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies. *Enterprise Information Systems*, vol. 14, no 9–10, pp. 1279–1303. DOI: 10.1080/17517575.2019.1633689.
10. Cox M., Ellsworth D. (1997) Application-controlled demand paging for out-of-core visualization. Proceedings of the *8th IEEE Visualization Conference (IEEE Vis 1997), Phoenix, AZ, USA, 19–24 October 1997*, pp. 235–244. DOI: 10.1109/VISUAL.1997.663888.
11. Addo-Tenkorang R., Helo P.T. (2016) Big data applications in operations/supply-chain management: A literature review. *Computers & Industrial Engineering*, vol. 101, pp. 528–543. DOI: 10.1016/j.cie.2016.09.023.
12. Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Byers A.H. (2011) *Big data: The next frontier for innovation, competition, and productivity*. Available at: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation# (accessed 15 December 2020).
13. Laney D. (2001) 3D data management: *Controlling data volume, velocity, and variety*. Technical report. META Group.
14. Russom P. (2011) *Big data analytics*. TDWI best pracrice report, 4th quarter. Renton, WA: TWDI.
15. Kwon O., Sim J.M. (2013) Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, vol. 40, no 5, pp. 1847–1857. DOI: 10.1016/j.eswa.2012.09.017.
16. McAfee A., Brynjolfsson E. (2012) Big data: the management revolution. *Harvard Business Review*, vol. 90, no 10, pp. 61–67.
17. Oracle (2013) Oracle: *Big data for the enterprise*. White paper. Redwood Shores, CA: Oracle Corp.
18. Wamba S.F., Akter S., Edwards A., Chopin G., Gnanzou D. (2015) How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, vol. 165, pp. 234–246. DOI: 10.1016/j.ijpe.2014.12.031.
19. White M. (2012) Digital workplaces: Vision and reality. *Business Information Review*, vol. 29, no 4, pp. 205–214. DOI: 10.1177/0266382112470412.
20. IDC (2013) *Big data in 2020*. Available at: https://www.emc.com/leadership/digital-universe/2012iview/big-data-2020.htm (accessed 15 December 2020).
21. Boyd D., Crawford K. (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, vol. 15, no 5, pp. 662–679. DOI: 10.1080/1369118X.2012.678878.
22. Chen H., Chiang R., Storey V. (2012) Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, vol. 36, no 4, pp. 1165–1188. DOI: 10.2307/41703503.
23. Dremel C., Herterich M., Wulf J., Brocke J. (2020) Actualizing big data analytics affordances: A revelatory case study. *Information & Management*, vol. 57, no 1, article no 103121. DOI: 10.1016/j.im.2018.10.007.
24. Ghasemaghaei M. (2020) The role of positive and negative valence factors on the impact of bigness of data on big data analytics usage. *International Journal of Information Management*, vol. 50, pp. 395–404. DOI: 10.1016/J.IJINFOMGT.2018.12.011.
25. Mikalef P., Krogstie J., Pappas I., Pavlou P. (2020) Exploring the relationship between big data analytics capability and competitive performance: The mediating roles of dynamic and operational capabilities. *Information & Management*, vol. 57, no 2, article no 103169. DOI: 10.1016/j.im.2019.05.004.

26. Strawn G.O. (2012) Scientific research: How many paradigms? *Educause Review*, vol. 47, no 3, pp. 26–34.
27. Golchha N. (2015) Big data – The information revolution. *International Journal of Applied Research*, vol. 1, no 12, pp. 791–794.
28. Wamba S.F., Akter S. (2015) Big data analytics for supply chain management: A literature review and research agenda. *Enterprise and Organizational Modeling and Simulation (EOMAS 2015)* (J. Barjis, R. Pergl, E. Babkin, eds.). Lecture Notes in Business Information Processing, vol. 231, pp. 61–72. DOI: 10.1007/978-3-319-24626-0_5.
29. Mital R., Coughlin J., Canaday M. (2014) Using big data technologies and analytics to predict sensor anomalies. Proceedings of the *Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS 2014), Wailea, Maui, Hawaii, 15–18 September 2014*, pp. 84.
30. Chen M., Mao S., Zhang Y., Leung V.C. (2014) *Big data: Related technologies, challenges and future prospects*. Springer. DOI: 10.1007/978-3-319-06245-7.
31. Marjani M., Nasaruddin F., Gani A., Karim A., Hashem I.A.T., Siddiqa A., Yaqoob I. (2017) Big IoT data analytics: Architecture, opportunities, and open research challenges. *IEEE Access*, vol. 5, pp. 5247–5261.
32. Chen C.P., Zhang C.-Y. (2014) Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, vol. 275, pp. 314–347. DOI: 10.1016/j.ins.2014.01.015.
33. Pfaffl M.W. (2001) A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Research*, vol. 29, no 9, p. e45. DOI: 10.1093/nar/29.9.e45.
34. Jourdan Z., Rainer P.K., Marshall T.E. (2008) Business intelligence: An analysis of the literature. *Information Systems Management*, vol. 25, no 2, pp. 121–131. DOI: 10.1080/10580530801941512.
35. Bifet A., Holmes G., Kirkby R., Pfahringer B. (2010) MOA: Massive online analysis. *Journal of Machine Learning Research*, no 11, pp. 1601–1604.
36. Kang Y.-S., Park I.-H., Rhee J., Lee Y.-H. (2016) MongoDB-based repository design for IoT generated RFID/sensor big data. *IEEE Sensors Journal*, vol. 16, no 2, pp. 485–497. DOI: 10.1109/JSEN.2015.2483499.
37. Jiang L., Xu L.D., Cai H., Jiang Z., Bu F., Xu B. (2014) An IoT-oriented data storage framework in cloud computing platform. *IEEE Transactions on Industrial Informatics*, vol. 10, no 2, pp. 1443–1451. DOI: 10.1109/TII.2014.2306384.
38. O'Leary D.E. (2013) 'Big data', the 'internet of things' and the 'internet of signs'. *Intelligent Systems in Accounting, Finance and Management*, vol. 20, no 1, pp. 53–65. DOI: 10.1002/isaf.1336.
39. Hagstrom M. (2012) High-performance analytics fuels innovation and inclusive growth: use big data, hyperconnectivity and speed to intelligence to get true value in the digital economy. *Journal of Advanced Analytics*, no 2, pp. 3–4.
40. Kenny J. (2014) *Big data can have big impact on supply chain management: The use of data analytics is underused in supply chain management to minimize risk exposure*. InsideCounsel.
41. Vasan S. (2014) Impact of big data and analytics in supply chain execution. *Supply Chain Digital*. Available at: https://www.supplychaindigital.com/logistics-1/impact-big-data-and-analytics-supply-chain-execution (accessed 15 December 2020).
42. Heudecker N., Buytendijk F., Kart L. (2013) Survey analysis: *Big data adoption in 2013 shows substance behind the hype*. Available at: https://www.gartner.com/en/documents/2589121/survey-analysis-big-data-adoption-in-2013-shows-substanc (accessed 15 December 2020).
43. Kamble S., Gunasekaran A. (2018) Big data-driven supply chain performance measurement system: a review and framework for implementation. *International Journal of Production Research*, vol. 58, no 1, pp. 65–86. DOI: 10.1080/00207543.2019.1630770.
44. Dubey R., Gunasekaran A., Childe S.J. (2019) Big data analytics capability in supply chain agility: The moderating effect of organizational flexibility. *Management Decision*, vol. 57, no 8, pp. 2092–2112. DOI: 10.1108/MD-01-2018-0119.
45. Mayer-Schönberger V., Cukier K. (2013) *Big data: A revolution that will transform how we live, work, and think*. New York: Eamon Dolan/Mariner Book.
46. Rozados I.V., Tjahjono B. (2014) Big data analytics in supply chain management: Trends and related research. Proceedings of the *6th International Conference on Operations and Supply Chain Management (OSCM), Sanur, Bali, 10–12 December 2014*, pp. 1096–1107. DOI: 10.13140/RG.2.1.4935.2563
47. Waller M.A., Fawcett S.E. (2013) Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, vol. 34, no 2, pp. 77–84. DOI: 10.1111/jbl.12010.
48. Bates D.W., Saria S., Ohno-Machado L., Shah A., Escobar G. (2014) Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, vol. 33, no 7, pp. 1123–1131. DOI: 10.1377/hlthaff.2014.0041.

49. Hopkins J. (2016) A comparative study examining academic cohorts with transnational migratory intentions towards Canada and Australia. *Higher Education Quarterly*, vol. 70, no 3, pp. 246–263. DOI: 10.1111/hequ.12091.
50. Kim G.-H., Trimi S., Chung J.-H. (2014) Big-data applications in the government sector. *Communications of the ACM*, vol. 57, no 3, pp. 78–85. DOI: 10.1145/2500873.
51. Sellitto C., Hawking P. (2015) Enterprise systems and data analytics: A fantasy football case study. *International Journal of Enterprise Information Systems*, vol. 11, no 3, pp. 1–12.
52. Hopkins J., Hawking P. (2017) Big data analytics and IoT in logistics: a case study. *International Journal of Logistics Management*, vol. 29, no 2, pp. 575–591. DOI: 10.1108/IJLM-05-2017-0109.
53. Sarma S., Brock D.L., Ashton K. (2000) *The networked physical world: Proposals for engineering the next generation of computing, commerce & automatic-identification*. White paper. Available at: https://cocoa.ethz.ch/downloads/2014/06/None_MIT-AUTOID-WH-001.pdf (accessed 15 December 2020).
54. Greengard S. (2015) *The internet of things*. Cambridge, MA: MIT Press.
55. Xu L.D., He W., Li S. (2014) Internet of things in industries: A survey. *IEEE Transactions on Industrial Informatics*, vol. 10, no 4, pp. 2233–2243. DOI: 10.1109/TII.2014.2300753.
56. Ben-Daya M., Hassini E., Bahroun Z. (2019) Internet of things and supply chain management:a literature review. *International Journal of Production Research*, vol. 57, no 15–16, pp. 4719–4742. DOI: 10.1080/00207543.2017.1402140.
57. Atzori L., Iera A., Morabito G. (2010) The internet of things: A survey. *Computer Networks*, vol. 54, no 15, pp. 2787–2805. DOI: 10.1016/j.comnet.2010.05.010.
58. Gubbi J., Buyya R., Marusic S., Palaniswami M. (2013) Internet of things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, vol. 29, no 7, pp. 1645–1660. DOI: 10.1016/j.future.2013.01.010.
59. Leminen S., Rajahonka M., Wendelin R., Westerlund M. (2020) Industrial internet of things business models in the machine-to-machine context. *Industrial Marketing Management*, vol. 84, pp. 298–311. DOI: 10.1016/j.indmarman.2019.08.008.
60. Zhou K., Liu T., Zhou L. (2015) Industry 4.0: Towards future industrial opportunities and challenges. Proceedings of the *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2015), Zhangjiajie, China, 15–17 August 2015*, pp. 2147–2152. DOI: 10.1109/FSKD.2015.7382284.
61. Li L. (2011) Application of the internet of thing in green agricultural products supply chain management. Proceedings of the *2011 Fourth International Conference on Intelligent Computation Technology and Automation (ICICTA), Shenzhen, China, 28–29 March 2011*, vol. 1, pp. 1022–1025. DOI: 10.1109/ICICTA.2011.256.
62. Kopetz H. (2011) *Real-time systems. Design principles for distributed embedded applications*. Springer.
63. Xia F., Yang L., Wang L., Vinel A. (2012) Internet of things. *International Journal of Communication Systems*, vol. 25, no 9, pp. 1101–1102. DOI: 10.1002/dac.2417.
64. Wortmann F., Flüchter K. (2015) Internet of things. *Business & Information Systems Engineering*, vol. 57, no 3, pp. 221–224. DOI: 10.1007/s12599-015-0383-3.
65. Postcapes (2017) *IoT standards and protocols*. Available at: https://www.postscapes.com/internet-of-thingsprotocols/ (accessed 15 September 2020).
66. Lee I., Lee K. (2015) The internet of things (IoT): Applications, investments, and challenges for enterprises. *Business Horizons*, vol. 58, no 4, pp. 431–440. DOI: 10.1016/j.bushor.2015.03.008
67. Zeinab K.A.M., Elmustafa S.A.A. (2017) Internet of things applications, challenges and related future technologies. *World Scientific News*, vol. 67, no 2, pp. 126–148.
68. Davies R. (2015) *The internet of things opportunities and challenges*. Available at: https://epthinktank.eu/2015/05/21/the-internet-of-things-opportunities-and-challenges/ (accessed 15 December 2020).
69. Porter M.E., Heppelmann J.E. (2014) How smart, connected products are transforming competition. *Harvard Business Review*, vol. 92, pp. 11–64.
70. Uckelmann D., Harrison M., Michahelles F. (2011) An architectural approach towards the future internet of things. *Architecting the internet of things*. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-19157-2_1.
71. Sun C. (2012) Application of RFID technology for logistics on internet of things. *AASRI Procedia*, vol. 1, pp. 106–111. DOI: 10.1016/j.aasri.2012.06.019.
72. Bandyopadhyay D., Sen J. (2011) Internet of things applications, challenges and related future technologies. *Wireless Personal Communications*, no 58, pp. 49–69. DOI: 10.1007/s11277-011-0288-5.
73. Langley D.L., van Doorn J., Ng I.C.L., Stieglitz S., Lazovik A., Boonstra A. (2020) The internet of everything: Smart things and their impact on business models. *Journal of Business Research*, vol. 122, pp. 853–863. DOI: 10.1016/j.jbusres.2019.12.035.

74. Yuvaraj S., Sangeetha M. (2016) Smart supply chain management using internet of things (IoT) and low power wireless communication systems. Proceedings of the *IEEE 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 23–25 March 2016*, pp. 555–558. DOI: 10.1109/WiSPNET.2016.7566196.
75. Tao F., Ying Z., Xu L.D., Zhang L. (2014) IoT-based intelligent perception and access of manufacturing resource toward cloud manufacturing. *IEEE Transactions on Industrial Informatics*, vol. 10, no 2, pp. 1547–1557. DOI: 10.1109/TII.2014.2306397.
76. Gnimpieba Z.D.R., Nait-Sidi-Moh A., Durand D., Fortin J. (2015) Using internet of things technologies for a collaborative supply chain: Application to tracking of pallets and containers. *Procedia Computer Science*, vol. 56, pp. 550–557. DOI: 10.1016/j.procs.2015.07.251.
77. Verdouw C.N., Robbemond R.M., Verwaart T., Wolfert J., Beulens A.J.M. (2018) A reference architecture for IoT-based logistic information systems in agri-food supply chains. *Enterprise Information Systems*, vol. 12, no 7, pp. 755–779. DOI: 10.1080/17517575.2015.1072643.
78. Decker C., et al. (2008) Cost-benefit model for smart items in the supply chain. *The internet of things* (C. Floerkemeier, M. Langheinrich, E. Fleisch, F. Mattern, S.E. Sarma, eds.). Lecture Notes in Computer Science, vol. 4952, pp. 155–172. DOI: 10.1007/978-3-540-78731-0_10.
79. Xu L.D. (2011) Information architecture for supply chain quality management. *International Journal of Production Research*, vol. 49, no 1, pp. 183–198. DOI: 10.1080/00207543.2010.508944.
80. Chen R.-Y. (2015) Intelligent IoT-enabled system in green supply chain using integrated FCM method. *International Journal of Business Analytics*, vol. 2, no 3, pp. 47–66. DOI: 10.4018/IJBAN.2015070104.
81. Yan J., Xin S., Liu Q., Xu W., Yang L., Fan L., Chen B., Wang Q. (2014) Intelligent supply chain integration and management based on cloud of things. *International Journal of Distributed Sensor Networks*, vol. 10, no 3, article no 624839. DOI: 10.1155/2014/624839.
82. Parry G.C., Brax S.A., Maull R.S., Ng I.C.L. (2016) Operationalising IoT for reverse supply: the development of use-visibility measures. *Supply Chain Management*, vol. 21, no 2, pp. 228–244. DOI: 10.1108/SCM-10-2015-0386.
83. Thoben K.D., Wiesner S., Wuest T. (2017) "Industrie 4.0" and smart manufacturing – A review of research issues and application examples. *International Journal of Automation Technology*, vol. 11, no 1, pp. 4–16.
84. Fleisch E. (2010) *What is the internet of things? An economic perspective, business process and applications.* White Paper WP-BIZAPP-053. Available at: https://cocoa.ethz.ch/downloads/2014/06/None_AUTOIDLABS-WP-BIZAPP-53.pdf (accessed 15 December 2020).
85. Abdel-Basset M., Nabeeh N.A., El-Ghareeb H.A., Aboelfetouh A. (2020) Utilising neutrosophic theory to solve transition difficulties of IoT-based enterprises. *Enterprise Information Systems*, vol. 14, no 9–10, pp. 1304–1324. DOI: 10.1080/17517575.2019.1633690.
86. Gonzalez J. (2011) *The impact of the internet of things on business and society.* Foundation de la Innovation Bankinter.
87. Poirier C.C., Bauer M.J. (2000) *E-supply chain: Using the internet to revolutionize your business.* San Francisco: Berrett-Koehler.
88. Vervest P., Preiss K., van Heck E., Pau L.-F. (2004) The emergence of smart business networks. *Journal of Information Technology*, vol. 19, no 4, pp. 228–233. DOI: 10.1057/palgrave.jit.2000024.
89. Wirtz B.W., Weyerer J.C., Schichtel F.T. (2019) An integrative public IoT framework for smart government. *Government Information Quarterly*, vol. 36, no 2, pp. 333–345. DOI: 10.1016/j.giq.2018.07.001.
90. Janssen M., Luthra S., Mangla S, Rana N.P., Dwivedi Y.K. (2019) Challenges for adopting and implementing IoT in smart cities: An integrated MICMAC-ISM approach. *Internet Research*, vol. 29, no 6, pp. 1589–1616. DOI: 10.1108/INTR-06-2018-0252.
91. Abdel-Basset M., Manogaran G., Mohamed M. (2018) Internet of things (IoT) and its impact on supply chain: A framework for building smart, secure and efficient systems. *Future Generation Computer Systems*, vol. 86, pp. 614–628. DOI: 10.1016/j.future.2018.04.051.
92. Rong K., Hu G., Lin Y., Shi Y., Guo L. (2015) Understanding business ecosystem using a 6C framework in Internet-of-Things-based sectors. *International Journal of Production Economics*, vol. 159, pp. 41–55. DOI: 10.1016/j.ijpe.2014.09.003.
93. Cegielski C.G., Jones Farmer L.A., Wu Y., Hazen B.T. (2012) Adoption of cloud computing technologies in supply chains. *International Journal of Logistics Management*, vol. 32, no 2, pp. 184–211. DOI: 10.1108/09574091211265350.
94. Suguna S.K., Kumar S.N. (2019) Application of cloud computing and internet of things to improve supply chain processes. *Edge computing: From hype to reality.* Springer, pp. 145–170.
95. Arlbjorn J.S., de Haas H.D., Munksgaard K.B. (2011) Exploring supply chain innovation. *Logistics Research*, vol. 3, no 1, pp. 3–18. DOI: 10.1007/s12159-010-0044-3.

96. Lee I. (2015) The Internet of Things (IoT) for supply chain innovation: a conceptual framework and analysis of Fortune 200 companies. *Asia Pacific Journal of Innovation and Entrepreneurship*, vol. 9, no 1, pp. 81–103.
97. Nozari H., Najafi E., Fallah M., Lotfi F.H. (2019) Quantitative analysis of key performance indicators of green supply chain in FMCG industries using non-linear fuzzy method. *Mathematics*, vol. 7, no 11, article no 1020. DOI: 10.3390/math7111020.
98. Kaur J., Sidhu R., Awasthi A., Srivastava S.K. (2019) A Pareto investigation on critical barriers in green supply chain management. *International Journal of Management Science and Engineering Management*, vol. 14, no 2, pp. 113–123. DOI: 10.1080/17509653.2018.1504237.
99. Rehman M.H., Yaqoob I., Salah K., Imran M., Jayaraman P.P., Perera C. (2019) The role of big data analytics in industrial Internet of Things. *Future Generation Computer Systems*, no 99, pp. 247–259. DOI: 10.1016/j.future.2019.04.020.
100. Yu M., Nagurney A. (2013) Competitive food supply chain networks with application to fresh produce. *European Journal of Operational Research*, vol. 224, no 2, pp. 273–282. DOI: 10.1016/j.ejor.2012.07.033.
101. Ji G., Hu L., Tan K.H. (2017) A study on decision-making of food supply chain based on big data. *Journal of Systems Science and Systems Engineering*, no 26, pp. 183–198. DOI: 10.1007/s11518-016-5320-6.
102. Osman A.M.S. (2019) A novel big data analytics framework for smart cities. *Future Generation Computer Systems*, no 91, pp. 620–633.

About the authors

Hamed Nozari
Ph.D.;
Lecturer, Department of Industrial Engineering, Islamic Azad University, Central Tehran Branch, Hamila Blvd., Poonak Sqr., Tehran 1469669191, Iran;
E-mail: Ham.nozari.eng@iauctb.ac.ir
ORCID: 0000-0002-6500-6708

Mohammad Fallah
Ph.D.;
Associated Professor, Department of Industrial Engineering, Islamic Azad University, Central Tehran Branch, Hamila Blvd., Poonak Sqr., Tehran 1469669191, Iran;
E-mail: Mohammad.fallah43@yahoo.com
ORCID: 0000-0001-5541-4284

Hamed Kazemipoor
Ph.D.;
Assistant Professor, Department of Industrial Engineering, Islamic Azad University, Central Tehran Branch, Hamila Blvd., Poonak Sqr., Tehran, 1469669191 Iran;
E-mail: Hkazemipoor@yahoo.com
ORCID: 0000-0002-1848-8927

Seyed Esmaeil Najafi
Ph.D.;
Assistant Professor, Department of Industrial Engineering, Islamic Azad University, Science and Research Branch, Daneshgah Blvd., Simon Bulivar Blvd., Tehran 1477893855, Iran;
E-mail: E.najafi@srbiau.ac.ir
ORCID: 0000-0002-3125-4439