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

Vol. 16 No. 1 – 2022

BUSINESS INFORMATICS

HSE SCIENTIFIC JOURNAL

CONTENTS

Publisher:
National Research University
Higher School of Economics

The journal is published quarterly

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

Acting Editor-in-Chief
E. Zaramenskikh

Computer Making-up:
O. Bogdanovich

Website Administration:
I. Khrustaleva

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

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

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

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

ABOUT THE JOURNAL

Business Informatics is a peer reviewed interdisciplinary academic journal published since 2007 by National Research University Higher School of Economics (HSE), Moscow, Russian Federation. The journal is administered by HSE Graduate School of Business. The journal is published quarterly.

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

The journal publishes papers in the areas of, but not limited to: modeling of social and economic systems, digital transformation of business, innovation management, information systems and technologies in business, data analysis and business intelligence systems, mathematical methods and algorithms of business informatics, business processes modeling and analysis, decision support in management.

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

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

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

EDITORIAL BOARD

ABOUT THE HIGHER SCHOOL OF ECONOMICS

Consistently ranked as one of Russia's top universities, the Higher School of Economics (HSE) is a leader in Russian education and one of the preeminent economics and social sciences universities in Eastern Europe and Eurasia.

Having rapidly grown into a well-renowned research university over two decades, HSE sets itself apart with its international presence and cooperation.

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

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

Since 2013, HSE has been a member of the 5-100 Russian Academic Excellence Project, a highly selective government program aimed at boosting the international competitiveness of Russian universities.

ABOUT THE GRADUATE SCHOOL OF BUSINESS

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

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

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

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

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

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

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

DOI: 10.17323/2587-814X.2022.1.7.21

Simulation of migration and demographic processes using FLAME GPU

Valery L. Makarov [a]
E-mail: makarov@cemi.rssi.ru

Albert R. Bakhtizin [a]
E-mail: albert@cemi.rssi.ru

Gayane L. Beklaryan [a]
E-mail: glbeklaryan@gmail.com

Andranik S. Akopov [b,a]
E-mail: aakopov@hse.ru

Nikita V. Strelkovskii [c]
E-mail: strelkon@iiasa.ac.at

[a] Central Economics and Mathematics Institute, Russian Academy of Sciences
Address: 47, Nakhimovsky Prospect, Moscow 117418, Russia

[b] HSE University
Address: 20, Myasnitskaya Street, Moscow 109028, Russia

[c] International Institute for Applied Systems Analysis (IIASA)
Address: 1, Schlossplatz, Laxenburg A-2361, Austria

Abstract

This article presents an approach to modeling migration and demographic processes using a framework designed for large-scale agent-based modeling – FLAME GPU. This approach is based on the previously developed simulation model of interaction between two communities: migrants and natives that is implemented in the AnyLogic simulation software. The model has had a low dimensionality of the discrete space representing the operating environment of the agent populations and a deterministic decision-making system of each agent. At the same time, the presence of multiple interactions between agents and transitions between their states determines a high computational complexity of such a model. The use of FLAME GPU makes it possible to conduct extensive simulation experiments with the model, mainly due to the parallelization of computational processes at the level of each agent, as well as the implementation of the mechanism of multiple computations using Monte Carlo techniques. The developed framework is used to study the impact of the most important parameters of the model (e.g., rate of migration, governmental expenditures on integration, frequency of creation of new workplaces, etc.) on the key outputs of the modeled socio-economic system (in particular, population size, share of migrants, number of assimilated migrants, GDP growth rate, etc.). The proposed approach can be used to develop decision-making systems for planning the hiring of new employees based on the forecast dynamics of migration and demographic processes.

Keywords: agent-based modeling, migration and demographic processes, population dynamics, large-scale modeling, parallel computing on GPU, supercomputer modeling, decision support

Citation: Makarov V.L., Bakhtizin A.R., Beklaryan G.L., Akopov A.S., Strelkovskii N.V. (2022) Simulation of migration and demographic processes using FLAME GPU. *Business Informatics*, vol. 16, no. 1, pp. 7–21. DOI: 10.17323/2587-814X.2022.1.7.21

Introduction

In modern times, many companies and organizations are faced with a deficit of labor resources and the need to form a long-term plan for hiring new employees taking into account forecasts of migration flows and demographic processes. In conditions of the degradation of the internal demographic situation, many firms try refocusing themselves to attract migrants. However, due to the presence of many barriers created by an insufficient level of proficiency in the local language, lack of necessary qualifications and some other factors, there are natural limitations in attracting external labor which can only be overcome through assimilation and integration processes implemented in conditions of financial support from the government.

The development of decision-making systems for planning the hiring of new employees and creation of new workplaces can be based on simulation models that take into account the forecast dynamics of the labor market. For instance, in the context of the spread of epidemics, the government is able to introduce restrictions on the inflow of external labor, leading to a deficit of labor resources in sectors of the economy focused on migrants. On the other hand, the development of high-tech enterprises necessitates the creation of new workplaces that are attractive for highly skilled natives. At the same time, providing the rational coexistence of two interacting communities, migrants and indigenous people, is an important challenge for business and government.

As a result, the problem of studying and forecasting migration and demographic processes using simulation methods is being updated. Such methods, in particular, the agent-based approach (ABM), make it possible to construct and investigate the behavior of a digital community consisting of agents with their own individual rules of behavior.

Among the well-known agent-based models of discrete type (i.e., with a discrete space of agents' existence), one can single out the well-known Sugarscape model' [1], which has become widespread as a tool for analyzing the attractiveness of local areas with resources ('conditional sugar') for agents. The model of 'nomads and farmers' [2, 3], in which some agents ('conditional farmers') create resources, while others destroy them in order to expand personal space ('conditional nomads') should be noted. Also, the models of population segregation of the Schelling class [4, 5], models of movement of an ensemble of unmanned vehicles [6, 7], models of migration and demographic processes [8–10], etc. are well known.

Such systems as NetLogo, AnyLogic [11], as well as systems designed for supercomputer agent-based modeling Repast HPC [12], MASS CUDA [13, 14], FLAME GPU [15–18], etc. can be highlighted among the simulation ABM-tools intended for general purposes.

Most of these systems differ in the way of software implementation of agents: either using only one central processor (for example, NetLogo, AnyLogic), or using a multi-cluster architecture based on CPU (e.g., Repast HPC) and MPI (Message Passing Interface) or they use GPU (Graphics Processing Unit) [19]. Among these platforms is FLAME GPU 2[1], which is characterized by a number of advantages. The framework is open source software, supports the ability to visualize a model using OpenGL [20]. Moreover, it allows multiple runs of an ensemble of models [21–24] on a personal computer (PC) using Visual Studio C++ and a Linux operating system on supercomputer systems based on NVIDIA CUDA[2] (e.g., NVIDIA QUADRO RTX, NVIDIA Tesla, etc.). As a result, a flexible approach to the development of ABMs on conventional PCs and computational experiments on GPU-clusters is provided.

The FLAME GPU framework has been used to develop agent-based models in many fields ranging from biology to economics. As far as the authors of this article know, at the time of writing, the FLAME GPU platform was used to simulate migration processes in only two works [25, 26]. Other closest works are modeling the behavior of agents in the 'sugar' model described above [27]. Modeling migration processes causes additional technical complexity since dynamic creation of agent-migrants during the simulation is needed.

This work is aimed at developing the agent-based model of the dynamics of population and consists of two interacting communities: natives and migrants with the software implementation using the supercomputer simulation system as FLAME GPU. The proposed approach made it possible to carry out a series of variation experiments and to identify important relationships in the dynamics of the migration and demographic processes under study.

1. General description of the model

An artificial socio-economic system, consisting of native and foreign populations interacting with each other through both personal contacts of the 'agent-agent' type and through message exchanges is considered. In such a system, agents are both individuals (indigenous people and migrants) and resources that have the 'ability' to assess the nearest agents and send them information about their state and correspondence to agent interests. At the same time, high technology resources correspond more to natives, and low-technology resources are associated with migrants.

Thus, an important feature of the implementation of the simulation model of the interaction of communities of migrants and natives is the mechanism of continuous messaging between agents and resources supported in the FLAME GPU.

As before, multiparticle interactions between agents are simulated in two-dimensional discrete space with a relatively small dimensionality of 100 × 100 cells and a capacity of not more than 10 000 agents. At the same time, the implementation of the model in the FLAME GPU is aimed largely at increasing the time-efficiency of the model in conditions of performing multiple recalculations using methods of the Monte Carlo type. The dimensionality of the model discrete space, which limits both the number of available workplaces and agents associated

[1] https://flamegpu.com/

[2] https://developer.nvidia.com/

with them, can be significantly increased almost without loss of the time-efficiency, mainly due to the parallelization of computational procedures ('agent-level functions') with the use of graphics processing units (GPUs). In such a discrete space, each agent can occupy only one cell with or without a workplace at each moment. At the same time, the complexity of the parallel implementation of the model under consideration and the difficulties of automatic synchronisation of agents in the FLAME GPU necessitate additional control of this rule and eliminate possible collisions if they occur.

As in the past, the creation of 'high-tech' and 'low-tech' workplaces, which are targeted by natives and migrants respectively, is provided. Both types of jobs provide the individual agents who occupy them with an increase in personal comfort. At the same time, the neighborhood with agent-migrants is negatively affected on the personal comfort level of natives, which is caused by the existing cultural differences and psychological characteristics of the agents.

All workplaces are created centrally and uniformly in a random way in all free cells of discrete space with different probabilities set for 'high-tech' and 'low-tech' jobs, respectively. Despite the fact that this approach is the most costly for the state, it can help to avoid a shortage of jobs, which is especially important at high rates of migration. In addition, a greater number of evenly distributed jobs allows a significant increase in the number of mutual contacts between indigenous people and migrants, which has a positive effect on the level of proficiency in the local language, the possibility of obtaining a relevant local education, etc., all of which helps to reduce the time required for assimilation and integration.

The contribution of 'high-tech' and 'low-tech' workplaces, usually occupied by natives and migrants, to economic growth (GDP) and government transfers (GT) is different. The ratio of GDP to labor resources units in the 'high-tech' sectors of the economy is significantly higher than in the 'low-tech' branches. At the same time, the creation of 'low-tech' workplaces leads to additional government spending, which also increases with the growth in the number of 'unemployed' agents.

The model provides for an inflow of migrants with the subsequent 'transformation' of arriving agents into indigenous people after the period required for assimilation has expired (1–30 years). Immigration is mainly due to the 'gravity effect' [10], which sets a reinforcing feedback between the number of available non-assimilated migrants and the inflow intensity of new agents. At the same time, in the model, the share of new immigrants (of the number of existing ones), as well as the costs of education and integration, are the key control parameters. Therefore, the migration process in such a system can be considered as 'controlled' and 'manned' by a decision-maker (i.e., the government).

At each simulation moment, agents search for the nearest workplace corresponding to their type. At the same time, a feature of implementation of the model in the FLAME GPU is the reverse order of doing this procedure, i.e., resources (jobs) search for the most suitable agents for themselves, and, in the case of a positive outcome, assign them their coordinates as target cells, while blocking access to all other agents.

In addition, agents-natives and agent-migrants search for the most suitable partner for marriage and childbirth (i.e., taking into account age, marital status, etc.). While the personal comfort level of an agent is below the threshold level, it looks for a workplace. If the agent comfort level is equal to or higher than the threshold level and it does not have a partner yet, then it searches for a partner for marriage and childbirth (taking into account, suitable age and other required agent characteristics).

Thus, all agent- individuals can be in a stationary state, a state of searching for a workplace, a state of searching for a partner, a state of being ready to have children, etc. At the same time, agent-migrants can also transfer to an assimilation state after a certain time interval that is an endogenous characteristic of the model. Agent-natives are characterized by higher values of thresholds (in particular, their minimum level of personal comfort is higher regarding the appropriate level of agent-migrants), which determine their transition to new states, for example, the state of searching for a workplace, a stationary state, the birth of children, etc.

The abstract description of the problem statement and model dependencies (without taking into account the effect of assimilation and integration) are presented in [8] in detail.

2. Software implementation

The main computational procedures and functions of the proposed simulation model taking into account the conditional sequence of their execution are described with *Table 1*. Functions of the **FLAMEGPU_STEP_FUNCTION** type are implemented at each model time at the central processing unit (CPU) level and the functions of the **FLAMEGPU_AGENT_FUNCTION** class are sequentially executed in parallel computations using graphic processors (GPUs). At the same time, higher performance of agent-based model implementation in comparison with the traditional approach is achieved by parallelizing the operations logic of each agent and exchanging messages with each other taking into account their spatial location.

In *Table 1*, ‘agent data’ refers to characteristics of agents for natives and migrants (e.g., gender, age, marital status, agent type, etc.), and ‘resource data’ are related to characteristics of workplaces (e.g., resource type, ‘occupied / vacancy’, etc.).

The developed simulation model supports two main ways of performing computational procedures:

- single runs, executed for one selected scenario with fixed values of control parameters and visualization of the state of agents using the Open GL libraries [20];
- multiple runs implemented using the method of the Monte Carlo class [21–24] due to the parallel launch of the simulation model in the so-called ‘ensemble’ mode. This approach allows you to vary the values of the control parameters in specified ranges, in particular, using uniform, normal, and other distribution functions with their own characteristics.

The visualization of agent states in FLAME GPU is performed using Open GL and, in particular, is a lattice of a given dimension, in the cells of which agents (i.e. migrants and natives) and resources (i.e. ‘high-tech’ and ‘low-tech’ workplaces) have been placed. In addition, there are free cells that do not contain resources and agents. At the same time, if an agent of working age occupies a cell that does not have a workplace, then it is considered as unemployed and the level of its personal comfort will gradually decrease. Note that the visualization of agents’ states and their dynamics, i.e. moving to new cells of the discrete space is realized at each moment of the model time. Such an approach makes it possible to qualitatively assess the populations’ development, considering the individual choice of the most preferred workplaces by agents, as well as to study the segregation effects, etc.

3. Results of numerical experiments

All computations were performed with a **FORSITE DSWS PRO supercomputer based on the QUADRO RTX 6000** over a time interval of 80 years. The total number of resource agents in the model is fixed (10 000) and it is limited by the dimension of a given discrete space (100 × 100). The number of native and migrant agents ranges from 0 to 10 000, and is the result of a simulation experiment. The values of the main parameters of the model are presented in *Table 2.*

Table 1.

Basic computational procedures and functions of the simulation model

Function name	Appointment	Input messages	Output messages
FLAMEGPU_INIT_FUNCTION (init_function)	The model initialization. Forming initial populations of natives, migrants and workplaces.	No	No
FLAMEGPU_STEP_FUNCTION (BasicOutput)	The arrival of new agent– migrants, birth of new agents (natives and migrants) in married couples (with more probability) and for single agents (with the lesser probability).	No	No
FLAMEGPU_STEP_FUNCTION (AgentUpdate)	The evaluation (i.e., collecting) of simulation results computed over the ensemble of agents at each moment of the simulation.	No	No
FLAMEGPU_EXIT_CONDITION (exit_condition)	Check doing the criterion of stopping the simulation.	No	No
FLAMEGPU_AGENT_FUNCTION (check_all_agents, MsgArray2D, MsgNone)	Checking and resolving the potential collisions caused by accidental placement of some agents in one cell of discrete space.	Agent data	No
FLAMEGPU_AGENT_FUNCTION (workplaces_creation, MsgNone, MsgNone)	Creation of new workplaces based on existing population of resources. The destroying of workplaces that are to be disap–pearance.	No	Resource data
FLAMEGPU_AGENT_FUNCTION (update_cell, MsgArray2D, MsgArray2D)	Information propagation among agents about available resources (workplaces). The check of a resource occupancy by another agent.	Agent data	Resource data
FLAMEGPU_AGENT_FUNCTION (check_cell, MsgArray2D, MsgArray2D)	Information propagation among other agents (natives and migrants) about existing agents (with their characteristics) and resources occupied by them. The identification of a resource type occupied by the agent.	Resource data	Agent data
FLAMEGPU_AGENT_FUNCTION (agent_to_agent_contacts, MsgArray2D, MsgNone)	The determination of the frequency of contacts of the 'agent–agent' type (within the 8–cells 'Moore neighborhood') to estimate (recalculate) the level of local language knowledge among migrants, and the personal comfort level of natives decreasing due to contacts with migrants.	Agent data	Agent data
FLAMEGPU_AGENT_FUNCTION (looking_for_partner, MsgArray2D, MsgArray2D)	The function of searching for the closest partner corresponding to specified criteria (e.g., the gender, age, marital status, etc.).	Agent data	Agent data
FLAMEGPU_AGENT_FUNCTION (get_married, MsgArray2D, MsgNone)	Getting married with an agent who sent a message with a unique identifier (ID).	Agent data	No
FLAMEGPU_AGENT_FUNCTION (looking_for_resource, MsgAr–ray2D, MsgArray2D)	The function of searching for an agent that is closely located regarding each workplace among agents which are in a workplaces search state. Assigning a target cell with a resource to the selected agent.	Agent data	Resource data
FLAMEGPU_AGENT_FUNCTION (update_agent_state, MsgNone, MsgNone)	Updating the state of each agent depending on the values of its characteristics (e.g., the personal comfort level, age, marital status, etc.).	No	No
FLAMEGPU_AGENT_FUNCTION (moving_trasaction, MsgArray2D, MsgNone)	A movement transaction of an agent in discrete space in order to occupy a chosen workplace, based on data about the target cell transmitted by the corresponding resource.	Resource data	No

Table 2.

The main parameters of the model

Parameter name	Minimum	Maximum
Share of new migrants (of the number of existing agent–migrants)	0.1	0.5
Share of government expenditure on education in GDP per capita	0.1	0.5
Lifetime of 'high–tech' workplaces	5	15
Lifetime of 'low–tech' workplaces	5	15
Frequency of creation of new workplaces	5	15
Life expectancy of natives	70	90
Life expectancy of migrants	60	80
Minimum age for marriage and childbirth of natives	18	30
Minimum age for marriage and childbirth of migrants	18	30
Minimum level of personal comfort for natives	3	10
Minimum level of personal comfort for migrants	3	10
Retirement age	60	75

Figures 1–4 show frequency diagrams for the most important characteristics of the system under study obtained using the Monte Carlo class method, aggregated with the proposed agent-based model through its control parameters and objective functions.

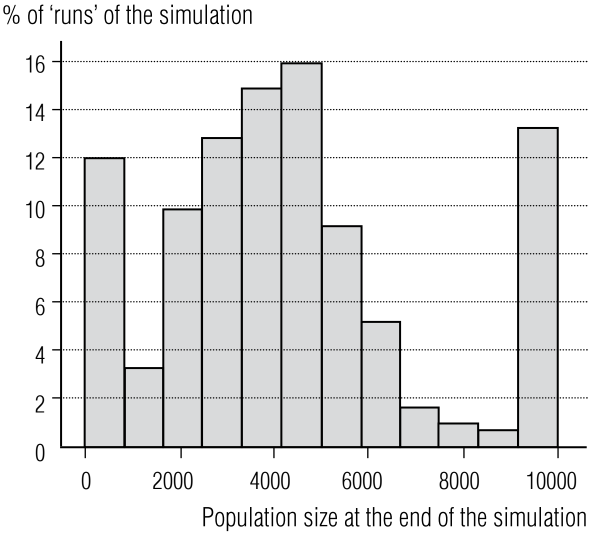

Fig. 1. Frequency diagram for population size.

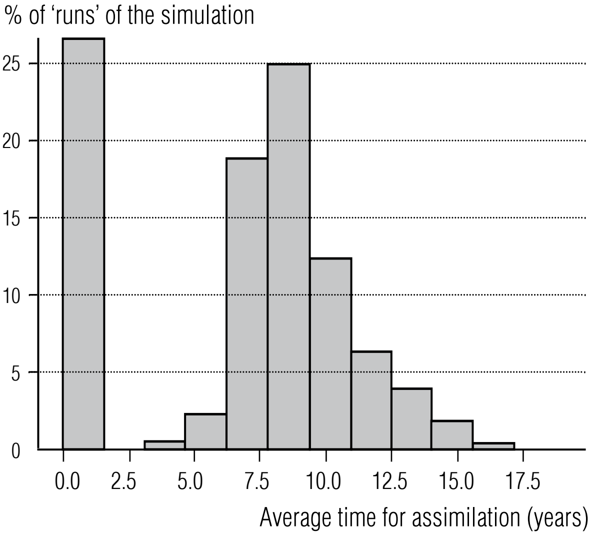

Fig. 2. Frequency diagram for the average time required for assimilation and integration.

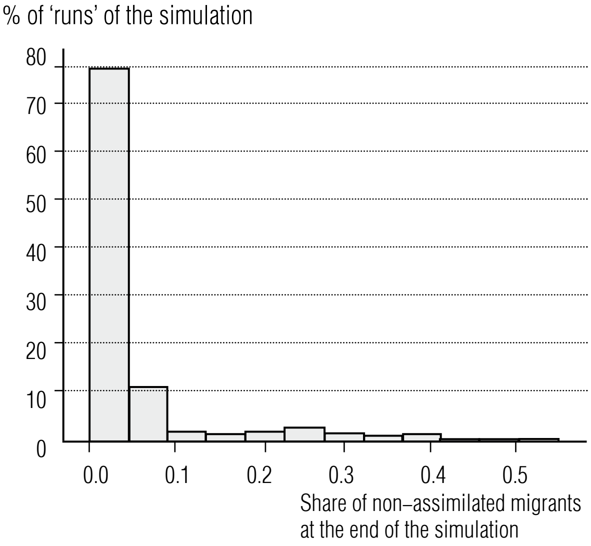

Fig. 3. Frequency diagram
for the share of non–assimilated migrants.

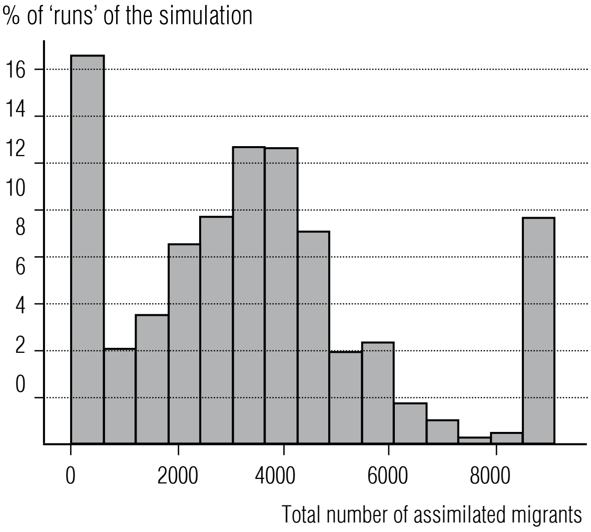

Fig. 4. Frequency diagram
for the total number of assimilated migrants.

In the process of conducting the numerical experiments, multiple runs of the model (more than 1000) were carried out and the scenarios most differing in the estimated characteristics were selected. They have been visualised with *Figs. 1–4.*

As follows from *Figs. 1–4*, the expected values of the modeled indicators have explicitly observed values. The frequent observability of the boundary values of indicators should be noted too. At the same time, it seems there are scenarios of some improvement in the required objective characteristics (e.g., the average time needed for assimilation), but they require a significant government expenditure on education, increasing the number of workplaces, etc.

As follows from *Fig. 5*, there is no unambiguous dependence of the share of non-assimilated migrants on the average time required for their assimilation – for the most frequent values of the first indicator (from 7 to 12 years), different values of the second are possible – from 0 to 0.55.

The data shown in *Fig. 6* demonstrate an almost linear dependence between the simulated population size (the total number of natives and agent-migrants) and the total number of assimilated migrants.

Further, the most important groups of scenarios for the evolutionary development of communities of migrants and indigenous people are considered:

- low-intensity and normal migration scenarios;
- scenarios of intensive and super-intensive migration.

The main characteristics of the scenarios to be studied are presented in *Table 3.*

In *Figs. 7–10* are shown the model dynamics of the key characteristics of the system under consideration over an 80-year simulation interval which is the result of the behavior of an ensemble of interacting agents-natives and migrants.

Figures 7–8 allow us to make the following important conclusion. With the existing patterns of agent behavior, a significant increase in the population size can be achieved only under conditions of intensive and super-intensive migration. However, such scenarios will cause a significant increase in the share of migrants in the population, which may lead to an increase in social tension.

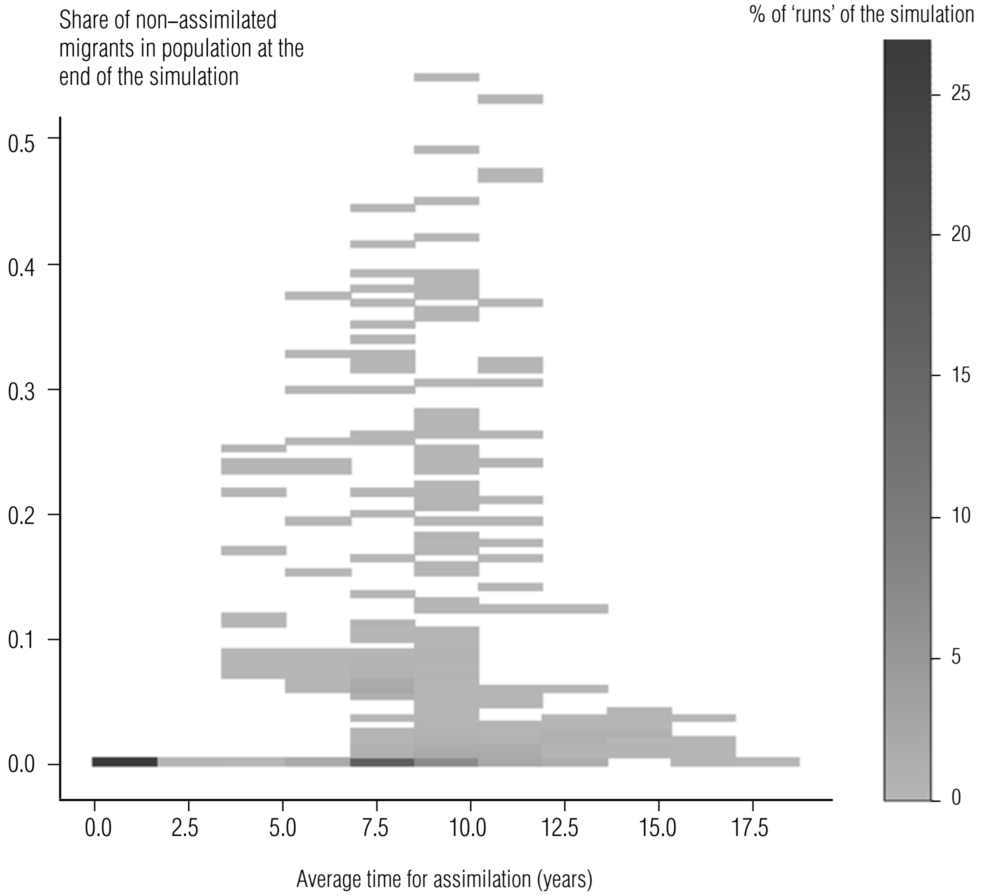

Fig. 5. Two–dimension frequency diagram for the average time needed for the assimilation and integration and share of non–assimilated migrants.

Table 3.

Studied scenarios and model assumptions

Group of scenarios	Scenario number	Share of new migrants	Share of government expenditure on education and integration
Low–intensity (normal) migration scenarios	Scenario 1	0.1	0.1
	Scenario 2	0.1	0.25
	Scenario 3	0.1	0.5
Intensive migration scenarios	Scenario 4	0.2	0.1
	Scenario 5	0.2	0.25
	Scenario 6	0.2	0.5
Super–intensive migration scenarios	Scenario 7	0.3	0.1
	Scenario 8	0.3	0.25
	Scenario 9	0.3	0.5

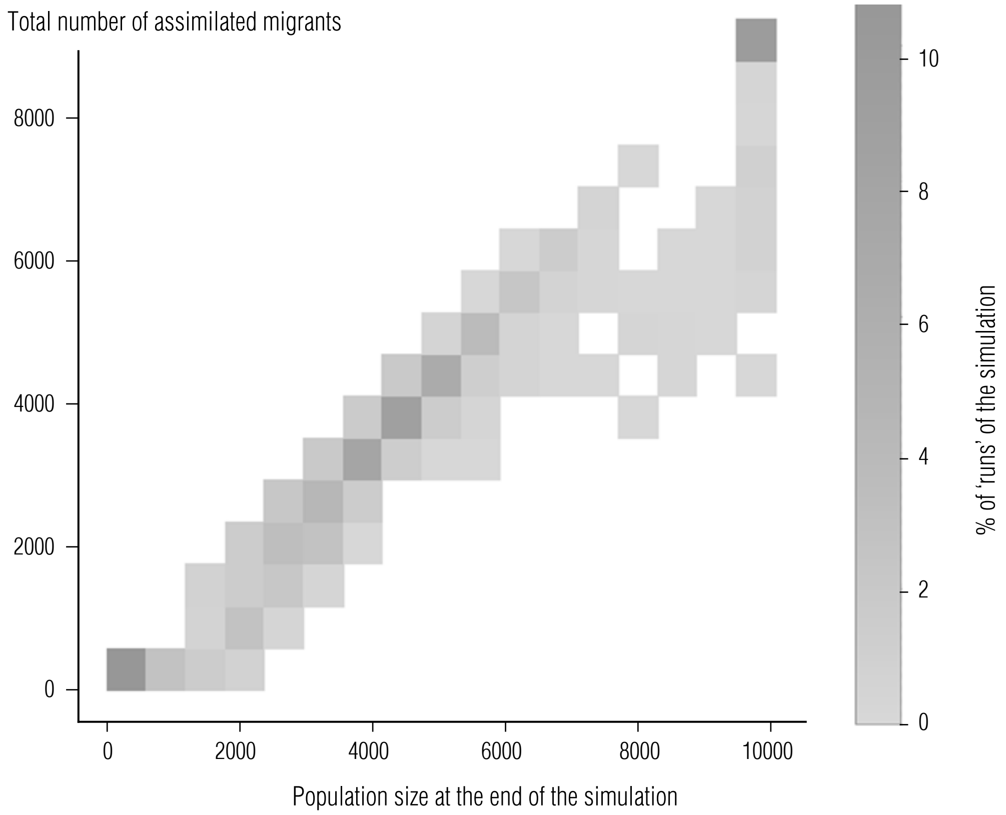

Fig. 6. Two-dimension frequency diagram for the population size and the total number of assimilated migrants in the model.

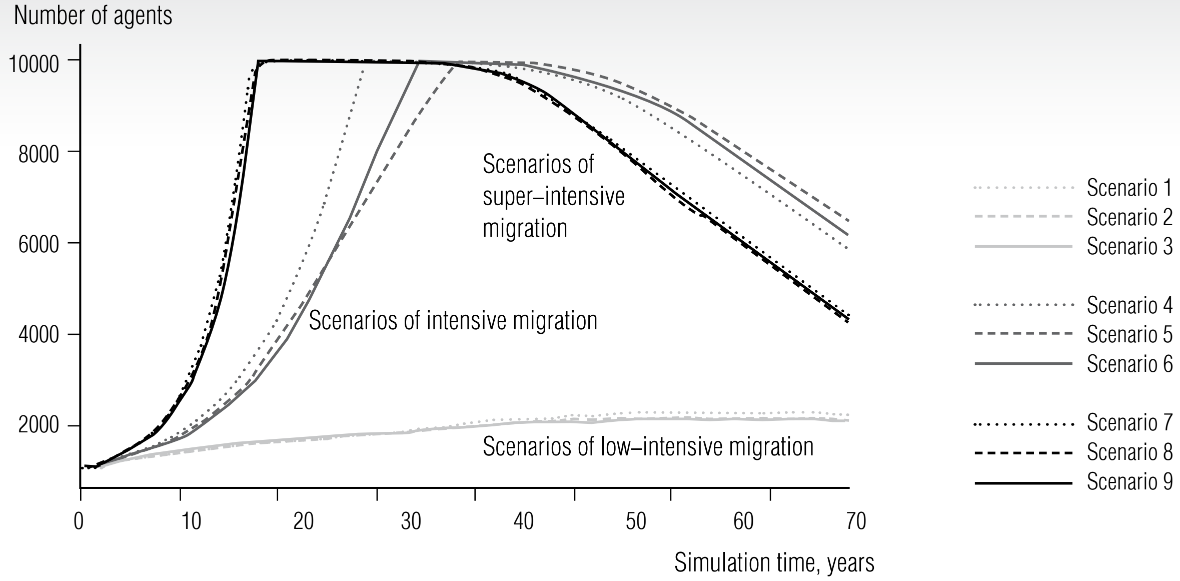

Fig. 7. Simulated population dynamics.

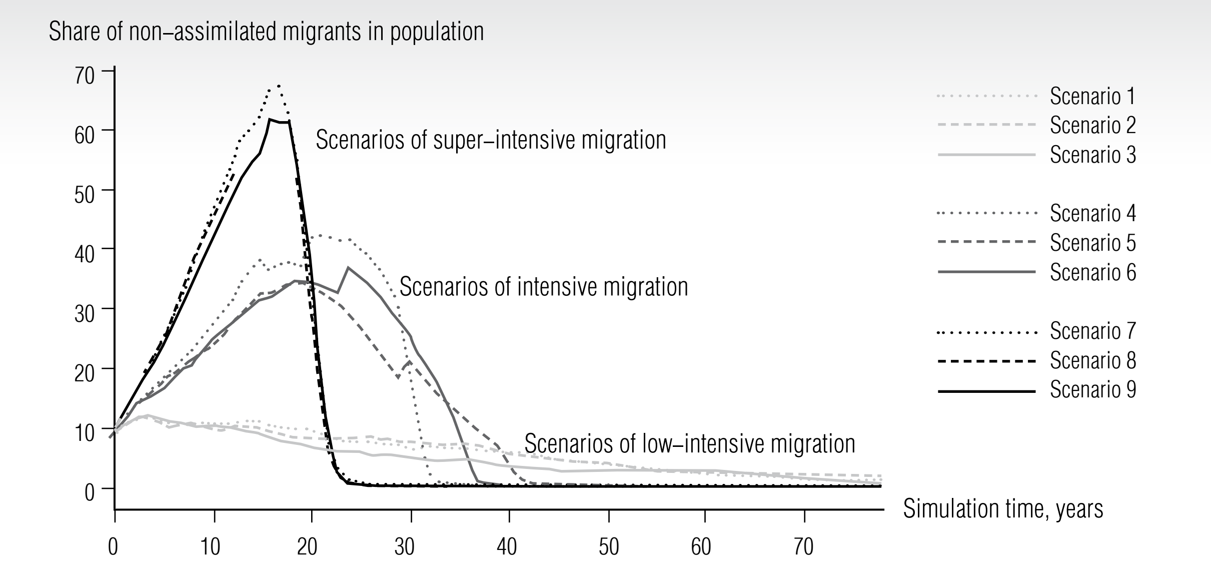

Fig. 8. Simulated dynamics of the share of non–assimilated migrants in the population.

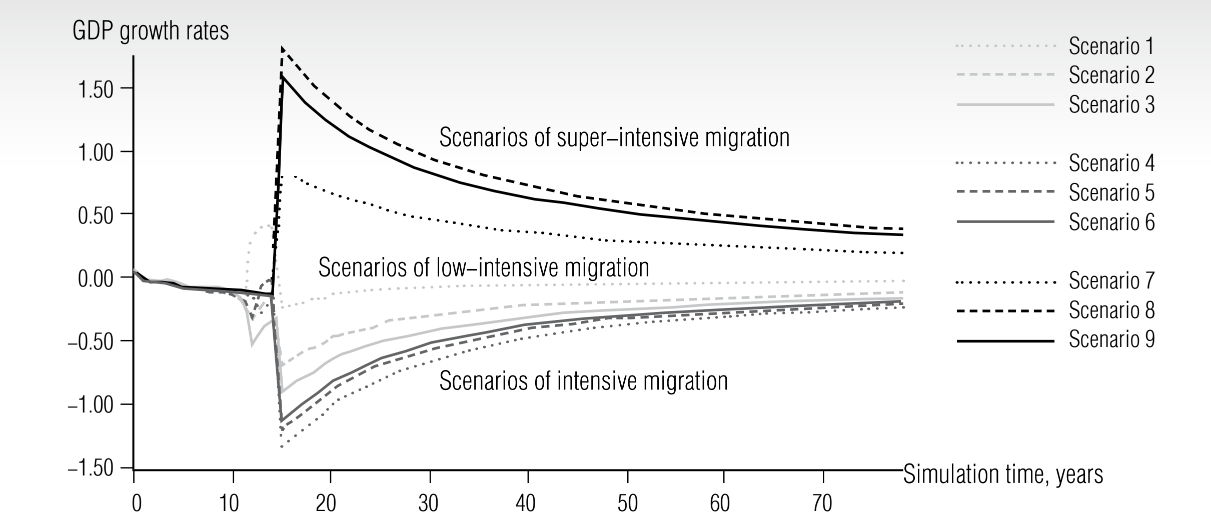

Fig. 9. Simulated dynamics of GDP growth rates.

Figure 9 shows that the highest rates of GDP growth can be achieved under scenarios of intensive migration, however, the subsequent shortage of resources leads to a gradual decrease in the rates of economic growth.

From *Fig. 10* it follows that scenarios of intensive and super-intensive migration cause a significant increase in government expenditure, mainly associated with the need to increase spending on education and integration of migrants, create appropriate jobs, pay unemployment benefits, etc.

Conclusion

This article presents a new approach to modeling migration and demographic processes using the FLAME GPU. The framework is intended

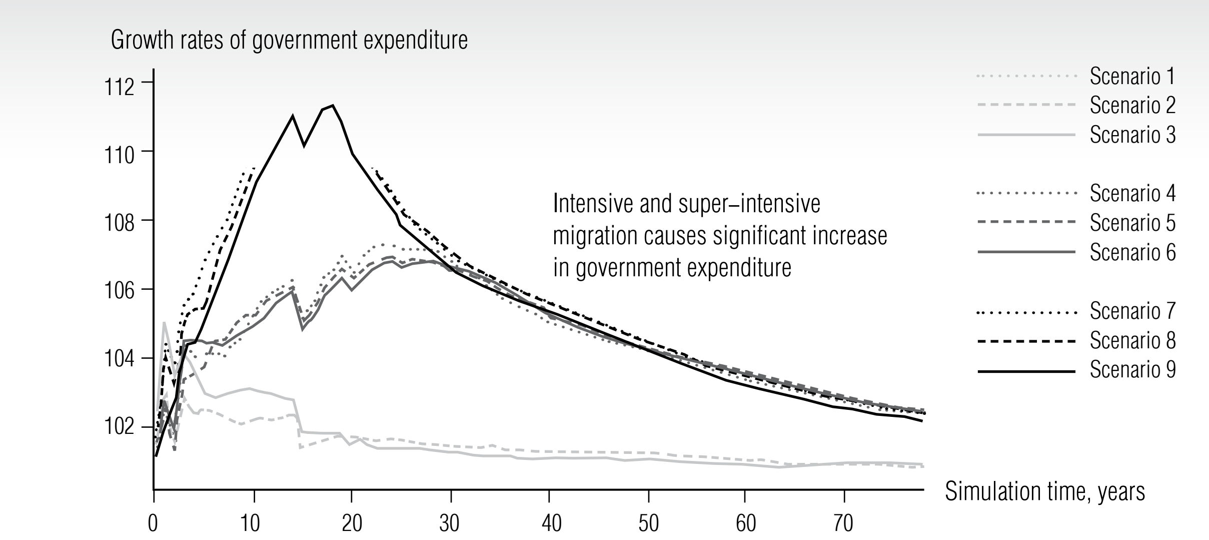

Fig. 10. Simulated dynamics of government expenditure.

for supercomputer agent-based modeling and it allows parallelizing the logic of the simulation model at the level of each agent, providing a significant increase in the time efficiency of the corresponding computational procedures.

As a result, using artificial data and methods of the Monte Carlo type, the most important characteristics of the model of interaction between natives and migrants were studied: the population size, average time needed for assimilation, share of non-assimilated migrants in the population, etc. The scenarios that provide a positive contribution to the economic and demographic growth have been found. At the same time, the implementation of such scenarios, based mainly on intensive migration, necessitates a significant increase in government expenditure on education and integration.

The proposed approach can be used to develop decision-making systems for planning hiring new employees based on the forecast dynamics of migration and demographic processes.

Further research will be aimed at complicating and detailing the model of interaction between migrants and indigenous people, using clustering methods for creating jobs, studying the effects of segregation, etc., using the FLAME GPU. ■

Acknowledgments

This study was funded by the Russian Foundation for Basic Research (RFBR) under the research project No. 18-51-14010.

References

1. Epstein J.M., Axtell R. (1996) *Growing artificial societies: Social sciences from the bottom up.* Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/3374.001.0001

2. Belousov F.A. (2018) Model of nomads and plowmen with a limited resource of spatial movement. *Ekonomika I Matematiceskie Metody*, vol. 54, no. 4, pp. 124–131 (in Russian). https://doi.org/10.31857/S042473880003336-8

3. Belousov F.A. (2017) Model of civilization with two types of reproduction of product (model of nomads and plowmen). *Ekonomika I Matematiceskie Metody*, vol. 53, no. 4, pp. 93–109 (in Russian).

4. Schelling T.C. (1971) Dynamic models of segregation. *The Journal of Mathematical Sociology*. Informa UK Limited, vol. 1, no. 2, pp. 143–186. https://doi.org/10.1080/0022250X.1971.9989794

5. Akopov A.S., Beklaryan A.L., Beklaryan L.A., Belousov F.A., Khachatryan N.K. (2020) Clustering in model of population segregation. *Artificial societies*, vol. 15, no. 4. (in Russian). https://doi.org/10.18254/S207751800012764-9

6. Akopov A.S., Beklaryan L.A., Beklaryan A.L., Belousov F.A. (2021) Simulation of motion of an ensemble of unmanned ground vehicles using FLAME GPU. *Information technologies*, vol. 27, no. 7, pp. 339–349 (in Russian). https://doi.org/10.17587/it.27.369-379

7. Akopov A.S., Beklaryan A.L. (2021) Scenario simulation of autonomous vehicles motion in artificial road network using FLAME GPU. *Artificial societies*, vol. 16, no. 1 (in Russian). https://doi.org/10.18254/S207751800014028-9

8. Makarov V.L., Bakhtizin A.R., Beklaryan G.L., Akopov A.S., Rovenskaya E.A., Strelkovskii N.V. (2020) Agent-based modelling of population dynamics of two interacting social communities: migrants and natives. *Ekonomika I Matematiceskie Metody*, vol. 56, no. 2, pp. 5-19 (in Russian). https://doi.org/10.31857/S042473880009217-7

9. Makarov V.L., Bakhtizin A.R., Beklaryan G.L., Akopov A.S. (2019) Agent-based model of migration to European Union countries with taking into account individual decision-making system. *Artificial societies*, vol. 14, no. 2 (in Russian). https://doi.org/10.18254/S207751800005804-3

10. Makarov V.L., Bakhtizin A.R., Beklaryan G.L., Akopov A.S., Rovenskaya E.A., Strelkovskii N.V. (2019) Aggregated agent-based simulation model of migration flows of the European Union Countries. *Ekonomika I Matematiceskie Metody*, vol. 55, no. 1. pp. 3–15 (in Russian). https://doi.org/10.31857/S042473880004044-7

11. Borshchev A. (2013) *The big book of simulation modeling: Multimethod modeling with AnyLogic 6.* AnyLogic North America.

12. Collier N., North M. (2013) Parallel agent-based simulation with repast for high performance computing. *Simulation*, vol. 89, no. 10, pp. 1215–1235. https://doi.org/10.1177/0037549712462620

13. Lysenko M., Roshan M.D. (2008) A framework for megascale agent based model simulations on graphics processing units. *Journal of Artificial Societies and Social Simulation*, vol. 11, no. 4. Available at: http://jasss.soc.surrey.ac.uk/11/4/10.html (accessed 30 June 2021).

14. Emau J., Chuang T., Fukuda M. (2011) A multi-process library for multi-agent and spatial simulation. *In Proc. of 2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing – PACRIM'11*, pp. 369–376. https://doi.org/10.1109/PACRIM.2011.6032921

15. Richmond P., Walker D., Coakley C., Romano D. (2010) High performance cellular level agent-based simulation with FLAME for the GPU. *Briefings in bioinformatics*, vol. 11, pp. 334–347. https://doi.org/10.1093/bib/bbp073

16. Kabiri C.M., Heywood P., Pennisi M. et al. (2019) Parallelisation strategies for agent based simulation of immune systems. *BMC Bioinformatics*, vol. 20, Article number: 579. https://doi.org/10.1186/s12859-019-3181-y

17. Kabiri Chimeh M., Richmond P. (2018) Simulating heterogeneous behaviours in complex systems on GPUs. *Simulation Modelling Practice and Theory*, vol. 83, pp. 3–17. https://doi.org/10.1016/j.simpat.2018.02.002

18. Kiran M., Richmond P., Holcombe M., Chin L., Worth D., Greenough C. (2010) FLAME: Simulating large populations of agents on parallel hardware architectures. *In Proceedings of Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '10)*, pp. 1633–1636.

19. Vouzis P.D., Sahinidis N.V. (2011) GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics*, vol. 27, no. 2, pp. 182–188. https://doi.org/10.1093/bioinformatics/btq644

20. Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. (1953) Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092. https://doi.org/10.1063/1.1699114

21. Hastings W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, vol. 57, no. 1, pp. 97–109. https://doi.org/10.2307/2334940

22. Fox W.P., Burks R. (2019) Monte Carlo simulation and agent-based modeling (ABM) in military decision-making. In: *Applications of Operations Research and Management Science for Military Decision Making. International Series in Operations Research & Management Science*, vol. 283. Springer, Cham. https://doi.org/10.1007/978-3-030-20569-0_8

23. Nianqiao J., Heng J., Jacob P. (2021) *Sequential Monte Carlo algorithms for agent-based models of disease transmission*. arXiv:2101.12156. https://doi.org/10.48550/arXiv.2101.12156

24. Mcreynolds T., Blythe D. (2005) Advanced graphics programming using OpenGL. In: *The Morgan Kaufmann series in computer graphics*, *Morgan Kaufmann*. https://doi.org/10.1016/B978-1-55860-659-3.X5000-8

25. Márquez C., César E., Sorribes J. (2013) Agent migration in HPC systems using FLAME. In: *Euro-Par 2013: Parallel Processing Workshops. Lecture Notes in Computer Science*, vol 8374, Springer, Berlin, Heidelberg, pp. 523–532. https://doi.org/10.1007/978-3-642-54420-0_51

26. Richey M.K. (2020) *Scalable agent-based modeling of forced migration*. Diss. George Mason University.

27. Kehoe J. (2015) *The Specification of Sugarscape*. arXiv:1505.060122015. https://doi.org/10.48550/arXiv.1505.06012

About the authors

Valery L. Makarov

Dr. Sci. (Phys.-Math.); Academician of Russian Academy of Sciences;

Academic Supervisor, Central Economics and Mathematics Institute, Russian Academy of Sciences, 47, Nakhimovsky Prospect, Moscow 117418, Russia;

E-mail: makarov@cemi.rssi.ru

ORCID: 0000-0002-2802-2100

Albert R. Bakhtizin

Dr. Sci. (Econ.); Corresponding Member of Russian Academy of Sciences;

Director, Central Economics and Mathematics Institute, Russian Academy of Sciences, 47, Nakhimovsky Prospect, Moscow 117418, Russia;

E-mail: albert@cemi.rssi.ru

ORCID: 0000-0002-9649-0168

Gayane L. Beklaryan

Cand. Sci. (Econ.);

Senior Researcher, Laboratory of Computer Modeling of Social and Economic Processes, Central Economics and Mathematics Institute, Russian Academy of Sciences, 47, Nakhimovsky Prospect, Moscow 117418, Russia;

E-mail: glbeklaryan@gmail.com

ORCID: 0000-0002-1286-0345

Andranik S. Akopov

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;

Chief Researcher, Laboratory of Dynamic Models of Economy and Optimization, Central Economics and Mathematics Institute, Russian Academy of Sciences, 47, Nachimovky Prospect, Moscow 117418, Russia;

E-mail: aakopov@hse.ru

ORCID: 0000-0003-0627-3037

Nikita V. Strelkovskii

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

Research Scholar, International Institute for Applied Systems Analysis, Laxenburg, Austria;

E-mail: strelkon@iiasa.ac.at

ORCID: 0000-0001-6862-1768

DOI: 10.17323/2587-814X.2022.1.22.41

Analysing the firm failure process using Bayesian networks

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

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

Abstract

This work analyses the firm failure process stages using the Bayesian network as a modelling tool because it allows us to identify causal relationships in the firm profile. We use publicly available data on French, Italian and Russian firms containing five samples corresponding to periods from one to five years before observation. Our results confirm that there is a difference between the stages of the failure process. For firms at the beginning of a lengthy process (3–5 years before observation), cumulative profitability is the key that determines liquidity. Then, as the process develops, leverage comes to the fore in the medium term (1–2 years before observation) for economies with more uncertainty. This factor limits the opportunities for making a profit, leading to further development of the failure. There are also national specifics that are caused, firstly, by the level of economic development and, secondly, economic policy uncertainty.

Keywords: financial failure, firm failure process, Bayesian network

Citation: Zelenkov Y. (2022) Analysing the firm failure process using Bayesian networks. *Business Informatics*, vol. 16, no. 1, pp. 22–41. DOI: 10.17323/2587-814X.2022.1.22.41

Introduction

The study of firm failure is one of the key issues in business research. It can be divided into two subdomains [1]: the first is failure prediction, and the second is theoretical and empirical investigations of the failure process. The firm failure process allows us to consider the behaviour of failing firms in the longer perspective [2, 3], while failure prediction studies often focus on financial performance only one or few years before distress [4, 5].

However, the short-term forecasting models' main weakness is that the firms' obligations often are longer than the period during which the risk of default is estimated with perfect accuracy [6]. Thus, the company should be analysed from a longer perspective. Moreover, while some firms with a certain financial performance profile fail, others with the same profile can overcome the difficulties and return to normal operations. Therefore, many authors argue the existence of different types of failure trajectories that may lead or may not lead a firm to default depending on its prehistory and current abilities [1, 7–9].

The existence of different firm failure processes (FFP) is a well-established fact supported by much theoretical and empirical research. However, there is no consensus in the scientific community not only about the exact definition of these processes but even about the number of variants. For example, Argenti [10] detected three failure trajectories of decline in firms' financial health; Ooghe and de Prijcker [8] describe four different types of failure processes; du Jardin [11] identifies seven types of FFP.

These differences are explained, firstly, by the methodology used, for example, works [8, 10] are based on the case method, and du Jardin [11] analyses empirical data using self-organising maps. Second, the authors view the process from different angles. Papers [10, 11] focus on financial results (this approach is also used in many other works [1, 2, etc.]), while [8] considers the problem through the lens of management efficiency.

In this work, we focus on defining the specifics of the various stages of a firm's failure process. Three research questions correspond to this goal:

- ♦ *RQ*1: How do the causal relationships between the financial ratios describing the firm's state change in different periods before the default?
- ♦ *RQ*2: Are there differences in firm failure processes that are determined by country specifics?
- ♦ *RQ*3: How does the degree of economic policy uncertainty affect the firm failure process?

We use the Bayesian Network as a modelling tool because it allows us to identify causal relationships in the firm profile. We use publicly available data on French, Italian, and Russian companies which present the firms' financial ratios from one to five years before the failure.

Our results confirm that there is a difference between the stages of the failure process (*RQ*1). For firms at the beginning of a lengthy process (3–5 years before observation), cumulative profitability is the key that determines liquidity. Then, as the process develops, leverage comes to the fore in the medium term (1–2 years before observation) for economies with more uncertainty. This factor limits the opportunities for making a profit, leading to further development of the failure. There are also national specifics that are caused, firstly, by the level of economic development (*RQ*2) and, secondly, by economic policy uncertainty (*RQ*3).

The rest of the paper is organised as follows. After reviewing sources analysing FFP, we present basic concepts of Bayesian networks. Next, we describe the datasets and pre-processing operations that are necessary to prepare the data for modelling. In the last part, we analyse the network structures obtained and discuss further research to extend the proposed approach.

1. Literature review

The first research into bankruptcy forecasting began in the 1930s [12]. These studies mainly focused on comparing individual rates of successful and unsuccessful firms. However, the number of published works was relatively small. The first multivariate model was presented by Altman [13], who used discriminant analysis based on five financial ratios. This model, also known as the Z-score, ushered in an era of intense research. Researchers have developed many predictive models using both statistical techniques and machine learning. It should be noted that models based on machine learning, in general, provide more excellent performance [14, 15]; however, the Z-score model with some modifications also remains relevant [4].

A common feature of predictive models is that they treat the failure prediction problem as a binary classification task. In this case, most often, the data of financial statements for a small number of periods before default is considered. In fact, these models are constructed on cross-sectional data; ratios from different time periods are often combined in one observation point; thus, the firm's individual dynamics are not taken in the account. Such an approach ignores the fact that companies change over time, all of which causes various problems and limitations [16]. In short, a firm's profile measured at time t cannot be reduced to measurements at time $t - 1$ alone, since the default, in most cases, is the result of a long process [9] and the discriminating power of ratios is unstable over time [17].

There are models based on observations of one, two or more years before failure at time t which are believed to be able to predict the state at years $t + 1$, $t + 2$, or even $t + 10$ [2, 18]. However, because they do not treat firm failure as a process, they have the same limitations as analysed in [16].

Some predictive studies use techniques that allow us to consider both the dynamics of firms' populations and their unique characteristics, such as panel regression [19] or survival analysis [20]. However, in general, the number of such works in the flow of research on predictive models is relatively small.

1.1. Firm failure process

Argenti [10] was perhaps the first who started to study the firms' failure process. He identified three patterns of decline and found that failing firms do not crash immediately after they decline. Some can delay the onset of bankruptcy for years.

D'Aveni [7] empirically tested Argenti's findings [10] that are based on case studies. According to both authors, the three failure processes are the following:

- Sudden decline, that is, the rapid collapse of the firm. This process of failure is typical of small or competitively disadvantaged firms that reoriented their strategy too boldly.
- Gradual decline, i.e., a slow and gradual process typical of bureaucratic and poorly managed firms that cannot adapt to the external environment.
- Lingering decline. This process is typical of firms that decline either rapidly or gradually but delay bankruptcy for several years. Such post-decline firms often centralise to be a threat rigid and exhibit strategic paralysis and downsizing activity.

Based on these earlier studies of failure processes, [1] postulate the existence of three types of failure process:

- short-run process when potential failure can be detected only about 1 year after the last reporting;
- mid-run process corresponds to a situation where the first signals of a potential failure can be detected 2–3 years before the default;
- long-run process when the potential failure can be detected more than 3 years before the default.

Short-term processes are more suitable for describing the situation when a firm with good performance declines suddenly. Mid-term and long-term processes, in turn, can describe two situations: the firm never becoming successful enough or the firm becomes worse step by step.

To empirically check their proposition, the authors of [1] analysed 1234 bankrupt manufacturing SME's from different European countries. They applied four clustering methods on the eight different sets of variables (presented in [4]) over the last five years before the bankruptcy. Their results confirm the existence of three types of failure processes, differing in time scale and, therefore, in decline rate.

Summarizing their results, the following can be noted. For short-run processes, the failure risk (FR) is observable only in year $t-1$ and negative annual profitability is the most important contributor. For mid-run processes, the failure risk can be detected in years $t-2$ and $t-3$, when negative annual profitability is also the most important contributor, followed by high leverage. For firms involved in a long-run process, the first signals can be detected up to $t-3$; they are annual and cumulative profitability and leverage. These ratios contribute to FR also in years $t-2$ and $t-1$. Liquidity as an indicator of FR is important only for the last stages of mid-term and long-term processes.

However, this view of failure processes is not the only one. For example, authors of [8] present four different processes explained through the lens of management: (1) unsuccessful start-up due to the lack of managerial or industry-related experience, (2) ambiguous growth of firms with over-optimistic management, (3) unbalanced growth induced by management's dazzle, and (4) an apathetic mature firm managed by people lacking motivation and commitment.

In a series of works [6, 9, 11] du Jardin models various failure processes (also called trajectories or profiles) using self-organizing maps (SOM). The basic idea of the application of SOM to study individual trajectories of firms is straightforward. Let us have panel data (observations of objects corresponding to measurements made in different time periods). If all the observations are classified on a SOM as if they were independent, it is possible to study the change of state of a given object along time [21]. To the best of our knowledge, this approach was first used in [22] to analyse the financial state of Spanish banks. The author noted that the trained SOM model groups the entities together according to their financial state similarities. Thus, new observations will be placed in a particular zone (bankrupts or non-bankrupts) according to the more activated neurons on the map. Therefore, it is possible to observe the bank's evolution using financial information from various years.

Briefly, du Jardin's methodology can be described as follows: at the first step, all firms are mapped to SOM, and the firm's observations at different points in time are considered independent. Then the trajectories of firms represented by the list of neurons that correspond to observations of one entity at sequential time points are built. In the last step, the trajectories are grouped into meta-classes, which can be viewed as processes leading or not leading to default. The author's results confirm that the generalisation error achieved with an SOM remains more stable over time than that achieved with conventional failure models (discriminant analysis, logistic regression, Cox's survival model, neural networks and ensemble methods). However, more importantly in the context of our discussion, du Jardin identifies a different number of processes that show the movement of firms between regions with different probability of failure: six in [9], seven in [11], and eight in [6]. This difference can be explained by the impact of the data used (specific years and time lag before default) and the impact of the technique of trajectories' grouping. This may not be significant in terms of the model's performance since these works' primary goal is to improve prediction accuracy by considering the firm's prehistory. However, even if this approach improves the predictive capabilities, it does not allow us to analyse the failure processes, since it is based on the black-box model.

So, we can conclude that the existence of different firm failure processes is a well-established fact; however, there is no consensus regarding these processes' definition. To contribute to the solution of this problem we propose to use a causality modelling technique, namely Bayesian networks. Such an approach allows us to identify causalities between financial ratios at different periods before failure, which can shed light on the firm's dynamics.

1.2. Bayesian networks and causality modelling

Intuitively, causality can be defined as influence through which a cause contributes to the production of an effect, while the cause is partially responsible for the effect, and the effect is partially dependent on the cause [23]. Complex systems are characterised by the presence of multiple interrelated aspects, many of which relate to the reasoning task. Thus, one of biggest challenges is the extraction of causal relationships from empirical data and construction of models of complex systems that allow causal inference.

The declarative representation approach [24] is based on a causal model of the system about which we would like to reason. This model encodes our knowledge of how the system works and can be manipulated by various algorithms that can answer questions based on the model.

To communicate causal relationships, a causal model uses a combination of equations and graphs. Mathematical equations that express the form of causality (e.g., linear, or non-linear) are symmetrical objects, so relationships of variables can be inverted using simple manipulations. For this reason, equations are augmented with a diagram that declares the directions of causality [25]. Such a model can be built manually based on expert knowledge or automatically using machine learning algorithms [26].

According to [27], causal inference extends predictive modelling (which involves estimating the conditional distribution $p(\mathbf{Y}|\mathbf{X})$ of the variables $\mathbf{Y}$ and $\mathbf{X}$ on the basis of a random sample) to causal modelling, where the model should be able to estimate the conditional distribution $p(\mathbf{Y}|\mathbf{X}||\mathbf{M})$ when manipulated $\mathbf{M}$.

There are several approaches to the construction of causal models, in particular structural equation modelling (SEM) and Bayesian networks (BN). The SEM [28] is limited, first, in that it requires a priori hypotheses about causality in the system. Secondly, it supposes only linear types of relationships. Therefore, in our study, we will use Bayesian networks, since they are free of such disadvantages. The structure of the network and its parameters can be extracted from data; relationships between variables are probabilistic.

A Bayesian network encodes the joint probability distribution $\mathbb{P}(\mathbf{X})$ of a set of m random categorical variables, $\mathbf{X} = (X_1, ..., X_m)$, as a directed acyclic graph (DAG) and a set of conditional probability tables (CPT). More formally, it is a pair $\langle \mathcal{G}, \mathcal{P} \rangle$, where $\mathcal{G}$ is the DAG whose vertices correspond to the variables in $\mathbf{X}$ and arcs represent direct dependencies between variables, and $\mathcal{P}$ is a collection of functions that define the behaviour of each variable in $\mathbf{X}$ given its parents in the graph [27,29].

The representation of the full joint table $\mathbb{P}(\mathbf{X})$ takes exponential space in the number of variables m. This complexity is avoided thanks to the Markov condition, which states that in a

Bayesian network every variable is conditionally independent of its non-descendant and non-parents. Thus, for the set of random variables $\mathbf{X}$ in $\mathcal{G}$, a density $\mathbb{P}(\mathbf{X})$ is

$$\mathbb{P}(\mathbf{X}) = \mathbb{P}(X_1, ..., X_m) = \prod_{i=1}^{m} \mathbb{P}(X_i \,|\, parents(X_i)),$$

where $parents\,(X_i)$ denotes the set of variables $X_j \subset \mathbf{X}$, such that there is an arc from node j to node i in the graph.

In other words, each node in the graph $\mathcal{G}$ that corresponds to a variable has an associated CPT that contains the probability of each state of the variable given its parents in the graph. Such a presentation allows us to describe the structure of complex distribution compactly [30] and can be interpreted from two points of view [24]. First, the graph is a compact representation of a set of independencies that hold in the distribution. The other perspective is that the graph is a skeleton for factorising distribution: it breaks up the distribution into smaller factors, each over a much smaller space of possibilities.

Bayesian networks have many advantages [24]. First, this type of presentation is interpretable by a human. Second, such a structure allows us to answer queries, i.e., computing the probability of some variables given evidence of others (inference). Third, models can be constructed whether by a human expert or automatically by learning from data. In our study, we will use the latter approach – data driven learning.

From a formal perspective, the Bayesian network represents the underlying joint distribution, including probabilistic properties such as conditional independence. On the one hand, it is a more compact representation of complex multivariate distributions. On the other hand, a "good" network structure should correspond to causality, in that an edge $X \rightarrow Y$ often suggests that X "causes" Y, since each value x of X specifies a distribution over the values of Y [24, 25].

The process of construction of Bayesian networks from data $\mathcal{D}$ includes two stages: generation of the graph representing the optimal structure of BN (structure learning), and definition of conditional probabilities (parameter learning). Many authors apply BN to problems in different domains: for example, business [31], ecology [32], healthcare [33], fault diagnosis in engineering systems [34] and many others [35].

It should also be noted that there is an extension of the BN model for longitudinal data, namely, dynamic Bayesian networks. However, dynamic models are based on the assumption that the process under study is stationary, i.e., its parameters do not change over time. According to [1] and other researchers, the firm failure process is not stationary. Therefore, in our study, we generate BN structures for different time intervals independently, supposing that comparing these structures will shed light on the peculiarities of FFP stages.

1.3. Inference and explanation in Bayesian networks

Inference is the process of computing new probabilistic information from a Bayesian network based on some evidence. It computes joint posterior probabilities for a set of variables given evidence which are the values on other variables. Inference is an NP-complete task, therefore there are algorithms that implement an exact inference but also algorithms for approximated inference that can converge slowly and even not exactly but that can in many cases be useful for applications. This capability allows us to use BN for supervised classification which aims at assigning labels to instances described by a set of predictor variables [36].

However, unlike many machine learning methods, a Bayesian network can be used not only for prediction but also for explanation [37]. Explanation tasks in Bayesian networks can be classified into three categories [38]:

- explanation of a model – presentation of the domain knowledge;
- explanation of reasoning – presentation of the results inferred and reasoning process that produced them;
- explanation of evidence, i.e., determination which values of the unobserved variables justify the available evidence.

Since our goal is to analyse the BN structures that model firms at different times before failure, explaining the model is the most important issue.

In [39] the authors give examples of model explanations. These explanations can include properties of nodes and their mutual influence that can be negative or positive. The influence of node A on node B is positive when higher values of A make high values of B more probable. The definitions of negative influence and negative link are analogous.

2. Data

Relevant data is needed to build causal models for firms that filed for default years after the measurement. In addition, an interesting question is the comparison of the FFP for different countries. Therefore, we chose three countries for analysis: France, Italy, and the Russian Federation.

To compare the economies on a macro-level, we use the Gross Domestic Product converted to constant 2017 international dollars using Purchasing Power Parity (PPP) rates and divided to the total population[1].

Figure 1 shows the change of this indicator for the selected time interval (2009–2019). France has the most stable economy; it exhibits constant GDP per capita growth during the whole period under study. The Italian economy is more volatile; after 2010, there was a recession, and growth resumed only in 2015. The Italian economy is driven in large part by small and medium-sized enterprises, many of them family-owned. Italy also has a sizable underground economy, which is estimated as much as 17% of GDP[2]. Based on these data, we can expect that the FFP model for French firms will be more stable than for Italian ones as they operate in a more stable environment. Note that according to the World Bank classification, both countries belong to the group of developed countries.

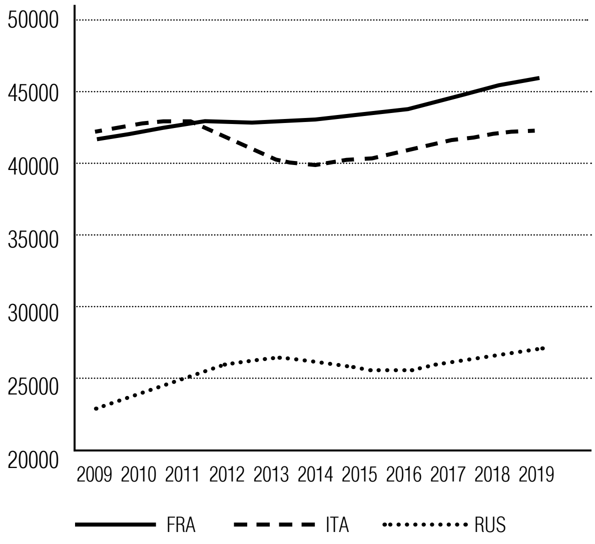

Fig. 1. Changes of GDP per capita (constant 2017 international dollars using PPP) for selected countries.

The Russian Federation belongs to the group of developing countries or economies in transition. The Russian economy is characterized by the significant share of the government-controlled sector and is largely regulated not by the market but by political decisions. A combination of falling oil prices, international sanctions, and structural limitations pushed Russia

[1] https://data.workdbank.org

[2] CIA (2021). The World Factbook. https://www.cia.gov/the-world-factbook/

into a deep recession in 2015, but GDP decline was reversed in 2017 as world oil demand picked up. All this leads to the highest growth of uncertainty. Under such conditions, it can be expected that the FFP model for Russian firms will change quite strongly at different stages.

We have collected the necessary data from the Bureau van Dijk Amadeus database[3] using the following search strategy:

- ♦ Data for companies that operate in 2009–2019. We excluded 2020 data to avoid the impact of external shocks related to the COVID-19 pandemic.
- ♦ The company belongs to a small and medium sized business (SMB) – the number of employees in the last available year is limited by values min = 10, max = 250.
- ♦ Good companies are companies which have Active status in the last available year
- ♦ Failed companies are companies that have one of the statuses: Active (default of payment), Active (insolvency proceedings), Bankruptcy, Dissolved (liquidation), and Dissolved.

For each country under consideration, we get five samples corresponding to year $t - n$, $n = 1, \ldots, 5$ before observation in year t. Each observation has a class label that indicates the firm's state at the end of the forecast period t: failure (*Class* = 1) or non-failure (*Class* = 0). The number of observations (samples) for each time period is presented in *Table 1*; the number of failed firms is indicated in brackets. As you can see, all datasets are unbalanced. The values of the imbalance (IB) ratio computed as the ratio of negative class observations to the number of failed companies are also given in *Table 1.*

2.1. Features selection

Since our goal is to build interpretable causal models, we must reduce the number of variables in the original dataset, leaving only those that provide the optimal balance of simplicity and completeness. Therefore, we will follow [1] approach, who used four variables that included the famous Altman's Z''-score model when analysing the bankruptcy process.

In a paper presenting the initial Z-score model, Altman [13] compiled a list of 22 potentially important financial ratios, classified into five standard categories: liquidity, profitability,

Table 1.

Dataset characteristics

		***t* - 5**	***t* - 4**	***t* - 3**	***t* - 2**	***t* - 1**
France	Samples	48024 (1509)	47163 (1503)	44151 (1439)	41798 (1382)	39720 (1313)
	IB ratio	30.825	30.379	29.682	29.245	29.251
Italy	Samples	55895 (5223)	56036 (5349)	56170 (5522)	56115 (5535)	55728 (5498)
	IB ratio	9.702	9.476	9.172	9.138	9.136
Russian Federation	Samples	44354 (1941)	43859 (2077)	43153 (2167)	43050 (2337)	42931 (2398)
	IB ratio	21.851	20.117	18.914	17.421	16.903

[3] Bureau van Dijk. Amadeus. https://amadeus.bvdinfo.com

Table 2.

Financial ratios in Altman's Z-score model

Category	Financial ratio	Definition	Comments
Liquidity	WCTA	Working Capital / Total Assets	Working capital is defined as the difference between current assets and current liabilities, so this ratio is a measure of the net liquid assets of the firm relative to the total capitalisation [13]. The liquidity role is based on legal considerations, as the inability to pay the outstanding debt is a sufficient precondition for starting an official bankruptcy process [1].
Cumulative profitability	RETA	Retained Earnings / Total Assets	It is the measure of cumulative profitability over time which implicitly includes the age of a firm [13].
Annual profitability	EBITTA	Earnings before Interest and Taxes / Total Assets	It is a measure of the true productivity of the firm's assets, abstracting from any tax or leverage factors[13].
Leverage	BVETL	Book Value of Equity / Book Value of Total Liabilities	In the initial Z–score model, the Market Value of Equity was used but this approach is applicable only to publicly traded companies (Altman et al., 2017). This ratio measures the firm's ability to service liabilities using its own equity because additional debt, all other things being equal, increases bankruptcy likelihood [1].
Activity	STA (excluded)	Sales / Total Assets	Excluded from the revised Z"–score model because it is an industry–sensitive variable [4].

leverage, solvency, and activity. Only five financial ratios were included in the final discriminant function (*Table 2*). Note that higher values of all selected ratios correspond to a lower likelihood of bankruptcy.

Later, the author noted that the original model is applicable only to publicly traded companies since it includes the firm's market value [40]. For this reason, in the new version of the model, he substituted the market value of equity by book value (Z'-score model). The next significant improvement was the exclusion of Sales / Total assets ratio because it is an industry-sensitive variable (Z''-score model).

Thus, we will use four ratios (WCTA, RETA, EBITTA, and BVETL) included in Altman's Z''-score. One of the BN modelling preconditions is that there must be no latent variables (unobserved variables influencing the network's variables) acting as confounding factors. Based on the time-tested Altman model, we can confidently believe that this condition is met and there are no latent factors in the empirical data.

In [4], the authors tested the performance of the Z''-score model using a huge international dataset (more than 2.6 million observations of firms from 31 countries in the training sample). Overall, their results confirm that the model performs well despite its simplicity.

However, the authors made several important clarifications:

- the coefficients of the model must be reevaluated for each sample,
- the model based on logistic regression gives better results than the multiple discriminant analysis version.

Thus, we will use logistic regression as the basis for validating subsequent data transformations. *Table 3* presents the ROC AUC scores obtained using the logistic regression (LR) 10-fold cross-validation procedure for data that contains the above four Altman features (see the 'Raw data' line for each country). Note, the quality of prediction decreases as the interval between observation and evaluation increases because the classification approach ignores changes in the firm over time [16].

2.2. Discretisation

Another problem stems from the fact that the concept of a non-linear Bayesian network was developed to handle discrete or categorical data. There are three common approaches to extending the Bayesian network to continuous variables [41]. The first is to model the conditional probability density of continuous variables using parametric distributions, and then to redesign the BN learning algorithms based on the parameterisations [42]. The second approach is to use nonparametric distributions, such as Gaussian processes [43]. The third approach is discretisation, that is a process that transforms a variable, either discrete or continuous, into a finite number of intervals and associates with each interval a numerical, discrete value [44, 45].

Table 3.

ROC AUC scores (10-fold cross-validation)

Data	Model	Datasets				
		t – 5	*t* – 4	*t* – 3	*t* – 2	*t* – 1
France						
Raw data	LR	0.675(0.034)	0.686(0.034)	0.694(0.026)	0.705(0.024)	0.714(0.028)
Discretized data	LR	0.687(0.030)	0.696(0.029)	0.712(0.022)	0.723(0.019)	0.729(0.027)
	BN	0.686(0.028)	0.697(0.028)	0.712(0.016)	0.726(0.015)	0.727(0.032)
Italy						
Raw data	LR	0.695(0.039)	0.727(0.031)	0.744(0.026)	0.768(0.017)	0.802(0.012)
Discretized data	LR	0.716(0.027)	0.761(0.019)	0.777(0.015)	0.798(0.013)	0.826(0.009)
	BN	0.717(0.028)	0.757(0.019)	0.776(0.013)	0.809(0.021)	0.833(0.015)
Russian Federation						
Raw data	LR	0.658(0.023)	0.675(0.021)	0.691(0.014)	0.711(0.014)	0.742(0.010)
Discretized data	LR	0.676(0.024)	0.683(0.018)	0.695(0.011)	0.731(0.021)	0.757(0.020)
	BN	0.709(0.034)	0.740(0.038)	0.738(0.034)	0.743(0.030)	0.770(0.037)

The discretisation approach in the context of Bayesian networks can be divided into two parts. First, there are algorithms that discretise attributes based on interdependencies between class labels and attribute values, such as the entropy binning method [46]. These algorithms are based on classification problems. They are used to discretise all continuous variables before learning the Bayesian network structure. The next class of algorithms requires that the structure of the network be known in advance [41, 45, 47]. These algorithms start with some preliminary discretisation policy, then the structure learning algorithm is started to determine the locally optimal graph structure. The discretisation policy is then updated based on the learned network, and this cycle is repeated until convergence.

We carried out a series of experiments and found out that preliminary discretisation based on [46] allows us to learn networks with higher score. *Table 3* shows the logistic regression ROC AUC scores obtained on discretised data, which confirm the chosen approach to discretisation improves the performance of the model.

For reference, *Figs. 2* and *3* shows the distribution of raw and discretised data respectively of the Italy $t-1$ dataset. Note that the average values of the transformed ratios have changed because we now use the identification of intervals into which each variable is divided instead of the absolute values. For BN, this transformation is acceptable because the model uses joint probabilities. The number of intervals according to [46] is determined based on the joint distribution of the attribute discretised and the target variable.

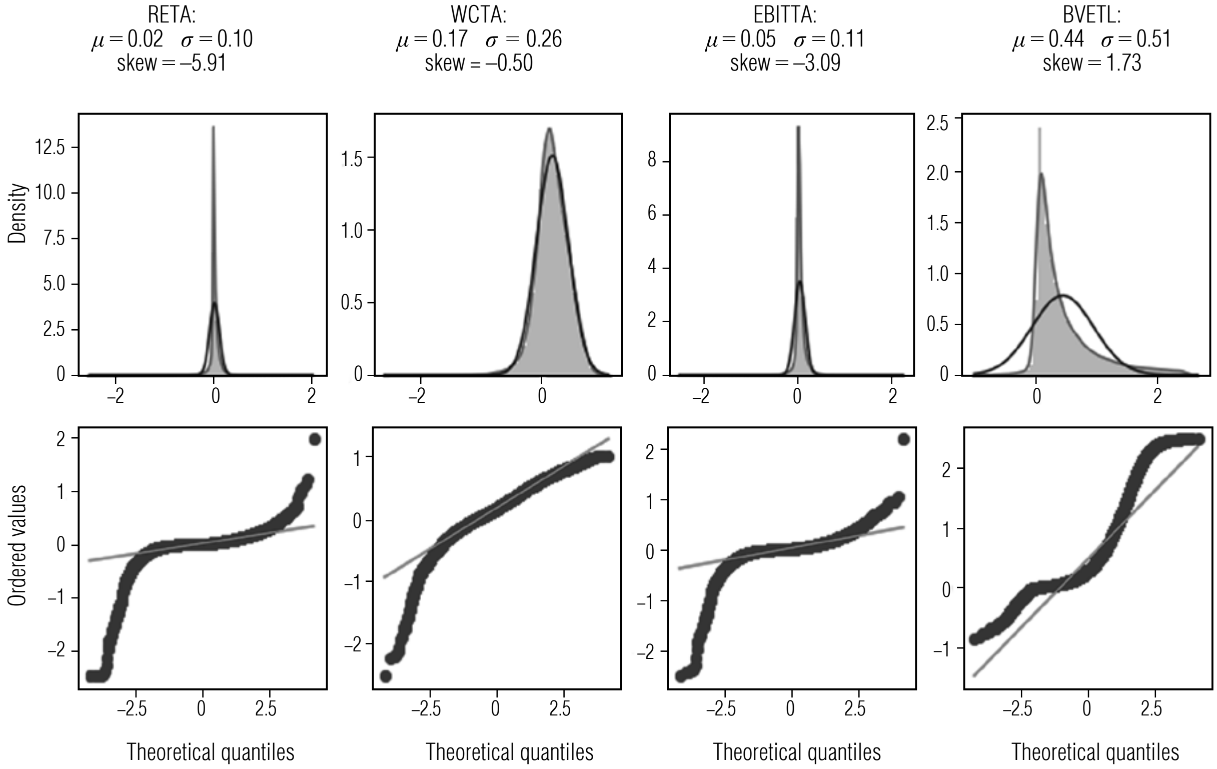

Fig. 2. Italy t – 1 dataset: distribution of continuous data.

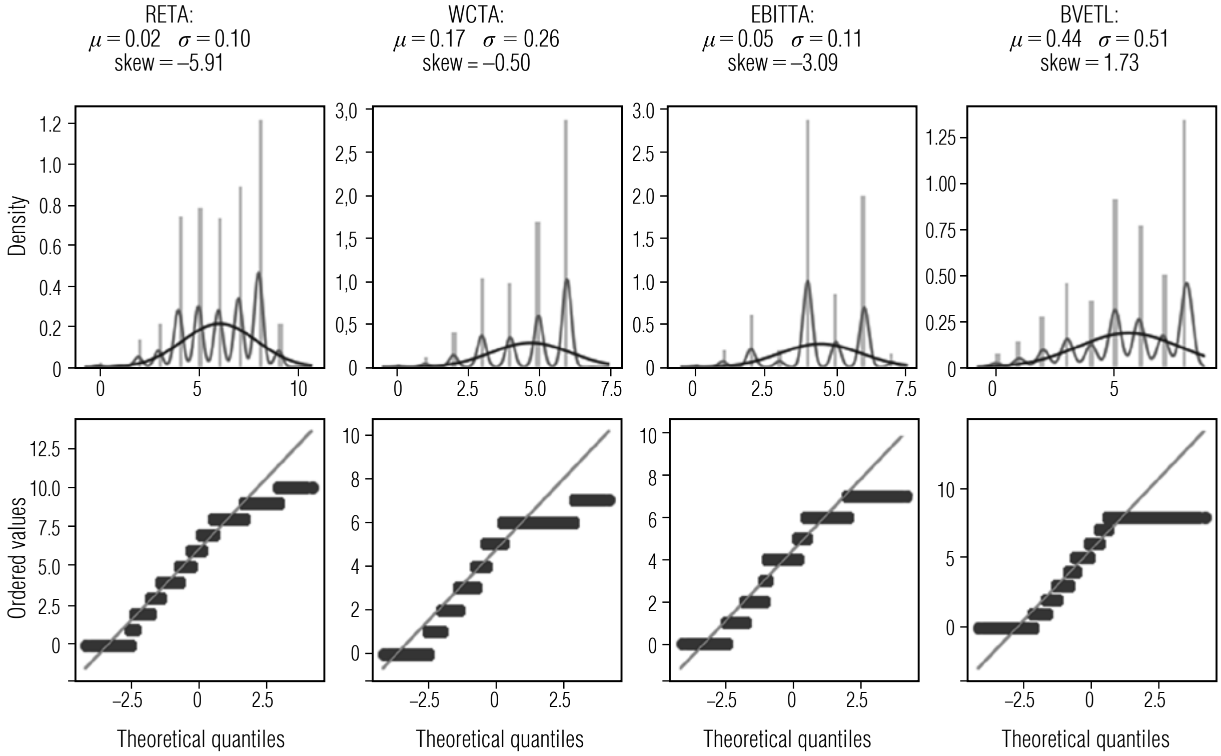

Fig. 3. Italy t – 1 dataset: distribution of discretised data.

3. Experiment and results

The process of construction of Bayesian networks from data $\mathcal{D}$ includes two stages: first, generation of the directed acyclic graph $\mathcal{G}$ representing the optimal structure of BN (structure learning), and next, definition of conditional probability tables $\mathcal{P}$ for each node in the graph (parameter learning). For our research, it is most important to study the structure of Bayesian networks corresponding to different periods before default. However, we performed both stages of learning, since CPT is important for the causal inference and, therefore, use of the model as a predictive tool.

Learning the structure of Bayesian networks can be complicated for two main reasons: (1) inferring causality and (2) the super-exponential number of directed edges that could exist in a dataset. Most methods for structure learning can be put into one of the following categories [24, 29]:

- score-based structure learning, with the goal to solve the optimisation problem

$$\underset{G \in \mathcal{G}}{\operatorname{argmax}}\ \mathit{score}\,(G, \mathcal{D}).$$

In other words, it is the task to find the best DAG according to some score function that measures its fitness to the data. Widely adopted scores are the Bayes Dirichlet equivalent uniform (BDeu), Bayesian Information Criterion (BIC), which approximates the BDeu, and Akaike Information Criterion (AIC).

- constraint-based structure learning family of algorithms that perform a series of statistical tests to find independences among the variables and build the DAG following these constraints.

According to [24], a score-based approach evaluates the complete network structure against the null hypothesis of the empty network. Thus, it takes a more global perspective, which allows us to trade off approximations in different part of the network. Therefore, we use score-based algorithms.

To evaluate the structure, we use the BIC score. Let us have a set of random variables $\mathcal{D}$. Let S be a candidate Bayesian network structure and Θ_S be a vector of parameters for S. Then

$$\mathrm{BIC} = \log P\left(\mathcal{D} | S, \Theta^*\right) - \frac{d}{2} \log N,$$

where Θ^* is the estimation of Θ_S;

d is the number of free parameters in S;

N is the dataset size.

The first term in the formula presents the logarithm of likelihood and the second one is the penalty for complexity.

BIC has two important properties that allow it to be used as a universal metric. Firstly, BIC is an equivalence invariant, i.e., it gives the same score to equivalent models. As the number of variables grows, the number of possible network structures also grows. This property of the BIC guarantees the assignment of the same score to equivalent networks. Secondly, BIC is locally consistent when the sample size is sufficiently large.

There are many different software packages and methods that they implement (e.g., see review in [29]). We use the pomegranate, that is an open-source Python library [48] implementing few score-based methods, in particular an exact algorithm A*[49], its greedy implementation and Chow-Liu [50] algorithm.

On the first step, we tested all the algorithms in the package to find the ones that give the best results on our data. According to the tests, the exact algorithm achieves the best performance. *Table 4* presents values obtained of BIC for the Bayesian Network. We also tested the Naive Bayes (NB) approach to ensure that the proba-

Table 4.

Bayesian Information Criterion (BIC) for the Bayesian Network (BN) and the Naive Bayes (NB) model

	***t* – 5**	***t* – 4**	***t* – 3**	***t* – 2**	***t* – 1**
			France		
NB	–201 163	–192 127	–180 853	–178 759	–159 080
BN	–171 651	–162 705	–152 862	–147 833	–129 372
			Italy		
NB	–360 293	–375 556	–380 321	–391 209	–394 780
BN	–316 636	–332 979	–324 984	–326 920	–329 246
			Russian Federation		
NB	–199 197	–249 535	–235 613	–240 508	–247 257
BN	–167 615	–214 696	–204 133	–206 716	–211 850

bilistic relationships between variables are beneficial. NB is the simplest form of the Bayesian network, derived from the assumption of mutual independence of exogenous variables. The results presented in *Table 4* confirm that the Naive Bayes method is inferior in accuracy to Bayesian networks. The corresponding BIC values are about 20% worse than those obtained for BN.

We also tested the performance of a classifier built based on this Bayesian network [36]. Given that the data is unbalanced, we used the decision threshold adjustment by introducing various penalties for misclassification errors [51]. So, for observation x, the predicted class label $\hat{y}(x) = 1$ if and only if $\mathbb{P}(x) \geq t$. Here the $\mathbb{P}(x)$ is an inference of BN when all variables except *Class* are known. Threshold t is computed as

$$t = C_{10}/(C_{10} + C_{01}),$$

where C_{ij} is a cost of predicting the class i when the true class is j. We set $C_{10} = 1$ and $C_{01} = IB$, where IB is the imbalance ratio of the training dataset.

Table 3 presents the ROC AUC scores obtained by 10-fold cross-validation; refer to the lines 'Discretised dataset / BN'. As we can see, the performance of the BN classifier at least comparable with the Logistic Regression for all datasets and outperforms it in most cases especially for uncertain economies (Russia and Italy).

4. Discussion

The network structures shown in *Figs. 4–6* allow us to draw some important conclusions about the features of different stages of the firm failure process. The *Class* variable that labels the firm state (0 for healthy firms and 1 for fail companies) is on the root of graphs. This can be easily interpreted as follows. The state of the firm is the root cause that determines the values of its financial ratios. This view is consistent with the problem of failure prediction when the state of the firm is computed by the values of financial ratios.

As follows from *Figs. 4–5*, for developed economies (Italy, France), the early stages of a long-term process (t – 5, t – 4) coincide. For the period t – 5, the cumulative profitability positively affects the difference between assets and liabilities, i.e., leverage (note, the numerator BVETL is the difference between Total Assets and Total Liabilities). Both factors then determine the firm's current liquidity and current profitability. Note that liquidity and annual profitability are independent. However, at stage t – 4, annual profitability becomes a factor affecting liquidity.

For a more predictable economy (France), the model will not change during the t – 4, t – 3 and t – 2 periods. One year before the financial failure, the network structure for France changes and becomes like the t – 5 period. Overall, we can conclude that cumulative profitability is a key factor in the success of French firms.

The model representing the short-term failure process of Italian firms (t – 2 and t – 1) is changing more radically. The key factor is the difference between assets and liabilities (leverage), which determines the firm's ability to generate profits and liquidity. Note also that the liquidity values are conditionally independent of the cumulative and annual profitability at these stages. Obviously, this is due to the higher uncertainty in the Italian economy. Firms unable to meet liabilities using their own assets cannot quickly remedy this situation by increasing productivity through borrowed resources.

For Russian companies, the key factors are leverage and cumulative profitability. Also note that in this case, the liquidity depends on all the variables under consideration (except the period t – 2). In general, the process can be described as follows. In the mid and long term

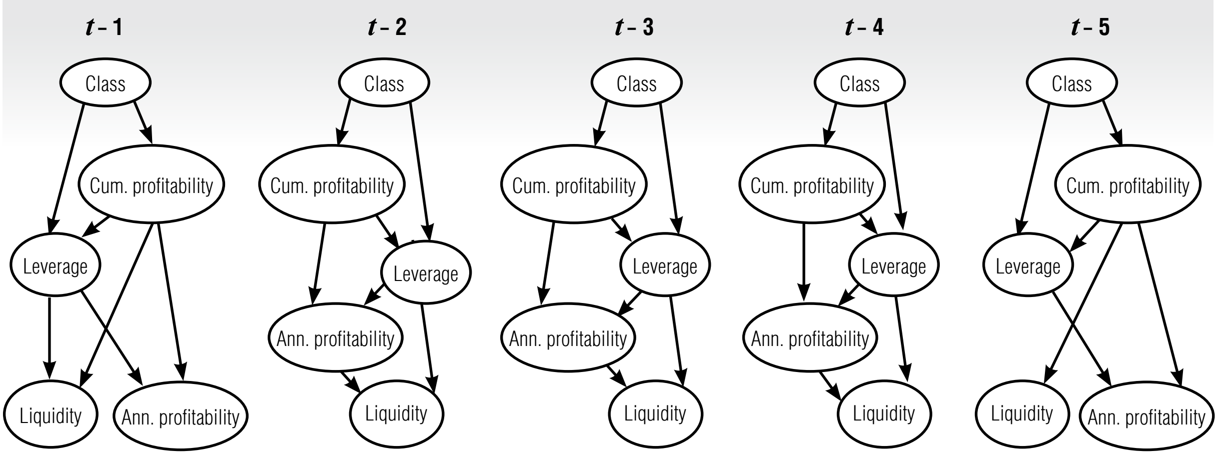

Fig. 4. France: Bayesian networks for different periods before default.

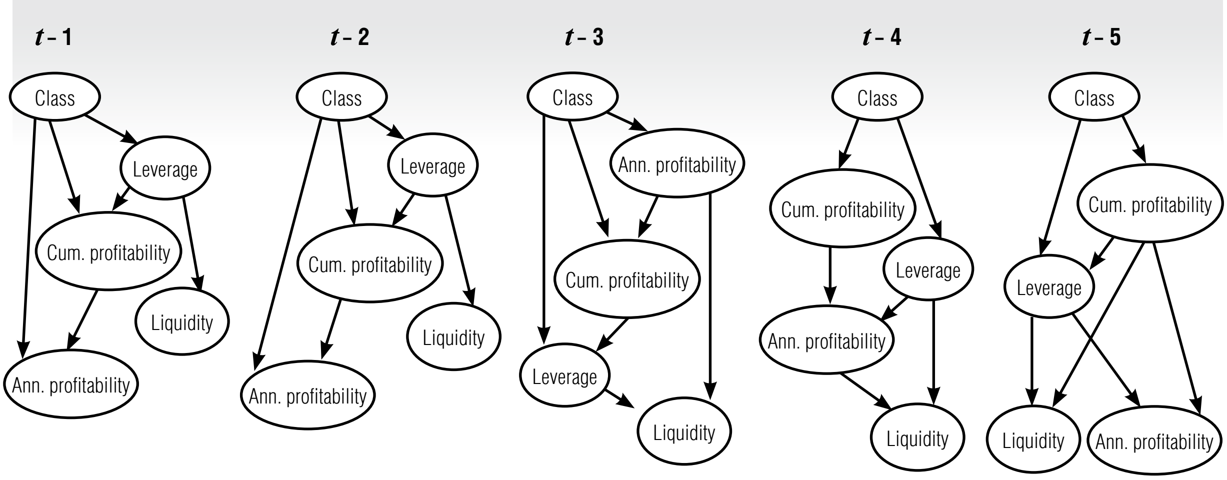

Fig. 5. Italy: Bayesian networks for different periods before default.

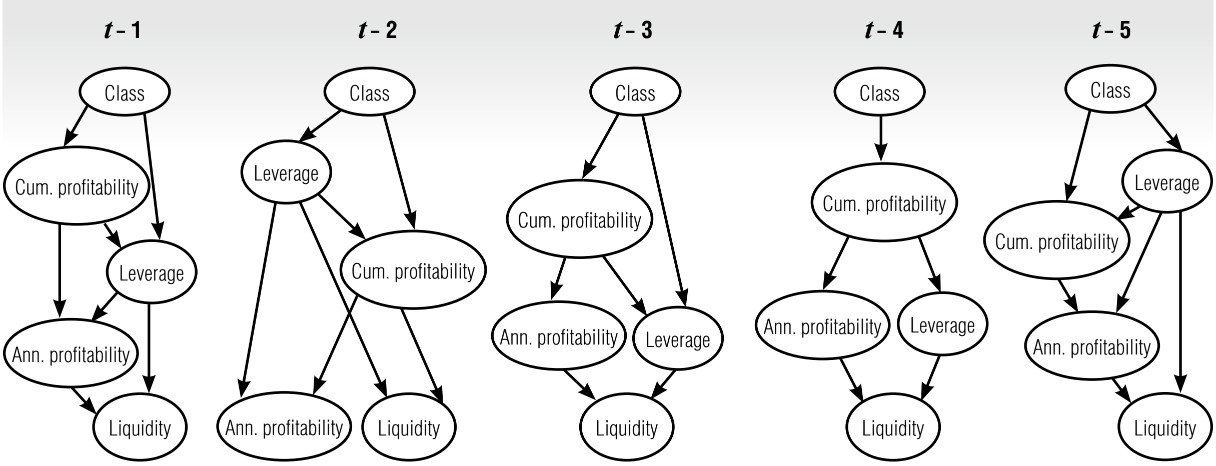

Fig 6. Russian Federation: Bayesian networks for different periods before default.

($t - 3$ and $t - 4$) cumulative profitability has a marginally positive effect on annual profitability. This can be explained by the fact that the RETA ratio implicitly reflects the firm's age [13] and its ability to generate profits sustainably. The accumulated profit also causes the number of external resources attracted. At the same time, annual profitability and leverage are conditionally independent; however, they completely determine liquidity. The conditional independence of the annual profitability and the volume of attracted resources can be explained by the fact that we are considering a fairly long process at these stages, the results of which will be evaluated in 3–4 years. Obviously, this process is more influenced by managerial decisions based on financial indicators that reflect long-term trends (cumulative profitability) than short-term results (annual profitability).

In stage $t - 2$, leverage becomes a key factor. This means that the ability to attract resources allows underperforming firms to increase profitability and increase liquidity and avoid financial disruptions in 2 years. In the year $t - 1$, cumulative profitability becomes a causal factor determining the leverage. This can be explained by the fact that potential lenders assess the firm's overall performance in the long term, which can limit the availability of borrowed resources. Annual profitability is caused by the ability to generate profit in the long term and service the debt. Leverage and annual profitability determine the current value of liquidity, which at this stage is the main indicator of potential financial failures.

The main conclusion drawn from the presented results is that the mutual influence of the factors that determine the state of the firm changes over time (*RQ*1). In general, for firms at the beginning of a lengthy process that could lead to failure, cumulative profitability is the key that determines other metrics such as liquidity and leverage. Then, as the process develops, in the medium term, the degree of self-sufficiency, as measured by leverage, come to the fore, especially for economies with higher uncertainty. In these stages, low values of these factors limit the opportunities for making a profit. This leads to further development of the failure.

However, there are national specifics that are caused, firstly, by the level of economic development (*RQ2*) and, secondly, economic policy uncertainty (*RQ*3). This specificity is manifested both in the change in the causal relationships between factors at different stages of the firm failure process and in the rate of change of models. The most robust set of models is obtained for France, which has the lowest uncertainty. For Russia, which is characterized by the maximum growth of economic uncertainty over the past 10 years, the models change most frequently and more radically.

Thus, the resulting graphs shed light on the specifics of the various stages of the failure process. As far as we know, our paper is the first attempt to analyze FFP based on Bayesian networks. However, our research in its current form has some issues that can be possibly viewed as limitations. In particular, we can note the following:

- The sample used contains cross-sectional data for different time periods before failure. It allows us to identify differences in causal relationships at stages of FFP; but it is impossible to trace the evolution of specific firms. To solve such a problem, panel data is needed. Analysis of data containing sequential periods for good and failed firms can provide more detailed information on the causality of a firm's decline. However, solving this problem requires another tool, which can be a Dynamic Bayesian Network.
- The analysed factors are limited to only four financial ratios presented in the Altman Z''-score. We accepted this limitation based on the requirements for simplicity of the model and its further interpretation. This made

it possible to draw important conclusions about the internal dynamics of a firm. However, in further research, the financial ratios list can be extended to get more complex and detailed models. It is also necessary to study the influence of other parameters, for example, corporate governance and environmental factors.

The next issue, which is of practical interest, is the definition of the current stage of the analysed firm's process. This information can be useful for predictive model which will compute the probability of default for a few future periods. This issue is also a topic of future research.

Conclusion

Our work's main goal was to demonstrate that Bayesian networks can serve as a reliable tool for analysing the dynamics of firms and studying the firm failure process. Our results, on the one hand, highlight the specifics of stages of the failure process for different economies. On the other hand, they allow us to build predictive models that surpass Altman's Z''-score using the same variables. As far as we know, the work presented is the first one using Bayesian networks for FFP analysis, so many issues remained outside our study's scope. Possible areas of research include:

- Building models on panel data describing the dynamics of a set of firms.
- Expansion of the number of analysed features.
- Modelling specifics of industries.
- Determining the stage of the process to predict failure in the long term.

All this opens a vast field for new studies, which, in the light of the results obtained, seem promising, since they can potentially make a significant contribution to the theoretical and empirical analysis of the firm failure process. ■

Acknowledgments

The work is a part of the project "Development of Quantitative Methods for Bankruptcy Prediction" supported by the Graduate School of Business of the HSE University.

References

1. Lukason O., Laitinen E.K. (2019) Firm failure processes and components of failure risk: An analysis of European bankrupt firms. *Journal of Business Research*, vol. 98, pp. 380–390. https://doi.org/10.1016/j.jbusres.2018.06.025

2. Altman E.I., Iwanicz-Drozdowska M., Laitinen E., Suvas A. (2020) A Race for Long Horizon Bankruptcy Prediction. *Applied Economics*, vol. 52, no. 37, pp. 4092–4111. https://doi.org/10.1080/00036846.2020.1730762

3. du Jardin P. (2021) Forecasting corporate failure using ensemble of self-organizing neural networks. *European Journal of Operational Research*, vol. 288, no. 3, pp. 869–886. https://doi.org/10.1016/j.ejor.2020.06.020

4. Altman E.I., Iwanicz-Drozdowska M., Laitinen E.K., Suvas A. (2017) Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-Score model. *Journal of International Financial Management & Accounting*, vol. 28, no. 2, pp. 131–171. https://doi.org/10.1111/jifm.12053

5. Zelenkov Y., Fedorova E., Chekrizov D. (2017) Two-step classification method based on genetic algorithm for bankruptcy forecasting. *Expert Systems with Applications*, vol. 88, pp. 393–401. https://doi.org/10.1016/j.eswa.2017.07.025

6. du Jardin P. (2017) Dynamics of firm financial evolution and bankruptcy prediction. *Expert Systems with Applications*, vol. 75, pp. 25–43. https://doi.org/10.1016/j.eswa.2017.01.016

7. D'Aveni R. (1989) The aftermath of organizational decline: A longitudinal study of the strategic and managerial characteristics of declining firms. *Academy of Management Journal*, vol. 32, no. 3, pp. 577–605. https://doi.org/10.5465/256435

8. Ooghe H., De Prijcker S. (2008) Failure processes and causes of company bankruptcy: A typology. *Management Decision*, vol. 46, no. 2, pp. 223–242. https://doi.org/10.1108/00251740810854131

9. du Jardin P., Séverin E. (2012) Forecasting financial failure using a Kohonen map: A comparative study to improve model stability over time. *European Journal of Operational Research*, vol. 221, no. 2, pp. 378–396. https://doi.org/10.1016/j.ejor.2012.04.006

10. Argenti J. (1976) *Corporate collapse: The causes and symptoms*. New York, NY: McGraw-Hill.

11. du Jardin P. (2015) Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research*, vol. 242, no. 1, pp. 276–303. https://doi.org/10.1016/j.ejor.2014.09.059

12. Bellovary J.L., Giacomino D.E., Akers M.D. (2007). A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial Education*, vol. 33, pp. 1–42.

13. Altman E.I. (1968) Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, vol. 23, pp. 589–609.

14. Barboza F., Kimura H., Altman E. (2017) Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, vol. 83, pp. 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

15. Zelenkov Y., Volodarskiy N. (2021) Bankruptcy prediction on the base of the unbalanced data using multi-objective selection of classifiers. *Expert Systems with Applications*, vol. 185, article ID 115559. https://doi.org/10.1016/j.eswa.2021.115559

16. Balcaen S., Ooghe H. (2006) 35 Years of Studies on Business Failure: An Overview of the Classic Statistical Methodologies and Their Related Problems. *The British Accounting Review*, vol. 38, pp. 63–93. https://doi.org/10.1016/j.bar.2005.09.001

17. Bardos M. (2007) What is at stake in the construction and use of credit scores? *Computational Economics*, vol. 29, no. 2, pp. 159–172. https://doi.org/10.1007/s10614-006-9083-x

18. Iwanicz-Drozdowska M., Laitinen E.K., Suvas A., Altman E.I. (2016) Financial and Nonfinancial Variables as Long-horizon Predictors of Bankruptcy. *Journal of Credit Risk*, vol. 12, no. 4, pp. 49–78. https://doi.org/10.21314/JCR.2016.216

19. Pizzi S., Caputo F., Venturelli A. (2020) Does it pay to be an honest entrepreneur? Addressing the relationship between sustainable development and bankruptcy risk. *Corporate Social Responsibility and Environmental Management*, vol. 27, no. 3, pp. 1478–1486. https://doi.org/10.1002/csr.1901

20. Zelenkov Y. (2020) Bankruptcy Prediction Using Survival Analysis Technique. In: *2020 IEEE 22nd Conference on Business Informatics (CBI)*, vol. 2, pp. 141–149. IEEE. https://doi.org/10.1109/CBI49978.2020.10071

21. Cottrell M. (2003) Some other applications of the SOM algorithm: how to use the Kohonen algorithm for forecasting. In: *Invited lecture at the 7th International Work-Conference on Artificial Neural Networks IWANN 2003.*

22. Serrano-Cinca C. (1998) Let financial data speak for themselves. In: Deboeck, G., Kohonen, T. (eds.) *Visual Explorations in Finance with Self-Organizing Maps*. Springer, pp. 3–23.

23. Bunge M. (2017) *Causality and modern science: Fourth revised edition*. Routledge, NY.

24. Koller D., Friedman N. (2009) *Probabilistic graphical models: Principles and techniques*. MIT Press, Cambridge: MA.

25. Pearl J. (2009) *Causality*. Cambridge University Press.

26. Zhao Q., Hastie T. (2021) Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics*, vol. 39, no. 1, pp. 272–281. https://doi.org/10.1080/07350015.2019.1624293

27. Spirtes P. (2010) Introduction to causal inference. *Journal of Machine Learning Research*, vol. 11, pp. 1643–1662.

28. Hair J.F., Hult G.T.M., Ringle C.M., Sarstedt M., Thiele K.O. (2017) Mirror, mirror on the wall: a comparative evaluation of composite-based structural equation modelling methods. *Journal of the Academy of Marketing Science*, vol. 45, no. 5, pp. 616–632. https://doi.org/10.1007/s11747-017-0517-x

29. Scanagatta M., Salmeron A., Stella F. (2019) A survey on Bayesian network structure learning from data. *Progress in Artificial Intelligence*, vol. 8, pp. 425–439. https://doi.org/10.1007/s13748-019-00194-y

30. Sucar L.E. (2021) *Probabilistic graphical models: Principles and applications*. Springer Nature. Cham, Switzerland.

31. Ekici A., Ekici S.O. (2021) Understanding and managing complexity through Bayesian network approach: The case of bribery in business transactions. *Journal of Business Research*, vol. 129, pp. 757–773. https://doi.org/10.1016/j.jbusres.2019.10.024

32. Marcot B.G., Penman T.D. (2019) Advances in Bayesian network modelling: Integration of modelling technologies. *Environmental Modelling & Software*, vol. 111, pp. 386–393. https://doi.org/10.1016/j.envsoft.2018.09.016

33. McLachlan S., Dube K., Hitman G.A., Fenton N.E., Kyrimi E. (2020) Bayesian networks in healthcare: Distribution by medical condition. *Artificial Intelligence in Medicine*, vol. 107, article ID 101912. https://doi.org/10.1016/j.artmed.2020.101912

34. Cai B., Huang L., Xie M. (2017) Bayesian networks in fault diagnosis. *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2227–2240. https://doi.org/10.1109/TII.2017.2695583

35. Pourret O., Naïm P., Marcot B. (2008) *Bayesian Networks: A Practical Guide to Applications*. Wiley, Hoboken.

36. Bielza C., Larranaga P. (2014) Discrete Bayesian network classifiers: A survey. *ACM Computing Surveys*, vol. 47, no. 1, article ID 5. https://doi.org/10.1145/2576868

37. Yuan C., Lim H., Lu T.C. (2011) Most relevant explanation in Bayesian networks. *Journal of Artificial Intelligence Research*, vol. 42, pp. 309–352. https://doi.org/10.1613/jair.3301

38. Lacave C., Díez F.J. (2002) A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, vol. 17, no. 2, pp. 107-127.

39. Lacave C., Luque M., Diez F. (2007) Explanation of Bayesian networks and influence diagrams in Elvira. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 4, 952–965. https://doi.org/10.1109/TSMCB.2007.896018

40. Altman E.I. (1983) *Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy*. Hoboken: Wiley.

41. Chen Y.C., Wheeler T.A., Kochenderfer M.J. (2017) Learning discrete Bayesian networks from continuous data. *Journal of Artificial Intelligence Research*, vol. 59, pp. 103–132. https://doi.org/10.1613/jair.5371

42. Weiss Y., Freeman W.T. (2001) Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation*, vol. 13, no. 10, pp. 2173–2200. https://doi.org/10.1162/089976601750541769

43. Ickstadt K., Bornkamp B., Grzegorczyk M., Wieczorek J., Sheriff M.R., Grecco H.E., Zamir E. (2010) Nonparametric Bayesian network. *Bayesian Statistics*, vol. 9, pp. 283–316.

44. Kurgan L.A., Cios K.J. (2004) CAIM discretization algorithm. *IEEE transactions on Knowledge and Data Engineering*, vol. 16, no. 2, pp. 145–153. https://doi.org/10.1109/TKDE.2004.1269594

45. Lustgarten J.L., Visweswaran S., Gopalakrishnan V., Cooper G.F. (2011) Application of an efficient Bayesian discretization method to biomedical data. *BMC bioinformatics*, vol. 12, no. 1, pp. 1–15.

46. Fayyad U.M., Irani K.B. (1993) Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, In: *Proceedings of 13th International Joint Conference on Artificial Intelligence (IJCAI'93)*, pp. 1022–1027.

47. Friedman N., Goldszmidt M. (1996) Discretization of continuous attributes while learning Bayesian networks. In: *Proceedings of 13-th International Conference on Machine Learning (ICML)*, pp. 157–165.

48. Schreiber J. (2018) Pomegranate: fast and flexible probabilistic modeling in python. *Journal of Machine Learning Research*, vol. 18, no. 164, pp. 1–6.

49. Yuan C., Malone B., Wu X. (2011) Learning optimal Bayesian networks using A* search. In: *22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2186–2191.

50. Chow C.K., Liu C.N. (1968) Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467.

51. Elkan C. (2001) The foundations of cost-sensitive learning. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'01)*, pp. 973–978.

About the author

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

DOI: 10.17323/2587-814X.2022.1.42.55

Information-logical model of express analysis of the state of the enterprise that meets the requirements of standards and regulations, based on publicly available data

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

Liudmila V. Zhukova
E-mail: lvzhukova@hse.ru

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

Abstract

The last 10 years have witnessed an explosive growth in the volume of information posted on the Internet and the digital economy, as well as the formation of official databases of various public authorities. The availability of a large information base open for research has facilitated the development of new methods and approaches to solving analytical problems. Building management and decision-making support systems based on the use of united disparate open data sources allows end users to make the most effective decisions. This is the approach that underpins business growth and managerial maturity at all levels — there is no alternative. Such an approach ultimately creates the conditions for further growth of the economy as a whole. This paper proposes the information and logical model of express analysis of compliance of socio-economic condition of the enterprise with the regulatory requirements of the control and supervisory authorities on the basis of open, publicly available information. The conclusions drawn on the basis of express analysis serve as a basis for deciding on the need for a more detailed, in-depth analysis of the state of individual enterprises.

Keywords: express analysis, information-logical model, enterprise status, regulatory requirements, control and supervisory bodies, open sources of information, publicly available data, structured and unstructured data

Citation: Bogdanova T.K., Zhukova L.V. (2022) Information-logical model of express analysis of the state of the enterprise that meets the requirements of standards and regulations, based on publicly available data. *Business Informatics*, vol. 16, no. 1, pp. 42–55. DOI: 10.17323/2587-814X.2022.1.42.55

Introduction

This paper proposes an information and logical model of express analysis of the compliance of the socio-economic condition of the enterprise with regulatory requirements on the part of the control and supervisory authorities on the basis of publicly available information. An information-logical model is built on the basis of the proposed concept, one of the important features of which is that the concept takes into account any requirements of different regulators, both quantitative and qualitative, imposed on economic objects of different types (enterprises, organizations, educational institutions, etc.) [1]. For different types of enterprises and the requirements imposed on them by the regulator, it is necessary to form different sets of components based on publicly available information, the aggregation of which, using the developed search table, results in the calculation of the value of the integral indicator, which is the basis of the express-analysis. Each component characterizes different aspects of the company's activities: economic, social, financial, technical, etc., and is evaluated in accordance with the methods of machine learning, mathematical statistics and econometrics [2, 3].

Both structured and unstructured information is used to carry out a rapid analysis of the state of an enterprise. Unstructured information is pre-structured using various methods of textual information processing [4–6].

The dynamics of the environment are increasing, the stability of the external environment is decreasing, and the requirements for a rapid response to crises are increasing. The amount of information that needs to be processed to make this or that decision is consistently increasing, at the same time the requirements for the quality, security and relevance of this information are becoming stricter.

The simultaneous use of structured and unstructured statistical data makes it possible to obtain a more accurate qualitative assessment of the object of study, taking into account changes not yet reflected in official statistical reports, which are provided with a certain periodicity and an inevitable time lag.

The result of the express-analysis is an assessment of compliance of the socio-economic condition of the enterprise with the regulatory requirements of the regulator. The conclusions drawn on the basis of the express analysis serve as a justification for deciding on the need for a more detailed, in-depth analysis of individual enterprises.

In recent years, research papers on various economic and mathematical studies have increasingly focused on the use of modern digital technologies for processing large volumes of structured, weakly structured and unstructured data from open Internet sources, machine learning and artificial intelligence methods in decision support models [7–10].

The use of innovative digital capabilities to collect and analyze publicly available information from the Internet allows us to perform additional analysis of the quality characteristics of various enterprises and other research objects. Such open data analysis can be carried out with the help of an auxiliary independent research object evaluation tool created on the basis of analysis of large volumes of structured, weakly structured and unstructured data from open internet sources, and to compare the results with the official research methodology on internal or official statistical data.

Control measures taken on the basis of official statistical information may come with a long delay, because between the end of the reporting period and the transfer of official statistical data on the state of the object to the public authorities may take from 3 to 8 months, which makes it difficult to respond promptly in force majeure situations.

In scientific research, many authors propose various economic and mathematical models based on official statistical information [11]. Most of them are econometric models or models that use machine learning techniques. As a rule, the available statistical data are divided into groups (demographic, social, financial, etc.), ranked, or somehow combined into a single integral indicator, and the factors are assigned weights. Often the result of such a study is an integral indicator (coefficient), which is useful for comparing objects. Such tools rely heavily on internal data or on an existing statistical base [12].

The use of structured and unstructured data analysis from open internet sources is the most comprehensive and versatile way to fully analyze the state of the economic object of study in a comprehensive way. It provides objective information on the current situation without intermediate processing based on the analysis of a wide variety of relevant data stored in the public domain in all Internet sources. If necessary, the results of external data analysis can be correlated with the results of similar analytical activities carried out using internal data. In addition, the results of analysis from open publicly available sources can complement official or internal data in some aspects of the subject's activities.

The advantage of using open data is the ability to obtain information at any periodicity (without reference to the regularity of updating, officially published statistical reporting), to expand and check compliance of the actual socio-economic condition of the object of research with official data.

1. Classification of publicly available information sources

All publicly available information can be represented in the form of different types of data. Currently, all existing data can be divided into:

1) structured;

2) poorly structured;

3) quasi-structured;

4) unstructured.

Structured data refers to data that is organized in a certain way, has a given structure, and describes a specific subject area. Taken together, this allows for reliable and in-depth analysis of this data. This information is most often presented in the form of tables.

Loosely structured data is data that does not follow a clear structure of tables and relationships in the database, but contains special delimiters (tags) that allow us to do semantic separation of the entire data set. Examples include XML documents.

Quasi-structured data is data in an unstructured format, which requires a lot of time to be processed by special tools. An example of such data is a website page.

Unstructured data is data that does not have a specific form and is not strictly fixed. At the moment this is the predominant data format due to the development of the information society. Approximately 80% of all currently available information is unstructured. Examples of such data are images, video, audio and textual information from social media.

Depending on the type of data, it requires its own preprocessing and processing methods. Most methods in mathematical statistics and econometrics are based on the analysis of structured information. Machine learning methods, neural networks allow us to analyze weakly structured, quasi-structured and unstructured data, identifying patterns in them. Furthermore, with various preprocessing procedures, these data can be reduced to structured data and incorporated into classical mathematical models.

If unstructured data is represented by text, pre-processing it using vectorization and classification methods allows us to bring it to a structured form.

The information to form the research base for the express analysis can be obtained from different sources, differing in status, frequency of updating and the degree of reliability of the information provided. *Table 1* presents the classification of publicly available information sources according to the reliability of the source.

Table 1.

Classification of sources of publicly available information

Source of information	Characteristics of information source	Example of information source	Type of information	Update on source
Official data generators and aggregators	Websites of federal and regional statistical bodies, websites of ministries and agencies that publish thematic data under the information disclosure regulations, the reliability of which is confirmed by the relevant public authority.	rosstat.gov.ru zakupki.gov.ru fssp.gov.ru cbr.ru wciom.ru	Structured data.	As a rule, the frequency of updates is once a quarter, or less frequently.
The websites and social media pages of the research subjects	Websites of enterprises, organizations of all forms of ownership, websites of platforms on which they are obliged to post information about their activities. The credibility of the information is usually confirmed only by the object of the research itself.	technomoscow.ru uniconf.ru tinkoff.ru 57.mskobr.ru	All data types.	Constant updating.
Unofficial data generators	Websites of organizations engaged in activities related to the research subjects and publishing data about them in open sources. Credibility is ensured by internal monitoring and control of information.	cian.ru hse.ru/rlms	Predominantly structured data.	According to the approved methodology, updates can be carried out either at set intervals or on an ongoing basis.
Unofficial data aggregators	Russian and international data aggregators, usually providing data for scientific and other studies. Credibility is ensured by internal monitoring.	bankodrom.ru banki.ru avtostat.ru data.worldbank.org.	Predominantly structured data.	Updates are usually carried out at intervals that correspond to the frequency with which official data are updated.
Unofficial internet sources of expert studies	Russian and international websites of expert organizations, rating agencies, personal pages of recognized experts. The reliability of the data is ensured by the reputation of the expert.	raexpert.ru ra–national.ru	All types of data.	The update is carried out in accordance with the source's internal rules.
Unofficial publicly available internet sources	Social media pages, blogs, comments on content, informal community pages. The validity of the data is usually not subject to verification.	moneyzz.ru pedsovet.su	Predominantly unstructured or weakly structured data.	Constant updating.

The information and logical model of building an integral indicator for express analysis of compliance of the socio-economic condition of the enterprise with the regulatory requirements of the control and supervisory authorities proposed in this article is based on the conceptual model of express analysis set out in [1]. A distinctive feature of the proposed conceptual model is that the authors propose to take into account the requirements of regulatory authorities as a starting point, while most Russian and foreign studies assess the state of the research object based on the requirements imposed on the object by its owners or investors. Another advantage of the conceptual model is the use of publicly available data, i.e. the possibility to obtain information at any time without being bound to the periods of updating the officially published statistical reports, and the possibility to check the compliance of the actual state of the research object with the official data. The proposed information-logical model is a combination of an algorithm for calculating the individual components of the integral index by using mathematical, econometric and statistical methods, the characteristics of input and output information at each stage, and, actually, the algorithm of calculating the values of the integral index by using a logical function based on a search table.

2. Components of the integral index and methods of their estimation

The integral index is a flexible express-analysis tool based on publicly available structured and unstructured data. The construction algorithm for the Integral Indicator is based on the aggregation of the individual values of each component in the set using a look-up table. Each component is estimated using mathematical, econometric and statistical methods, such as: logistic regression model, clustering and grouping methods, thematic modelling methods, etc.

The flexible toolkit of express analysis for management decision-making developed on the basis of the conceptual model is a sequence of five stages, starting from the requirements on the part of control and supervisory authorities, development and evaluation of a set of components characterizing the research object, their aggregation into a single integral indicator based on the search table, and ending with the monitoring and ranking of research objects according to the results of calculations [1].

Depending on the type of research object (industrial enterprise, banking organization, educational institution, etc.), based on the requirements of various regulators, a list of data sources for rapid analysis is formed: websites of research objects, news sources, electronic platforms or information aggregators, websites of state authorities, etc. The research database is created on the basis of the information from these sources. The flexibility of the tools proposed in the article is due to the fact that the list of components necessary for rapid analysis can be supplemented depending on the type of research subject, the frequently changing requirements of regulatory and supervisory authorities and an increasing number of publicly available information sources.

Table 2 provides a list of the possible components identified by the authors relating to the four blocks of types of input information for component calculation, types of variables of the calculated value of each component (according to the metrics proposed by Robert S. Kaplan and David P. Norton), and methods for estimating component values [13].

Various estimation methods are used to estimate the components of the integral indicator based on information about the survey objects from the database.

Table 2.

Components of the integral indicator and methods of their estimation

№	Components	Type of input information	Type of variable component calculated value	Method of estimation
Characterization of the financial condition of the object of study				
1	Probability of financial disadvantage	structured	categorical, ordinal	logistic regression model
The status identity of the object of study				
2	Status of the object of study in terms of scale	structured	categorical	cluster analysis
3	Status of the object of study as belonging to an abnormal group	structured	categorical	cluster analysis
Characteristics of the external information environment				
4	Media activity in relation to the object of study	weakly structured, quasi–structured and unstructured data	quantitative	semantic analysis
5	Positive tone of references to the subject of the study in online sources		quantitative	semantic analysis
6	Negative tone of references to the subject of the study in online sources		quantitative	semantic analysis
Regulatory requirements for the condition of the object of study				
7	Compliance with the requirements of public authorities	structured	binary or categorical	statistical and index analyses

2.1. Component 1. Probability of financial distress

Represents the probability of an unfavorable financial condition of the research object (bankruptcy, revocation of a license for financial reasons). In order to estimate this probability, a logistic regression model is applied based on financial statements data and their volatility indicators: standard deviation and variance, data on macroeconomic variables, data on public procurement as a supplier or buyer, In general, a logistic regression model takes the form [1]:

$$P\left(Y=1 \mid x, m, v\right)=\frac{1}{1+e^{-z}},$$

$$z=\beta_0+\sum \beta_i x_i+\sum \gamma_j m_j+\sum \varphi_k v_k,$$

where:

$P\left(Y=1 \mid x, m, v\right)$ – the conditional probability of the financial condition of the object under investigation being adverse;

β_0 – constant;

x_i – the variables that characterize the financial condition of the subject of the study;

m_j – variables characterizing the environment external to the object of study (macroeconomic factors);

v_k – non-quantitative indicators of the subject's performance;

β_i, γ_j, φ_k – regression coefficients to be estimated.

2.2. Components 2 and 3. Study object status by scale and abnormal group membership

Represents the clustering results to determine whether the survey object belongs to one of the classes. These components allow us to take into account specific features of all objects of the study type in terms of location, scale, type of activity, etc. The specifics of the obtained cluster of objects are taken into account, all of which allows us to assess more objectively the state of the enterprise in relation to objects from its class.

Clustering algorithms are divided into two types:

1. Hierarchical methods.

2. Non-hierarchical methods.

Hierarchical clustering methods are of two types [14, 15]:

1. Agglomerative (combining).

In this category of methods the initial objects are combined and the number of clusters is reduced [16]. This approach is carried out "bottom-up": creating small clusters and combining them into larger ones.

2. Divisive (decoupling).

Divisive type algorithms are characterized by the initial condition of having one cluster. This initial cluster is divided into smaller clusters. Dividing algorithms work top-down.

The disadvantage of these methods is the computational complexity on high dimensional data. A characteristic feature of hierarchical clustering methods is that observations once in a cluster cannot move to another cluster when further combining (disjoining) objects, in contrast to non-hierarchical methods.

The main distinctive idea of non-hierarchical clustering methods is to determine the center of the cluster and group all objects that are at a distance from the cluster center within a given threshold value [14, 15]. The group of non-hierarchical clustering methods includes algorithms of *k*-means family [16].

For high-dimensional data with an unknown number of clusters, the BIRCH (two-step or two-stage clustering) method based on *k*-means method is proposed. Two-step clustering does not require the number of clusters to be specified, since in the first step the optimal number of clusters is determined, and then the partitioning into homogeneous groups already takes place. This method makes it possible to analyze large amounts of both quantitative and qualitative data and works well with small memory sizes.

The quality of the resulting clustering can be evaluated using the silhouette measure *Sil* [17]:

$$Sil = \frac{1}{N}\sum\nolimits_{c_k \in C}\sum\nolimits_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{\max\left(a(x_i, c_k), b(x_i, c_k)\right)},$$

where:

Sil – the overall value of the silhouette measure of clustering of all data;

N – total number of objects in the sample;

C – set of all clusters;

c_k – *k*-th cluster on the set *C*;

x_i – *i*-th object, $i \in [1, N]$;

$a(x_i, c_k) = \frac{1}{|c_k|}\sum\nolimits_{x_i \in c_k}\left\|x_i - x_j\right\|$ –

the average distance from object $x_i \in c_k$ to other objects x_j in that cluster ck (compactness);

$|c_k|$ – number of objects in a cluster c_k;

$b(x_i, c_k) = \min_{c_l \in C \setminus c_k}\left\{\frac{1}{|c_l|}\sum\nolimits_{x_j \in c_l}\left\|x_i - x_j\right\|\right\}$ –

average distance from the site to objects x_j from another cluster c_l: $k \neq l$, $k, l \in [1, C]$.

Silhouette measure *Sil* takes values on the interval –1 to +1, where:

1 – all observations are located exactly in the centers of their clusters;

–1 – all observations are located at the centers of some other clusters;

0 – the observations are located at equal distances on average from the center of their cluster and the center of the nearest cluster.

2.3. Components 4, 5 and 6. Media activity in relation to the research object, positive and negative tone of mentions of the research object in Internet sources

The evaluation of these components is an analysis of unstructured or weakly structured data, predominantly textual. The semantic analysis to assess the meanings of the components characterizing media activity and the tone of the references to the object of research requires a preliminary preprocessing of this data, technical and linguistic data cleaning, compilation of the vocabulary of the words used in the texts.

Tone is the author's emotional attitude towards some object expressed in the text [18, 19]. One way to determine the tonality is to search for the emotional component in the text by the previously formed tonal dictionaries using linguistic analysis. The application of ready-made dictionaries to purified textual data allows us to classify textual units (sentences, words) into three categories: ambivalent, positive and negative. Semantic analysis of media activity, text categorization and application of machine learning techniques require text vectorization.

Vectorization is the process of converting textual documents into a numeric vector. The choice of vectorization method usually depends on a specific case, conditions, available hardware and technological tools. New methods and algorithms that improve vectorization quality and processing speed are constantly appearing and make it possible to introduce natural language processing into a model.

Currently the most popular algorithm implemented in many statistical packages, is the Bag-of-Words. Bag-of-words is a vector representation of an unordered set of words into a vector of dimension *n* [20–23]. Schematically, the algorithm can be represented as follows.

The whole text can be represented as a set of processed words, that is, individual terms (t_j), which with the help of this algorithm are translated into numerical data from the space R^n.

$$B: \text{words} \to R^n,$$

$$B\,(\textit{'some text in the Internet'}) = (w_{i,1}, w_{i,2}, ..., w_{i,n}),$$

where:

t_j – term j;

w_{ij} – the weight of term j in the document; the weight of the documents is rationed so that $0 < w_{ij} < 1$, для $\forall i$;

n – number of terms in space.

The document is then set up as follows:

$$d = (w_1, w_2, ..., w_{|V|}),$$

where:

d – document vector;

$|V|$ – the number of unique terms in the document.

The weight of a term can be set in several ways:

1. In a binary way:

$$w_i = \begin{cases} 1, t_i \in d \\ 0, t_i \notin d \end{cases}.$$

2. According to the number of occurrences of the term:

$$w_i = n_i,$$

where n_i – the number of occurrences of the term in the document.

3. Term Frequency – TF.

$$w_i = tf(t_i, d) = \frac{n_i}{\sum_{k=1}^{|V|} n_k},$$

where:

tf – thermal frequency;

n_i – the number of occurrences of the term in the document;

$\sum_{k=1}^{|V|} n_k$ – number of terms in the document.

4. Term Frequency – Inverse Document Frequency (TF–IDF).

Representation in the form of two parameters:$w_{ij} = tf_i \cdot idf_i$, where tf_{ij} – is the ratio of the number of terms t_i on paper d_j to the total number of terms in this document, $iidf_i$ – the number inverse of the number of documents in which the term occurs t_i. Thus, the more often a word occurs in this document, but less often in all documents in general, the greater the weight of that term in the document:

$$tf\left(t_i, d\right) = \frac{n_i}{\sum_{k=1}^{|V|} n_k},$$

$$idf\left(t, d\right) = \log \frac{|D|}{|d_i \supset t_i|},$$

where:

$d_i \supset t_i$ – the number of documents in which it occurs t_i;

$|D|$ – the number of documents in the enclosure.

The weight is then calculated as follows:

$$w_i = tf - idf(t_i, d, D) = tf(t_i, d) \cdot idf(t, d).$$

After vectorization, semantic text analysis algorithms are applied to determine tone, main themes, media activity, etc.

To calculate the values of the components, statistical methods are used to summarize the information about the object of study, e.g. by directly counting the occurrence of positive and negative words, the overall tone of the text is determined.

2.4. Component 7. Compliance with government requirements

This component is defined as a binary or ordinal indicator calculated using indices and statistical indicators. It represents an estimate of the number of irregularities in the activity of the object of study, in case normative and threshold values are given by the control or oversight state authorities.

A consolidated representation of the above is the information-logical model of express analysis of the compliance of the socio-economic state of the object of research with the requirements of control and supervisory authorities (*Fig. 1*). In *Fig. 1*, stage 3, which is key in the algorithm for calculating the integral indicator, is shown in general form. Detailed elaboration of stage 3 of the information-logical model is presented in *Fig. 2.* In this stage, the components of the integral index are evaluated and the values of the integral index itself are calculated depending on the values of each component in the set.

The interquartile range of IQR for the sample size n is proposed to transform the values of the component which characterize the media activity (component 4), the tone of the reference about the research object in Internet sources (components 5 and 6) and compliance with the requirements of public authorities (component 7). Here:

$F_n(x)$ – selective distribution function;

$IQR = Q_3 - Q_1$, where $Q_3 = 0.75$; $Q_1 = 0.25$.

The proposed information and logical model was tested on the basis of data from a group of industrial enterprises and financial sector enterprises.

A rapid analysis was conducted to match the need for financial assistance for 506 industrial enterprises registered in Moscow and the feasibility of its provision to federal and regional authorities. The express analysis was based on

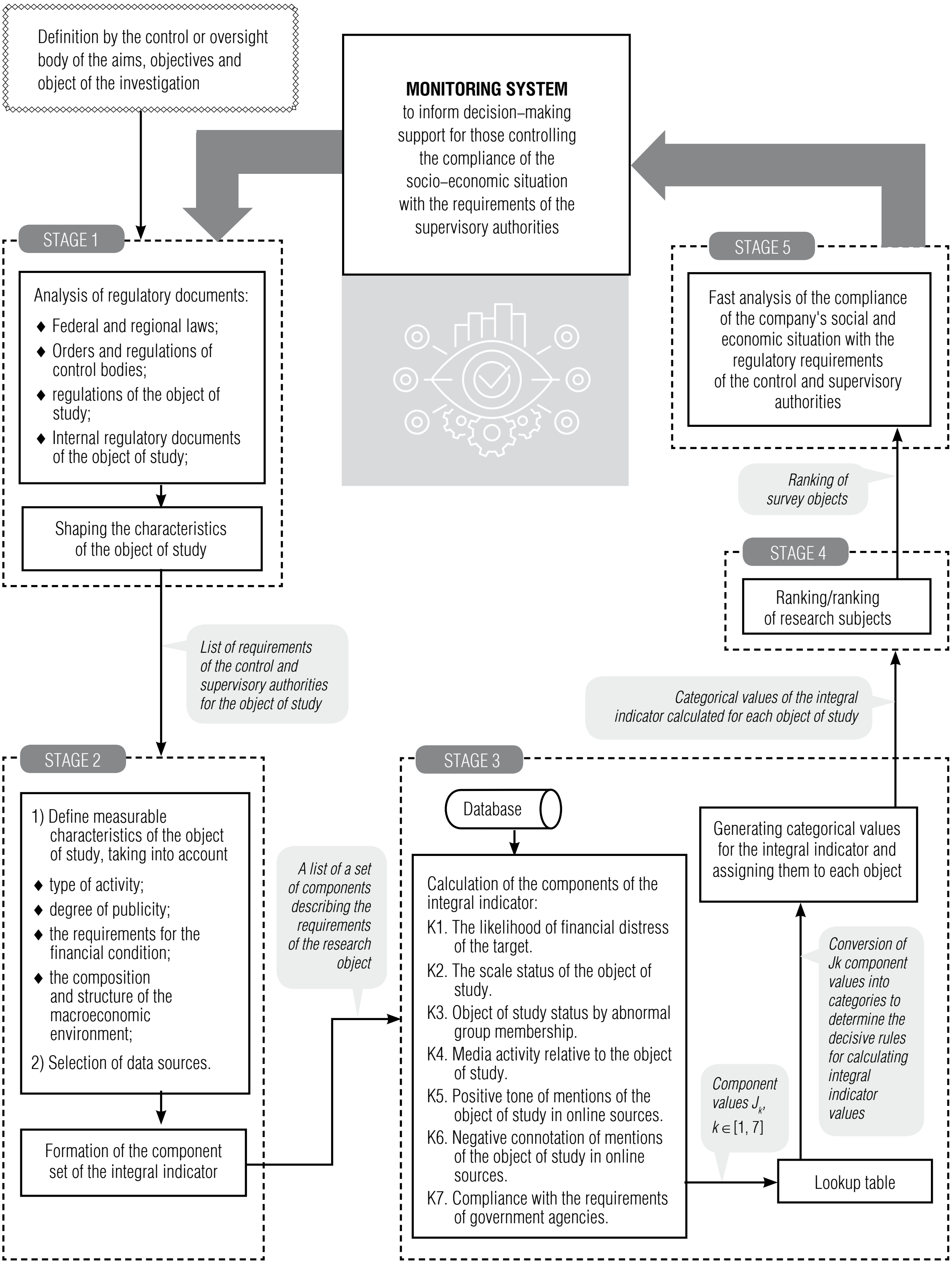
Definition by the control or oversight body of the aims, objectives and object of the investigation
MONITORING SYSTEM
to inform decision–making support for those controlling the compliance of the socio–economic situation with the requirements of the supervisory authorities
STAGE 1
Analysis of regulatory documents:
♦ Federal and regional laws;
♦ Orders and regulations of control bodies;
♦ regulations of the object of study;
♦ Internal regulatory documents of the object of study;
Shaping the characteristics of the object of study
List of requirements of the control and supervisory authorities for the object of study
STAGE 2
1) Define measurable characteristics of the object of study, taking into account
♦ type of activity;
♦ degree of publicity;
♦ the requirements for the financial condition;
♦ the composition and structure of the macroeconomic environment;
2) Selection of data sources.
Formation of the component set of the integral indicator
A list of a set of components describing the requirements of the research object
STAGE 3
Database
Calculation of the components of the integral indicator:
K1. The likelihood of financial distress of the target.
K2. The scale status of the object of study.
K3. Object of study status by abnormal group membership.
K4. Media activity relative to the object of study.
K5. Positive tone of mentions of the object of study in online sources.
K6. Negative connotation of mentions of the object of study in online sources.
K7. Compliance with the requirements of government agencies.
Component values J_k, $k \in [1, 7]$
Lookup table
Conversion of Jk component values into categories to determine the decisive rules for calculating integral indicator values
Generating categorical values for the integral indicator and assigning them to each object
Categorical values of the integral indicator calculated for each object of study
STAGE 4
Ranking/ranking of research subjects
Ranking of survey objects
STAGE 5
Fast analysis of the compliance of the company's social and economic situation with the regulatory requirements of the control and supervisory authorities

Fig. 1. Information–logical model of the algorithm for calculating the components of the integral indicator.

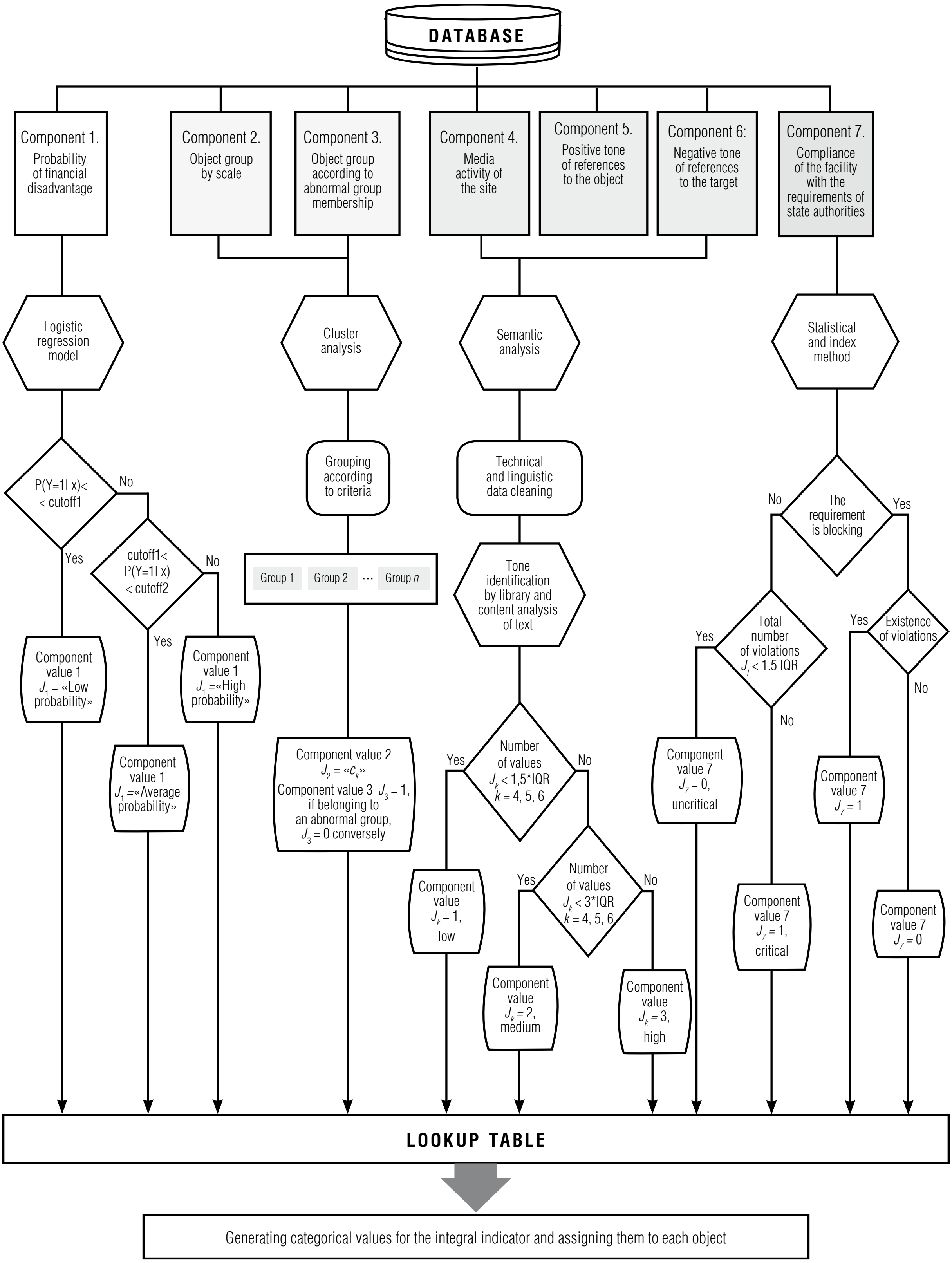
DATABASE
Component 1. Probability of financial disadvantage
Component 2. Object group by scale
Component 3. Object group according to abnormal group membership
Component 4. Media activity of the site
Component 5. Positive tone of references to the object
Component 6: Negative tone of references to the target
Component 7. Compliance of the facility with the requirements of state authorities
Logistic regression model
Cluster analysis
Semantic analysis
Statistical and index method
P(Y=1| x)< < cutoff1
No
Yes
cutoff1< P(Y=1| x) < cutoff2
No
Yes
Component value 1 J_1 = «Low probability»
Component value 1 J_1 =«High probability»
Component value 1 J_1 =«Average probability»
Grouping according to criteria
Group 1
Group 2
···
Group *n*
Component value 2 J_2 = «c_k»
Component value 3 J_3 = 1, if belonging to an abnormal group, J_3 = 0 conversely
Technical and linguistic data cleaning
Tone identification by library and content analysis of text
Number of values J_k < 1,5*IQR k = 4, 5, 6
Yes
No
Component value J_k = 1, low
Number of values J_k < 3*IQR k = 4, 5, 6
Yes
No
Component value J_k = 2, medium
Component value J_k = 3, high
The requirement is blocking
No
Yes
Total number of violations J_j < 1.5 IQR
Yes
No
Existence of violations
Yes
No
Component value 7 J_7 = 0, uncritical
Component value 7 J_7 = 1, critical
Component value 7 J_7 = 1
Component value 7 J_7 = 0
LOOKUP TABLE
Generating categorical values for the integral indicator and assigning them to each object

Fig. 2. Detailed step 3 of the information–logical model of the algorithm for calculating the components of the integral indicator.

open data for 2016, 2017 and 2018. The results obtained were in line with the actual data for the following year on the assignment of subsidies and benefits by the Moscow City Government [24].

The proposed conceptual model of express analysis of the compliance of the socio-economic condition of the object of research with the stated requirements on the part of control and supervisory authorities has been tested to assess the socio-economic condition of a commercial bank. The controlling body in this case is the Central Bank of Russia — the supervisory authority in the banking sphere. In accordance with the CBR requirements for bank reliability, the values of the four components of the integral index were obtained and its value for each bank was calculated. The predictive ability of the constructed model was confirmed by their actual state as of March 2020 [1].

Conclusion

This article suggests an information and logical model for express analysis of compliance of the social and economic condition of the object with the requirements of the control and supervisory bodies with the use of open public data. The proposed information-logical model is based on the concept of using an integral indicator for rapid analysis of compliance of the socio-economic condition of the object, regardless of its type of requirements imposed on it by control and supervisory authorities.

The classification of information sources and methods of processing them depending on the type of data is given.

An algorithm is proposed for calculating the possible components allocated by the authors relating to the four blocks of input information types, types of variables of the calculated value of each component (in accordance with the metrics proposed by Robert S. Kaplan and David P. Norton), and methods for estimating component values.

The developed conceptual model has been tested to carry out a rapid analysis of the compliance of the socio-economic condition of two different types of facilities with the requirements imposed on them by the supervisory authorities on samples of 506 industrial enterprises [24] and 111 banks [1]. ■

References

1. Bogdanova T.K., Zhukova L.V. (2021) The concept for valuation the position of the control object based on a universal complex indicator using structured and unstructured data. *Business Informatics*, vol. 15, no. 2, pp. 21–33 (in Russian). http://doi.org/10.17323/2587-814X.2021.2.21.33
2. Krichevskiy M.L. (2019) Methods of machine learning in choosing a strategy of an enterprise. *Russian Journal of Innovation Economics*, vol. 9, no. 1, pp. 251–266 (in Russian). https://doi.org/10.18334/vinec.9.1.40093
3. Opekunov A.N., Kuzmina M.G. (2019) Principles of forming models for forecasting the probability of bankruptcy of enterprises using machining elements. *Models, Systems, Networks in Economics, Technology, Nature and Society*, no. 4, pp. 24–31 (in Russian).
4. Kasevich V.B. (1977) *Elements of general linguistics*. Moscow: Nauka (in Russian).
5. Savenkov P.A. (2019) Using methods and algorithms of machine learning in management decision support systems. *Bulletin of Science and Education*, nos. 1–2 (55), pp. 23–25 (in Russian). https://doi.org/10.24411/2071-6168-2019-10207
6. Popova S.V., Khodyrev I.A. (2012) Keyword extraction. *Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics*, no. 1 (77), pp. 67–71 (in Russian).

7. Eliseeva E.N. (2019) Financial instruments for assessing the insolvency of industrial enterprises. *Region: systems, economy, management*, no. 3 (46), pp. 132–140 (in Russian).

8. Medvedev D.A. (2019) Big data: the reasons for their emergence and how they can be used. *Science and Education Today*, no. 4 (39), pp. 14–16 (in Russian).

9. Pang B., Lee L. (2008) Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1–2, pp. 1–135. https://doi.org/10.1561/1500000011

10. Morozov A.N. (2018) Alternative sources of statistical information as the basis for political decision making. *Problems of State and Municipal Management*, no. 2, pp. 50–70 (in Russian).

11. Puzanov A.S., Trutnev E.K., Markvart E., Popov R.A., Safarova M.D. (2017) *Strategic planning and urban regulation at the municipal level*. Moscow: Delo (in Russian).

12. Andreeva N.A., Ugrimova S.N. (2019) To the question of the application of statistical methods of integral estimation of effectiveness of system of management of industrial enterprises. *Accounting and Statistics*, no. 1 (53), pp. 42–49 (in Russian).

13. Norton D. P., Kaplan R.S. (2008). *Balanced scorecard*. Moscow: Olymp-Business.

14. Chugunov V.R., Zhukova L.V., Kovalchuk I.M., Kovaleva A.S. (2017) Mathematical methods of data grouping for making management decisions in planning tasks. *Actual Problems of System and Software Engineering 2017. Proceedings of the 5th International Conference on Actual Problems of System and Software Engineering Supported by Russian Foundation for Basic Research. Project #17-07-20565*, pp. 333–341 (in Russian).

15. Baresyagin A.A., Kupriyanov M.S., Stepanenko V.V., Kholod I.I. (2007*) Data analysis technologies: Data Mining, Visual Mining, Text Mining, OLAP*. 2nd edition. St. Petersburg: BXV-Petersburg (in Russian).

16. Shalymov D.S. (2008) Stable clustering algorithms based on index functions and stability functions. S*tochastic optimization in computer science*, vol. 4, pp. 236–248 (in Russian).

17. Kaufman L., Rousseeuw P. (2005) *Finding groups in data: An introduction to cluster analysis*. Wiley-Interscience.

18. Semina T.A. (2020) Sentiment analysis: modern approaches and existing problems. *Social sciences and humanities. Domestic and foreign literature. Ser. 6, Linguistics*, no. 4, pp. 47–64 (in Russian).

19. Liu B. (2010) Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition* (eds. N. Indurkhya, F.J. Damerau). London: Chapman and Hall/CRC.

20. Bolshakova E.I., Vorontsov K.V., Efremova N.E., Klyshinsky E.S., Lukashevich N.V., Sapin A.S. (2017) *Automatic text processing in natural language and data analysis. Tutorial*. Moscow: HSE (in Russian).

21. Demidova L.A., Stepanov M.A. (2019) An approach to solving problem of the structural transformations detection in the time series' groups. *Cloud of science*, no. 2. pp. 201–226 (in Russian).

22. Popova S.V., Khodyrev I.A. (2012) Extraction of keyword combinations. *Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics*, no. 1 (77), pp. 67–71 (in Russian).

23. Krasnyansky M.N., Obukhov A.D., Solomatina E.M., Voyakina A.A. (2018) Comparative analysis of machine learning methods for solving the problem of classifying documents of scientific and educational institution. *Bulletin of Voronezh State University*, no. 3, pp. 173–182 (in Russian).

24. Zhukova L.V. (2021) Express-analysis of the state of industrial enterprises of Moscow using the universal comprehensive indicator. *Economic Science of Modern Russia*, vol. 4 (95), pp. 89–96 (in Russian).

About the authors

Tatiana K. Bogdanova

Cand. Sci. (Econ.);

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

E-mail: tanbog@hse.ru

ORCID: 0000-0002-0018-2946

Liudmila V. Zhukova

Assistant Professor, Department of Applied Economics, Faculty of Economic Sciences, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: lvzhukova@hse.ru

ORCID: 0000-0003-1647-5337

DOI: 10.17323/2587-814X.2022.1.56.71

Digital transformation of music aggregation and distribution companies: The case of Russia

Artem I. Altynov
E-mail: aaltynov@hse.ru

HSE University
Address: 11 Pokrovsky Bulvar, Moscow 109028, Russia

Abstract

Currently distributors ensure the operation of the whole value chain in the music industry, while most researchers focus on technological, streaming and copyright impact in light of digitalization. This paper tries to understand the influence of digitalization on business models and the role of music distributors in a value chain. Research identifies operational processes that were changed due to digitalization, barriers that arose, and actions taken to overcome them on the corporate level. Through retrospective case study based on an interview with the CEO, the experience of the Russian music distribution company Broma16 is analyzed. This paper describes the goals of business model transformation, drivers, performed activities and their results. The research derives four consequences of digitalization for the firm's business model and role in a value chain: the firm can be considered as a technological company rather than a music company. Distribution relies on digital instruments for management and marketing. Local intermediaries get more opportunities to enter foreign markets, but they have to perform innovations in these markets. Music distributors operate at the complicated intersection of copyright and technological aspects. The research applies a general theoretical framework for the study of digital transformation of business models. A similar approach can be used to do research on companies in music and other creative industries, and to conduct workshops with industry representatives. The paper provides value for practitioners in emerging music markets, for example, Brazil, Argentina and Mexico, due to the presentation of management practices towards digitalization and the consequences of transformation for the distribution role.

Keywords: digitalization, digital transformation, music distribution, Russia, case study

Citation: Altynov A.I. (2022) Digital transformation of music aggregation and distribution companies: The case of Russia. *Business Informatics*, vol. 16, no. 1, pp. 56–71. DOI: 10.17323/2587-814X.2022.1.56.71

Introduction

Listening to music is one of the main digital practices in Russia and in the world [1, 2]. Digitalization assumes the creation of entirely new business models that fully rely on digital technologies and data to enable innovations in products and services [3]. In the music industry, it is connected with streaming platforms. In 2019, worldwide revenue from sales of physical copies was equal to $4.4 billion, while revenue from streaming – $11.4 billion, representing more than 50% of total revenue of the industry [4].

The Russian music industry has a very optimistic future, since it follows global trends. Eastern markets including Russia are characterized by high growth rates, and this attracts Western firms [5]. The Russian streaming services developed by Yandex and VK successfully compete with international platforms iTunes and Spotify and capture almost 70% of the digital music market in the country [2]. 87% of Russian listeners use streaming services, while worldwide only 61% of listeners use streaming on average [6]. In terms of the popularity of streaming, Russia is overtaking developed countries, for example, the USA (68%), Canada (56%), UK (56%) [7]. The Russian music industry is unique due to the activity of listeners and the success of domestic companies that develop music services.

Streaming platforms cannot collaborate with all people producing music all over the world [8]. Distributors aggregate licenses from a big number of artists and present theirs rights, ensuring the operation of the whole value chain in the industry [9]. The study of how digitalization impacts on distributors' roles gives us an understanding of how the operational processes in the industry have changed. The main research question of this study is the following: how has digital transformation affected business models and the role of music distributors in a value chain? This paper answers these questions using the example of the Russian company Broma16 and identifies the main operational processes that were changed, the barriers that arose during the transformation, and decisions taken to overcome them. This paper is based on a retrospective case study of the company Broma16 that represents the rights of more than 1000 clients worldwide and has more than 25 years of experience in the field. The data was gathered through a semi-structured interview with the CEO of the company. The case study is led by theory grounded in the holistic business model framework proposed in [10] for the analysis of digitalization in creative industries.

Changes in the external environment were the main driver of transformations. Russian artists had no instrument for music management on streaming platforms, while distributors had to buy expensive software from foreign firms. Foreseeing the industry's future on the Western example, the company started the development of a platform for music management internally. Addressing the gap, Broma16 was among the first companies to provide digital distribution services for Russian artists. At the time, senior management decided to separate developments into an external company Heaven11, that now sells digital instruments for music distribution. Broma16 faced conservatism in the value chain connected with the lack of skills. Emerging music streaming markets in other countries presented growing potential, and both companies focused on foreign markets. Currently 90% of Broma16 clients are international artists and labels.

The transformation of music distributors is not the most popular area of research in the field, where all articles can be distinguished between several topics (*Table 1*). The first group studies the impact of technologies on production approaches, quality of products, existing and emerging markets, and consumption practices. Several articles are dedicated to the appearance of new intermediaries, curators, who rely heavily on data [11], freelance managers and designers [12]. The second group is aimed at streaming influences on consumers' experiences, music and media industries, artists and producers. Authors actively study business models and strategic management in leading streaming companies. The third group of articles is connected with copyright issues, while the main topics cover reasons and patterns of piracy, the negative impact on artists and approaches to reduce it, copyright policies on the national level. As the main areas for studies show, the development of digital technologies, dissemination of streaming services and copyright issues are key trends shaping the music industry.

The large gap in the analysis of intermediary roles is observed, while only several articles try to study this topic [9, 12, 19]. Authors strive to figure out how intermediaries emerge, which roles they perform, leaving behind an understanding of business models and their development under digitalization. Alternatively, for other topics technological impact is of interest to researchers, while we note the lack of articles dedicated to the study of intermediaries on the intersection of business and technological issues. To fill this gap, our study proposes the following hypothesis: distributors should be considered as technological, rather than music companies, whose work is shaped by vast amounts of data and legal issues.

The analysis we carried out fills this gap and offers insights into digitalization in Broma16 and four consequences for the firm's business model and the role of the company in the value chain. These consequences cover the importance of technologies in working processes of todays' distributors; new opportunities for local distributors due to digitalization; intersections of legal and technological issues in the work of distributors. The outcomes of the article can be relevant for countries where the rapid develop-

Table 1.

Key topics and sub-topics identified within the review of literature

Topic	Sub-topics	Examples of articles
Technological impact	Impact of digital technologies on production, promotion, distribution, consumption practices, industry structure and business models of artists, labels	[11, 13–15]
Streaming impact	Impact of streaming services on music expenditures, listening behavior, value chains, business and production practices	[8, 16–18]
Intermediaries	Emergence of music aggregators. Roles of intermediaries	[9, 12, 19]
Business of technologies	Business models of streaming, producer–oriented platforms. Partnerships between artists and platforms	[20–22]
Copyright issues	Phenomenon of piracy. Copyright policy. The fight of companies against piracy	[23–25]

ment of streaming market is observed in the recent years: Mexico, Brazil, Argentina [4, 7]. The practices reviewed are of interest in countries that try to establish domestic companies as leaders and protect their market from the influence of Apple, Spotify and other foreign firms, providing services along the value chain [4, 26]. These companies are focusing on Russia, but domestic firms set competition on the market, so their experience can be useful [4].

1. Digitalization of the music industry

In the early days of the industry, special laboratories produced records, while distributors were responsible for selling vinyl. After the identification of the possibility to gain higher revenues, they started to invest their own money in laboratories, which gradually were transformed into recording studios [12, 22]. Studios made records, while distributors started to shape the market and research the audience [14]. This led to the creation of groups responsible for the repertoire inside distribution firms, all of which attracted the attention of studios [12]. Studios decided to analyze demand and audience, despite the lack of distribution channels, and developed over the time into major labels that found new artists and created records [22]. Distributors were responsible only for the replication of copies [12]. The whole value chain was established (see *Fig. 1*). A similar value chain existed during all physical formats until the era of digitalization started in the early 2010s [22].

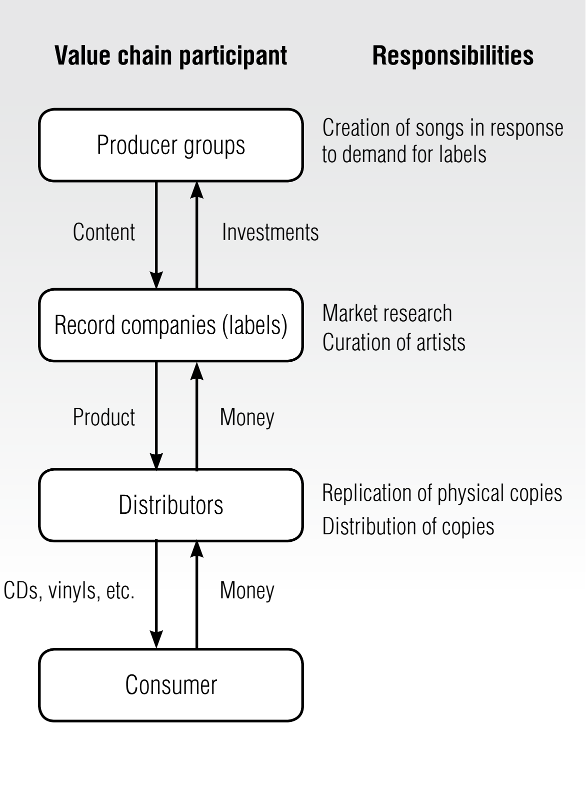

Fig. 1. Value chain in the music industry before digitalization [9, 12].

Digitalization (laptops and software, ready-made samples, electronic instruments) has transformed production and made it cheaper for independent musicians [11, 14, 20]. Production does not require investments, and labels have established a focus on marketing services and perspective artists, where the product is a marketing unit (content + marketing strategy) [7, 22]. Simultaneously musicians can perform management functions themselves and turn into professional artists without producers and labels [12, 20].

Now artists receive money not for sales, but for licensing, when they transfer rights on distribution to stores [14, 18]. Intermediaries are necessary, because the relationships between stores and artists are characterized by information asymmetry, where one side has bargaining power [8, 17]. The stronger party can capture a share of the margin by threatening partners [9, 16]. Hiring one intermediary allows many artists to cut transaction costs through the width of the music library that determines market power [8]. Streaming services are also interested in intermediaries, since they need to sign deals with millions of owners of copyrights (record labels or individual artists) all over the world [8, 9, 17]. Currently distributors stay in the center of the value chain, ensuring its functioning (see *Fig. 2*).

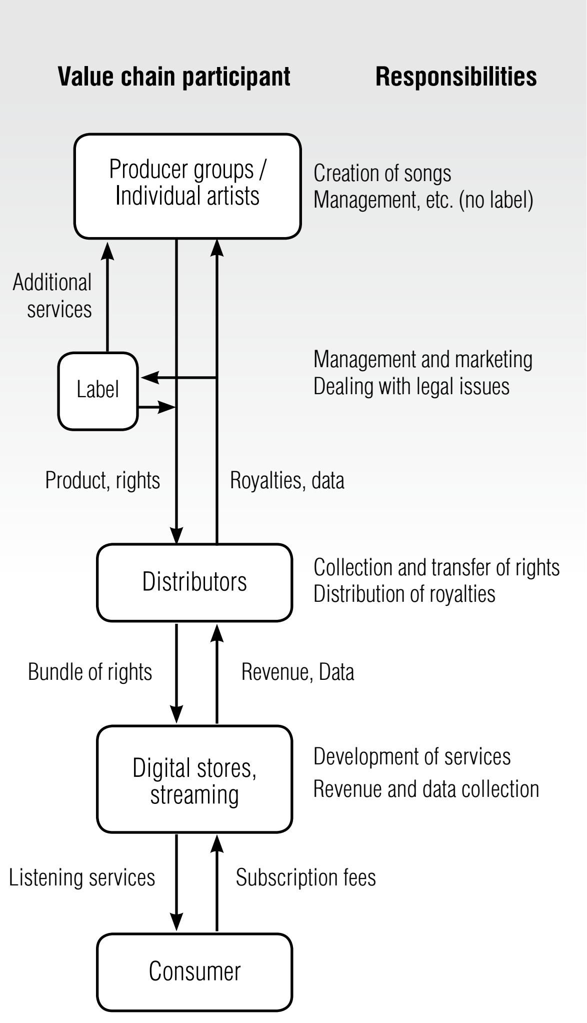

Fig. 2. Value chain in the music industry after digitalization [9, 12].

Distributors have more mechanisms to influence the market. While labels concentrate on prospective artists who bring profit, distributors do not care whose music they deliver to platforms, serving as many artists as possible to maximize revenue [14]. Thanks to the width of the library, distributors receive more market power in terms of communication with streaming platforms. Distributors are the only “musical” party that can limit the influence of technological companies, for example, by providing exclusive content.

2. Methods

This paper uses a case study as a research method for several reasons. Firstly, it represents an intensive study of a single unit for the purpose of understanding a larger class of similar units [27]. Focusing on the one object enhances homogeneity in case design, which is important to draw valid conclusions [28, 29]. Research applies the typical case technique to constitute a story as the most representative technique used to probe causal mechanisms [30]. The selection of the company Broma16 was guided by two criteria: the company’s experience in the music industry and adoption of digital technologies. Broma16 was founded more than ten years ago, and its staff has more than 25 years of experience in the music industry [31]. Broma16 is officially affiliated with the firm Heaven11, providing innovative digital solutions for music companies [32]. The case is limited within the ten year period since iTunes was officially launched in Russia in 2012 and it was the beginning of streaming development.

The research applies a specific theoretical framework to frame a case study [33]. A common framework provides the basis for formulating framing questions and identifying issues [29]. The research uses a holistic business model framework developed in [10] to analyze digitalization in companies in creative industries. This framework has been chosen instead of popular ones due to several advantages. First, it was constructed through the literature review of more than 70 definitions and concepts of a business model. Thus, it presents a synthesis of previous studies. Secondly, the framework was tested and finalized through 80 case-studies, and it was reviewed during three workshops with business representatives from creative industries (see *Fig. 3*). The business model covers all aspects, explaining how a firm does business and will operate in the future, and it is considered as an ideal type to plan innovations [34, 35].

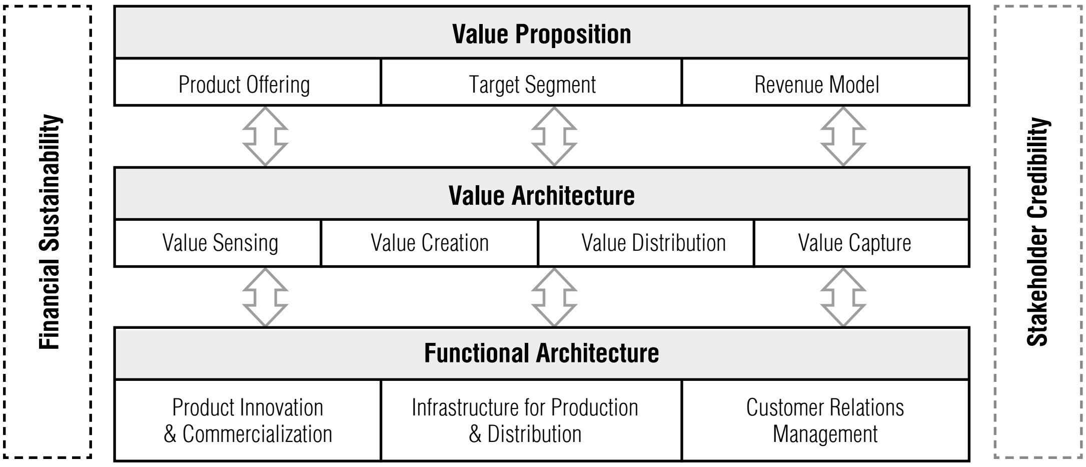

Fig. 3. The holistic business model framework [10].

This study focuses on the foundational layer of the framework, because it reflects transformed processes. The foundational layer is also called the functional architecture, since it consists of core activities of a firm [10]. The research targets three main aspects of the business model: product innovation and commercialization, infrastructure for production and distribution, customer relations management. Through the case study changes in these aspects will be reflected.

The interview tool was used for data collection to gain in-depth knowledge about the phenomenon and identify the most significant issues [29]. All internet materials about the company were used to formulate the background for the interview and the case study (*Appendix 1*). The proposed interview guideline does not follow a rigid structure (*Appendix 2*). To obtain relevant data, it was necessary to conduct interviews with top managers who have been working in the company a sufficient amount of time, since the case study is retrospective. Employees have a non-disclosure agreement (NDA) that covers all information about technologies. Interviewing employees on this topic would be unethical. To overcome the barrier, an online interview with the CEO was conducted. He has been at the forefront of the company's digitalization for the last 10 years.

Interviewing the CEO allowed us to understand the motivation of the company towards transformation and to explain why the enterprise architecture is to be developed. Based on the interview, the motivation could be discussed in the context of the Business Motivation Model, developed by the Object Management Group (OMG). This framework covers the goals of a company, explains how a firm intends to accomplish it, opportunities and threats for a business, its strengths and weaknesses. The model pays attention to the external environment, covering drivers of development [36]. An interview script to collect and interpret data in a narrative way was applied. Since only one interview was conducted, the application of coding was unnecessary and faced risks of missing important details.

3. Case description

Broma16 is a distribution company established in 2010 in Moscow. It represents the rights of 1000 clients, including labels, artists and other distribution companies, and over a million music recordings. The compa-

Group of services	**Music publishing**	**Music licensing**
Customers	♦ Authors ♦ Performers ♦ Producers ♦ Labels ♦ Distributors	♦ Production studios ♦ Theaters ♦ Telecom operators ♦ Radios ♦ Film, game and companies
Services	♦ Administration of copyrights, including adjacent rights ♦ Collection of royalties from streaming services, YouTube (cover versions or user-generated content) ♦ Tracking of plays at online and terrestrial radios, TV, in public places ♦ Handling of release details (artworks, title, additional information) ♦ Consulting and marketing services (pitching to playlists and TV, movie, games, ads production collectives	♦ Delivering licenses for streaming services, podcasts, online, NV and radio advertising, online and live performances, karaoke services, mass media products ♦ Delivering licenses for retail audio products (CDs and vinyls); audio products for promotion; audio-visual products (DVD, Blu-Ray); digital memory devises; background music services ♦ Delivering licenses to play music in public places

Fig. 4. Key services provided by Broma16 [31].

ny's activities can be distinguished between two groups (see *Fig. 4*). The first group is music publishing services for producers to license repertoire, manage copyrights, collect royalties. The company can help with release marketing, delivering flexible contracts with different bundles of services. The second group is delivery of licenses on behalf of publishers to production studios that develop music products (podcasts, advertising, etc.), physical products (CDs, etc.), background music services in public places [31].

The company established a technology-driven approach to deliver services based on a digital platform for repertoire management and licensing. The platform provides information on every transaction, regardless of geography and presents it in granular reports to ensure transparency of copyrights and distribution of royalties. The platform design allows them to control repertoire ownership over multiple territories, multiple right types, revenue streams. The platform is the core of the business model, enabling all services in the company's portfolio, but the situation was not always like that.

Initially the company was created to produce a repertoire of artists and deliver marketing services. Changes in the motivation of the company are illustrated by the following statement from the CEO: "… Market reorientation from albums to singles was obvious more than ten years ago, and company was aimed at the promotion of singles. But shift to the licensing model of distribution, started in 2012, required from producers to upload music and collect royalties, where the problem was to document the music."

Broma16 was among the first companies in Russia to provide such services for producers. Russian distributors had to buy software from foreign companies, and it was expensive. The firm started in-house development of a digital platform in 2010–2011. Since 2012, development has been realized within the affiliated company Heaven11 (see *Fig. 5*), which provides a common database and platform for music management. Currently all Broma16 services are enabled by the platform.

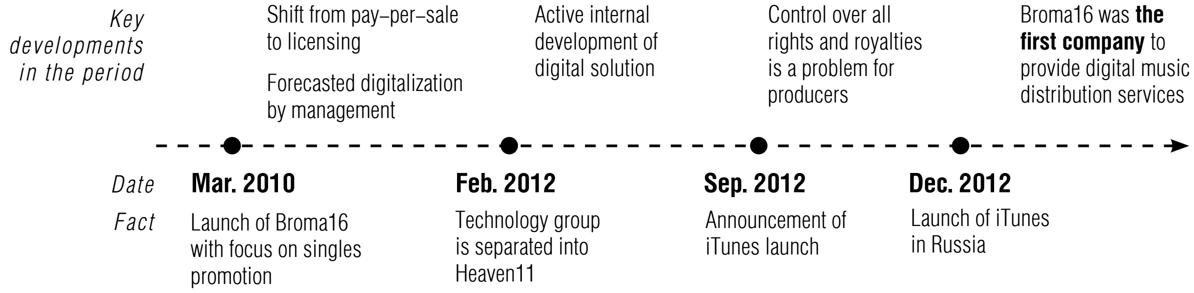
Key developments in the period
Shift from pay–per–sale to licensing
Forecasted digitalization by management
Active internal development of digital solution
Control over all rights and royalties is a problem for producers
Broma16 was **the first company** to provide digital music distribution services
Date
Mar. 2010
Feb. 2012
Sep. 2012
Dec. 2012
Fact
Launch of Broma16 with focus on singles promotion
Technology group is separated into Heaven11
Announcement of iTunes launch
Launch of iTunes in Russia

Fig. 5. Timeline with key developments during the digitalization of the company.

Heaven11 offers products to various parties who create, license or use music commercially [32]. The core product is the multi-territorial multi-rights database for music works, master recordings, record releases, audio-visual products and lyrics. The key technical components of the system include a common database for music works, supplemented with instruments for data importing/exporting in different formats, deduplication and cleaning, automation solutions for processing Digital Service Provider reports, songs identification. This makes it possible to send releases to platforms, track submissions and manage repertoire, automatically distribute royalties for all parties with detailed statements, prepare reports and documentation. Key clients of the company are publishers and labels, online music platforms, Collective Management Organizations.

4. Findings

Digital technologies have disruptively changed the business models of companies in creative industries, and Broma16 is also an example of this [34]. The firm focused on the promotion of singles, but the processes of value creation and customer appearance have been changed by the introduction of a platform for music management. These changes have strongly affected the business model of the company. The firm moved along the value chain from promotion to distribution.

To analyze transformations in the business model, the foundational layer in the holistic business model framework developed by [10] is considered (*Fig. 3*). Changes have affected all three components in the foundational layer. Broma16 was established for the promotion of singles, but due to the appearance of music streaming senior management decided to focus on distribution. Thus, the key product bundle has been fully changed. Currently the company stays after marketing services in a value chain (*Fig. 2*). Instead of providing marketing services, Broma16 deals with music documentation, licensing and royalties distribution.

Licenses have become the key product in the music industry, since they represent amounts of money passing from fans to artists and labels through streaming services and distributors. Working directly with licenses, distributors gain higher revenues and more market power. To establish this control, distributors have to serve as many clients as possible and digital technologies are the important instrument for this.

The platform for music management allowed Broma16 to enter the Russian music distribution market among the first. It proves that the infrastructure component became the key in the business model due to digitalization [10]. All customer relations have been automated on the platform. The group of potential customers has expanded from artists with promising singles to all stakeholders who want to sell or buy music licenses (*Fig. 4*). Currently Broma16 represents the rights of more than 1000 clients, including artists and labels.

The OMG Business Model Motivation framework is applied to understand transformations of the business model of Broma16. At the company's inception, the CEO's motivation was to promote singles because of their growing influence. The announcement of the iTunes launch caused changes in the goals of the firm. Russian artists and labels had no instrument to upload music on streaming services, track all purchases, collect and distribute royalties. So, the digital transformation of consumption practices identified gap in Russian music industry.

Appearance of new market niche for digital distribution can be considered as the main driver. Realizing this, senior management of the company decided to shift focus from the promoter role to the distributor role. Since 2011 the motivation of the Broma16 is to create convenient digital instrument for uploading music on streaming platforms.

But the company faced some threats for business development. Russian distributors had to buy software for music management from foreign companies. That time firms could not afford expensive foreign digital solutions. Broma16 had strengths, since the company was young. It gave flexibility for strategic development. Senior managers had more than 15 years of experience in the industry and foreseen the movement towards digital music consumption before the announcement of iTunes launch. Analyzing the music industry in western countries, senior management decided to start in-house development of the platform for music management. When Apple announced the iTunes launch in Russia, development was already separated into the company Heaven11. A successful industry forecast and pro-active development of the platform allowed both firms to be among the first companies on the market: Broma16 as a digital music distributor and Heaven11 as a technological provider for music distributors.

Another threat to the company's business was the conservatism of clients (artists and labels). According to the CEO: "... Broma16 was ready to provide today's set of services eight years ago, but artists were not ready to use the platform for music management. Even today some producers try to contact Broma16 to check music submission, though they can do it on the platform."

Skills in the use of professional software, platforms, content management solutions and databases are required from artists. The company experienced several approaches to solve the problem. First – educational services for clients, explaining, how to register songs in the database, manage content, use collected data for commercial purposes. Second – a focus on foreign markets. According to the CEO, "currently more than 90% of the clients of Broma16 are international artists and labels." In the early 2010s music streaming markets were emerging in a lot of countries. Due to market conditions, both companies have been successful on an international scale. Currently the companies have offices in Amsterdam (Broma16) and Dublin (Heaven11).

5. Discussion

External development of the industry stimulated transformations in Broma16, all of which is usual for creative industries [34, 37]. Opportunities to upload music on iTunes determined the need for a digital solution to collect and distribute royalties. Demand for new intermediaries is often connected with digitalization [35]. The lack of software for music management shows the increasing importance of infrastructure [38]. Broma16 presented a platform to address this gap. It resulted in the transformation of all components in the foundational layer of the business model, following the experience of other firms in the creative industries [39]. However, companies used to experiment with business models to build more flexible and adaptive ones with different revenue streams [35, 37, 40, 41]. In the case of Broma16, experiments were not observed. The company has little experience in testing different models since it concentrated on a specific gap in the industry, but the platform allowed the company to differentiate the sets of services, where flexibility raises a an opportunity to reach new clients [37, 40].

The study reveals several barriers for business model transformation. Russian distributors had to buy expensive software from foreign firms and IT costs are usually considered as the main barrier [39, 41]. The forecast of the shift towards digital music consumption and timely internal development of the platform allowed Broma16 to transform successfully, but usually it is difficult to predict the time needed for radical restructuring [34]. Internal development is not a common approach since technology partnerships are more typical in the sector [37, 42]. In-house projects can often face significant time and money challenges. But the case proves the important benefits of in-house development. It allowed the company to enter new technological markets, providing similar services for other firms. Heaven11 presented entirely new technology, where the lack of similar solutions and competitors allowed them to perform successfully. But such a case is rarely seen. Usually technology companies capture creative markets and control infrastructure for distribution, not vice versa [38, 41].

The second barrier connected with conservatism in the value chain is not so common, since customers are used to react rapidly on technological developments [36, 42, 43]. Broma16 provides educational services for clients, but in the short-term it could be useless due to the early stage of market development and the lack of demand [35]. But firms can focus on foreign markets, where it is easier to find partners and launch digital projects [37, 40]. Broma16 followed a similar approach. The results of the case study mostly follow the theory. Digitalization, driven by the external development of the industry, caused changes in the foundational layer of the business model. Broma16 changed its key product and concentrated on music distribution rather than music promotion, as it first planned. The firm faced problems with high IT costs, but this was overcome through internal development, which is not a typical path. It made it possible to establish the external company Heaven11 and to enter new technological markets. To deal with conservatism in the value chain, the company provided educational services and shifted its focus to foreign markets, which is a wide-spread practice.

Considering the experience of the music distribution company Broma16, several conclusions can be drawn regarding the influence of digitalization on a firm's business model and the role of the company in the value chain.

Consequence 1: Music distributors are technological companies, rather than music companies.

Digital technologies formed the key infrastructural part in the business models in creative industries [10]. The core platform for the Broma16 business model was developed by Heaven11, that can deliver products for other distributors. Hence, Heaven11 is the key company for the distribution role, since it technologically ensures operation of the value chain. Digitalization is the reason for the growing importance of IT capabilities in intermediary companies, where technology firms can perform mediator roles successfully [22, 41]. This raises new challenges towards considering distributors as technological companies, rather than music companies. However, innovations require expertise in the area to choose the right moment for the new solution launch, identify demand and prospects for the implementation. Digitalization contributes to new markets based on the combination of industry-specific and technological knowledge. For example, in the case of Broma16 it created a new intermediary "layer" of companies that provide technological services for distributors.

Consequence 2: Music distribution relies on digital instruments for management and marketing.

Since music has become digital, companies have to use digital instruments for the distribution and related services. In Broma16, all communications, management practices and statistics have been transferred to the platform. Artists can collect data on the vast number of metrics (skip rate, shares, saves, and other responses by the user) to evaluate performance [18]. Labels can establish data-driven marketing strategies [17, 20]. However, this poses additional pressure on producers [13]. Artists have to use digital marketing tools, develop social media, analyze data and metrics collected and counted by Broma16. They also need to develop a specific non-creative set of skills, what is called a "shift from musician-as-artist to musician-as-entrepreneur" [14]. However, this is relevant only for independent musicians when major artists transmit their work to labels [16]. As earlier it depends on the volume of resources (time, money) an artist or label is ready to invest in advertising.

Consequence 3: Local intermediaries get more opportunities to enter foreign markets, but they have to perform innovations new to these markets.

Digitalization leads to the globalization of creative industries, and it is easier for local companies to perform successfully in international markets [37, 40]. In the presented case study, both Broma16 and Heaven11 had to innovate and introduce new products for foreign markets to enter them. The platform for music management and common database for all songs was innovative for the Russian market in 2012. iTunes was launched in Russia in the end of 2012, while the company Heaven11 was officially registered in the beginning of the year. Providing digital distribution services for Russian artists among the early adopters, Broma16 could identify the growing demand in other countries and enter those markets. But the platform was also innovative for foreign markets, where the market of streaming was emerging. Local intermediaries can get more opportunities to enter foreign markets, but in order to realize these opportunities they have to present innovative solutions for these markets.

Consequence 4: Music distributors operate at the complicated intersection of legal and technological aspects.

The main problem for distribution is the documentation of music, since numerous authors with neighboring rights can produce a song, performed in different sources both digitally and physically. Due to the lack of a single register of all songs with authors and sales (plays), it is very hard to control the copyright on digital content. Distribution companies have to find solutions for the afore-mentioned problems. Distributors are required to keep a register of all songs, with all copyright holders, and identify all streams, plays and sales for every song. Companies have to regularly update this register for thousands of songs to distribute royalties. Broma16 has automated most of these processes within the digital platform for music management. At the same time, distributors need to control the copyrights of clients and protect them, pay taxes on revenues and royalties, considering local legislative systems. Often distributors adopt digital technologies for issues posed by digital distribution and legislation, rather than music, and operate at the intersection of these aspects.

To sum up, the hypothesis of the paper was confirmed. Distributors provide musicians with digital instruments for music management and promotion, as music consumption has moved to digital services. This allows us to consider distributors as technological companies. At the same time, distributors have to deal with vast amounts of data and legal issues to provide high-quality services to customers.

Conclusion

This paper identifies managerial practices towards digitalization in the Russian music distributor Broma16. Stimulated by developments in the external environment, internal transformation faced a lack of IT investments and skills among partners in the value chain. To overcome these barriers, one needs internal development of the digital solution through the involvement of IT specialists, refocusing on new technological and foreign markets, and delivery of educational services. The study derives four major consequences of digital transformation for the firm's business model and role in the value chain:

- the firm can be considered as a technological company, rather than a music company;
- music distribution heavily relies on digital instruments for management and marketing;
- local intermediaries get more opportunities to enter foreign markets, but they have to perform innovations new to these markets;
- music distributors operate at the complicated intersection of technological and legislative issues.

Within the previous experience, analyzed through the review of literature, the current research applies a theoretical framework for the study of digitalization of music distributors, which is not very common among researchers, especially in Russia. This paper covers the motivation of key stakeholders towards the transformation of the business model, key drivers and barriers, strengths and threats for the Broma16 business. A similar approach can be used to study digital transformation on the corporate level in other creative industries for comparative analysis.

The practices towards digital transformation management applied by one of the most innovative music distribution companies in Russia are presented. Since the research was theory-led, barriers, company faced, and actions taken to overcome them, were considered according to the previous experience. The paper shows how senior management has changed the motivation and goals of the company to reorient it in a new market niche for digital distributors. Industry forecast and proactive in-house development of a digital platform allowed them to capture the market niche and enter foreign markets. Music distributors in emerging markets and firms in other creative industries can get some insights and apply them to avoid critical lock-ins.

This study can be useful for practitioners who manage digitalization in creative industries. The consequences identified can also be relevant for companies where streaming markets rapidly emerge right now, especially for companies in Latin America, for example Brazil, Argentina, Mexico, Paraguay [7, 26]. The reviewed experience may be interesting for Asian countries (South Korea, China, Japan) that try to establish domestic leaders instead of leaving the market free for Western firms, Apple, Spotify and other firms providing different services along the value chain. These companies are focusing on Russia, but domestic firms show they have advantages [4].

The study has several limitations. Firstly, it is based on a single case, which does not allow us to derive conclusions about development that will cover all aspects of the industry. Researchers can use the same theoretical framework to conduct multiple case studies and get some insights about linkages and differences between companies, which will allow them to better understand the development of the industry. The second limitation arises from the inability to conduct in-depth interviews with employees of companies because of NDA agreements. To mitigate the barrier, further studies can use the same framework to conduct workshops with industry representatives for in-depth examination of the changes stimulated by digitalization. ■

References

1. Polyakova V., Fursov K. (2021) *Digital practices of people in Russia in a period of self-quarantine*. ISSEK series of regular bulletins "Digital Economy" (in Russian). Available at: https://issek.hse.ru/news/438496284.html (accessed 13 July 2021).

2. *Entertainment and media industry overview: Forecast for 2019–2023* (2019) PwC (in Russian). Available at: https://www.pwc.ru/ru/publications/media-outlook/mediaindustriya-v-2019.pdf (accessed 27 March 2021).

3. *A roadmap toward a common framework for measuring the digital economy. Report for the G20 digital economy task force* (2020) OECD. Available at: https://www.oecd.org/sti/roadmap-towards-a-common-framework-for-measuring-the-digital-economy.pdf (accessed 07 April 2021).

4. *Global music report* (2021) IFPI. Available at: https://www.ifpi.org/wp-content/uploads/2020/03/GMR2021_STATE_OF_THE_INDUSTRY.pdf (accessed 27 July 2021).

5. *The show must go on* (2020) Goldman Sachs. Available at: https://www.goldmansachs.com/insights/pages/infographics/music-in-the-air-2020/report.pdf (accessed 07 April 2021).

6. *The Russian market of music streaming services will continue to grow rapidly* (2019) J'Son & Partners (in Russian). Available at: https://www.sostav.ru/publication/ostalnoe-vidimost-po-prognozu-j-son-and-partners-rossijskij-rynok-muzykalnykh-strimingovykh-uslug-prodolzhit-bystryj-rost-37158.html (accessed 27 March 2021).

7. *Music consumer insight report* (2018) IFPI. Available at: https://www.ifpi.org/wp-content/uploads/2020/07/091018_Music-Consumer-Insight-Report-2018.pdf (accessed 27 March 2021).

8. Herbert D., Lotz A.D., Marshall L. (2019) Approaching media industries comparatively: A case study of streaming. *International Journal of Cultural Studies*, vol. 22, no. 3, pp. 349–366. https://doi.org/10.1177/1367877918813245

9. Galuszka P. (2015) Music aggregators and intermediation of the digital music market. *International Journal of Communication*, vol. 9, pp. 254–273.

10. Li F. (2020) The digital transformation of business models in the creative industries: A holistic framework and emerging trends. *Technovation*, vols. 92–93, article ID 102012. https://doi.org/10.1016/j.technovation.2017.12.004

11. Bonini T., Gandini A. (2019) "First week is editorial, second week is algorithmic": Platform gatekeepers and the platformization of music curation. *Social Media + Society*, vol. 5, no. 4, article ID 2056305119880006. https://doi.org/10.1177/2056305119880006

12. Hracs B.J. (2015) Cultural intermediaries in the digital age: The case of independent musicians and managers in Toronto. *Regional Studies*, vol. 49, no. 3, pp. 461–475. https://doi.org/10.1080/00343404.2012.750425

13. Morris J.W. (2015) Curation by code: Infomediaries and the data mining of taste. *European Journal of Cultural Studies*, vol. 18, nos. 4–5, pp. 446–463. https://doi.org/10.1177/1367549415577387

14. Eiriz V., Leite F.P. (2017) The digital distribution of music and its impact on the business models of independent musicians. *The Service Industries Journal*, vol. 37, nos. 13–14, pp. 875–895. https://doi.org/10.1080/02642069.2017.1361935

15. Koh B., Hann I.-H., & Raghunathan, S. (2019) Digitization of music: Consumer adoption amidst piracy, unbundling, and rebundling. *MIS Quarterly*, vol. 43, no. 1, pp. 23–45. https://doi.org/10.25300/MISQ/2019/14812

16. Vonderau P. (2019) The Spotify effect: Digital distribution and financial growth. *Television & New Media*, vol. 20, no. 1, pp. 3–19. https://doi.org/10.1177/1527476417741200

17. Kjus Y. (2016) Musical exploration via streaming services: The Norwegian experience. *Popular Communication*, vol. 14, no. 3, pp. 127–136. https://doi.org/10.1080/15405702.2016.1193183

18. Morgan B.A. (2020) Revenue, access, and engagement via the in-house curated Spotify playlist in Australia. *Popular Communication*, vol. 18, no. 1, pp. 32–47. https://doi.org/10.1080/15405702.2019.1649678

19. Street J., Laing D., Schroff S. (2018) Regulating for creativity and cultural diversity: The case of collective management organisations and the music industry. *International Journal of Cultural Policy*, vol. 24, no. 3, pp. 368–386. https://doi.org/10.1080/10286632.2016.1178733

20. Hesmondhalgh D., Jones E., Rauh A. (2019) SoundCloud and Bandcamp as alternative music platforms. *Social Media + Society*, vol. 5, no. 4, article ID 2056305119883429. https://doi.org/10.1177/2056305119883429

21. Fleischer R. (2017) If the song has no price, is it still a commodity? Rethinking the commodification of digital music. *Culture Unbound*, vol. 9, no. 2, pp. 146–162. https://doi.org/10.3384/cu.2000.1525.1792146

22. Meier L.M., Manzerolle V.R. (2019) Rising tides? Data capture, platform accumulation, and new monopolies in the digital music economy. *New Media & Society*, vol. 21, no. 3, pp. 543–561. https://doi.org/10.1177/1461444818800998

23. Burmester A.B., et al. (2016) Accepting or fighting unlicensed usage: Can firms reduce unlicensed usage by optimizing their timing and pricing strategies? *International Journal of Research in Marketing*, vol. 33, no. 2, pp. 343–356. https://doi.org/10.1016/j.ijresmar.2015.06.005

24. Mróz B. (2016) Online piracy: An emergent segment of the shadow economy. Empirical insight from Poland. *Journal of Financial Crime*, vol. 23, no. 3, pp. 637–654. https://doi.org/10.1108/JFC-04-2015-0022

25. Mazziotti G., Simonelli F. (2016) Another breach in the wall: copyright territoriality in Europe and its progressive erosion on the grounds of competition law. *Info*, vol. 18, no. 6, pp. 55–66. https://doi.org/10.1108/info-06-2016-0026

26. *Year-End Report* (2021) MRC Data, Billboard. Available at: https://www.musicbusinessworldwide.com/files/2021/01/MRC_Billboard_YEAR_END_2020_US-Final.pdf (accessed 27 July 2021).

27. Gerring J. (2004) What is a case study and what is it good for? *American political science review*, vol. 98, no. 2, pp. 341–354. https://doi.org/10.1017/S0003055404001182

28. Huyghe A., et al. (2014) Technology transfer offices as boundary spanners in the pre-spin-off process: The case of a hybrid model. *Small Business Economics*, vol. 43, no. 2, pp. 289–307. https://doi.org/10.1007/s11187-013-9537-1

29. Simons H. (2014) Case study research: In-depth understanding in context. *The Oxford handbook of qualitative research*, edited by Leavy P., pp. 455–470. https://doi.org/10.1093/oxfordhb/9780199811755.013.005

30. Seawright J., Gerring J. (2008) Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political research quarterly*, vol. 61, no. 2, pp. 294–308. https://doi.org/10.1177/1065912907313077

31. *Broma16* (2021) Official website. Available at: https://broma16.com/en/ (accessed 07 April 2021).

32. *Heaven11* (2021) Official website. Available at: https://www.heaven11.pro/#intro (accessed 07 April 2021).

33. Thomas G. (2011) A typology for the case study in social science following a review of definition, discourse, and structure. *Qualitative inquiry*, vol. 17, no. 6, pp. 511–521. https://doi.org/10.1177/1077800411409884

34. Øiestad S., Bugge M.M. (2014) Digitisation of publishing: Exploration based on existing business models. *Technological Forecasting and Social Change*, vol. 83, no. 1, pp. 54–65. https://doi.org/10.1016/j.techfore.2013.01.010

35. Landoni P. et al. (2020) Business model innovation in cultural and creative industries: Insights from three leading mobile gaming firms. *Technovation*, vols. 92–93, article ID 102084. https://doi.org/10.1016/j.technovation.2019.102084

36. *Business Motivation Model: Version 1.3* (2015) Object Management Group. Available at: https://www.omg.org/spec/BMM/1.3/PDF (accessed 27 July 2021).

37. Towse R. (2017) Economics of music publishing: Copyright and the market. *Journal of Cultural Economics*, vol. 41, no. 4, pp. 403–420. https://doi.org/10.1007/s10824-016-9268-7

38. Lee J.-A. (2019) Tripartite perspective on the copyright-sharing economy in China. *Computer Law and Security Review*, vol. 35, no. 4, pp. 434–452. https://doi.org/10.1016/j.clsr.2019.05.001

39. Kim H.-J., Kim B.-H. (2018) Implementation of young children English education system by AR type based on P2P network service model. *Peer-to-Peer Networking and Applications*, vol. 11, no. 6, pp. 1252–1264. https://doi.org/10.1007/2fs12083-017-0612-2

40. Lyubareva I., Benghozi P.-J., Fidele T. (2014) Online business models in creative industries: Diversity and structure. *International Studies of Management and Organization*, vol. 44, no. 4, pp. 43–62. https://doi.org/10.2753/IMO0020-8825440403

41. Nakano D., Fleury A. (2017) Recorded music supply network reconfiguration: The dual effect of digital technology. *International Journal of Manufacturing Technology and Management*, vol. 31, nos. 1–3, pp. 153–175. https://doi.org/10.1504/IJMTM.2017.10002918

42. Benghozi P.-J., Paris T. (2014) The cultural economy in the digital age: A revolution in intermediation? *City, Culture and Society*, vol. 7, no. 2, pp. 75–80. https://doi.org/10.1016/j.ccs.2015.12.005

43. Dobusch L., Schüßler E. (2014) Copyright reform and business model innovation: Regulatory propaganda at German music industry conferences. *Technological Forecasting and Social Change*, vol. 83, no. 1, pp. 24–39. https://doi.org/10.1016/j.techfore.2013.01.009

Appendix 1.
Background materials for interview

Source	Internet link
About the company	https://broma16.com/en/about/
Technology at the company	https://broma16.com/en/technology/
Services provided by the company	https://broma16.com/en/
News of the company	https://broma16.com/en/news/
Roster of the company	https://broma16.com/en/roster/artists/featured/
Accounts of the company in social media: Facebook, VKontakte	https://www.facebook.com/BroMa16/ https://vk.com/broma16
Presentation of the company's CEO at the conference "Coliseum 2019"	https://www.youtube.com/watch?v=NWCg3N_uEvw

Appendix 2.
Interview guideline

The interview was conducted as part of the non-commercial research. The purpose of the interview was to understand how the functions of distribution companies have been changed under digitalization. The interview was held in the online format. The length of the interview did not exceed two hours. Information about respondents was not disclosed, except for the employee's position and length of service. Raw data collected within the interview was not the subject of publication.

Part 1. Information about the respondent

1. Employee’s position:
2. Length of service in the company

Part 2. Technological development of the company

3. What role do information and digital technologies play in the company’s core business?

4. How has this role changed within the development of the company and industry?

5. How are information and digital tools used for everyday tasks?

6. Have there been any barriers or difficulties in implementing new information and digital technologies? How did the company solve them?

Part 3. Interaction with customer

7. How do the company’s relationships with its customers develop?

8. How are information and digital tools used to provide services to customers?

9. Did the technological development of the company/industry stimulate changes in the set of services?

Part 4. Future development

10. How often does the company face the necessity to implement new digital solutions? What are the reasons for this necessity?

11. Is there a technology on the market that the company will be required to implement in the near future?

12. What will be the drivers and barriers for implementing a new solution in the near future?

About the author

Artem I. Altynov

Research Assistant, Laboratory for Science and Technology Studies, Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, 11, Pokrovsky Bulvar, Moscow 109028, Russia;

E-mail: aaltynov@hse.ru

ORCID: 0000-0001-9209-438X

DOI: 10.17323/2587-814X.2022.1.72.82

Customer segmentation using *k*-means clustering for developing sustainable marketing strategies

Nidhi Gautam [a]
E-mail: nidhi.uiams@pu.ac.in

Nitin Kumar [b]
E-mail: nk2268913@gmail.com

[a] University Institute of Applied Management Sciences, Panjab University
Address: South Campus, Behind P.U. Alumni House Udyog Path, Punjab University Rd, Sector 25,
Chandigarh 160014, India

[b] Housing Development Finance Corporation Limited
Address: Dalhousie Road, near Shani Dev Mandir, Pathankot, Punjab 145001, India

Abstract

Sales and marketing is the indispensable department of an organization which leads to the generation of revenue and building customer relationship. Marketing is the process of finding the potential customers and sales is the process of converting those potential customers into real customers. Hence, it is imperative that marketing and sales go hand in hand. Developing marketing strategies needs proper market research which can cover the relevant pointers like demographics, culture, spending power, income and many more. The process of segmentation, targeting and positioning (STP) is carried out to develop marketing and sales strategies. STP is done by collection of the marketing intelligence. For this process, surveys are also used but data mining has far more effective and better results so far. Organizations tend to take risk because of the importance and relevance of the marketing and sales department. Most of the budget in the organizations is allocated for marketing and promotional activities. For making data-driven and accurate decisions, data mining is used in various fields to extract valuable information and patterns. This paper discusses the use of the data mining concept on marketing. This paper aims to analyze marketing data with *k*-means data mining clustering techniques and to find the relationship between marketing and *k*-means data mining clustering techniques.

Keywords: marketing, *k*-means, clustering, data mining tools, data mining techniques

Citation: Gautam N., Kumar N. (2022) Customer segmentation using *k*-means clustering for developing sustainable marketing strategies. *Business Informatics*, vol. 16, no. 1, pp. 72–82. DOI: 10.17323/2587-814X.2022.1.72.82

Introduction

Marketing is a value creation process according to Philip Kotler [1]. It is all about what you deliver to your customer. People consider marketing and sales as the same but both are different and have their own relevance and impact in the organization. Marketing is the process of finding the potential customer and sales is the process of converting those potential customers into real customers [2]. The concept of marketing can be explained using different marketing concepts. First, there is a production concept, which explains that supply will create the demand. Second, the sales concept emphasis on selling the product by hook or by crook. Third, the product concept gives importance to product innovation and differentiation. Fourth, the marketing concept, which believes in catering to the needs of the customer as customer, is king for them. Lastly, the societal marketing concept focuses on the customer as well as society and the environment. It totally depends upon the organization to decide which concept it wants to follow. Whenever anyone considers marketing strategies, marketing mix is used. According to Philip Kotler, marketing mix comprises four components i.e., product, price, place and promotion [1].

Data mining is a process of extracting knowledge from large amounts of data. It is a technique to find trends, patterns, correlations, anomalies in databases which can be helpful to make accurate future decisions. It is also known as knowledge discovery in databases (KDD) results. Data mining helps experts to understand the data and leads to better and data driven decisions. Data mining is an intersection of three fields: databases, artificial intelligence and machine learning. The steps of data mining includes data cleansing (for the removal of noisy and inconsistent data), data integration (to combine data from multiple sources for efficient data processing), data selection (to select and retrieve the relevant data for analysis), data transformation (to transform the collected data into a desirable form for further data processing), data mining (a technique applied on data for various pattern matching methods), pattern evolution (to determine important patterns which represent knowledge and data insights) and knowledge presentation (this is done by various visualization methods for knowledge representation pictorially or graphically). The various applications of data mining include market basket analysis (association mining). Market basket analysis is the method to discover relations or correlations among the set of data items. Classification analyzes a training set of objects with known labels and tries to form a model for each class based on the features in the data. Regression is to predict values of some missing data or to build a model for some attributes using other attributes of the data. Time Series Analysis analyzes time series data to find certain regularities and interestingness in data. Clustering is used to identify clusters embedded in the data. The task of clustering is to find clusters for which intra-cluster similarity is high and inter-cluster similarity is low. Outlier analysis is used to find outliers in the data, namely detect data which are very far away from average behavior of the data [3].

This paper focuses on the importance of data mining approaches, specifically clustering, for formalizing the marketing strategies by understanding customer needs and spending behavior. The paper highlights the importance of visualization tools to understand the important relationships between various parameters in the dataset. It shows that how a pair-plot can be helpful in identifying clusters and the silhouette score for deciding the *k* value for *k*-means clustering method.

The rest of the paper is organized as follows: Section 1 presents the motivation and research rationale of the study. Section 2 describes numerous existing models and related work using various clustering methods. Section 3 presents the research design and methodology followed for this work. Section 4 describes the results and discussions. The last Section concludes the paper with future recommendations.

1. Motivation and research rationale

This study aims to work on the transaction dataset of a store which is taken from an internet source [4] for making clusters of customers depending upon their income and spending score. This will lead to segmentation, targeting and positioning (STP) of the customers and their behavior patterns will help us to develop sustainable marketing strategies. Nowadays, everything is connected to the customer with the help of the Internet of things. You just need space to store data to perform analysis to keep track of the customers. This study can help organizations to gain new customers and maintain loyal customers. The main contributions of this paper are summarized as follows:

- to perform market segmentation of the customers depending upon their spending using unsupervised machine learning;
- to perform STP;
- to know the target customer and develop marketing strategies.

2. Existing models with literature review

Association rule mining is a data-mining concept which is used to optimize the patterns associated with dynamic behaviors of transactions made by customers when purchasing some specific products. The insights generated from this technique can be used by the retailing business for making data-driven decision-making. Using this algorithm, the frequent transactions made by the customers have been analyzed using the support and confidence of the customers in buying associated items. The analysis conducted by Association rule mining model can best be used in managing product placement on the shelves in the supermarket [5–7]. The study was conducted in order to make a market basket analysis by using association rules. The data used in the study was the sales data of a supermarket from the Vancouver Island University website. Data was analyzed in the Weka tool where the dataset contained 225 different products for analysis [8].

The study focused on small and medium enterprises (SMEs), where customer behavior was analyzed from the perspective of SMEs. The model proposed the integration of customer relationship management (CRM) and the data mining techniques to provide effective rules and new patterns for better decision-making. The model suggested that as the era of big data is unwinding itself, the data mining application may enhance accuracy of rules and patterns, all of which can further help the enterprises to improve customers' satisfaction and loyalty, reduce customer churn and so on. The suggested models can also help SMEs to classify the priority customer groups and offer them better facilities and ranges to retain them. The suggested models can help the enterprises to further improve their market share, position in the market and maintain a positive development process [9–11].

The model incorporated a new graphical display for clustering techniques. In this model, each cluster is represented by a silhouette. The silhouette is based on the comparison of its tightness and separation. This silhouette showed which objects lie well within their cluster and which ones are merely somewhere in between clusters. The whole clustering is displayed by combining the silhouettes into a single plot. The average silhouette width provides an evaluation of clustering validity and may help to select an ‘appropriate’ number of clusters [12–13].

Classification of aquifer vulnerability using *k*-means cluster analysis uses the application of the cluster analysis in ground vulnerability assessment using the *k*-means technique. In this study, a clustering technique is used because it removes some of the subjectivity associated with the indexing method. It creates a vulnerability map that does not rely on fixed weights and ratings and provides a more objective representation of the system's physical characteristics. The model was applied to an aquifer in Iran and compared with the standard DRASTIC approach using the water quality parameters nitrate, chloride and total dissolved solids (TDS) as surrogate indicators of aquifer vulnerability. The model having clustering techniques outperformed the other methods [14–16].

The paper discovered the segments of organic food consumers in Lebanon by using a market segmentation based on lifestyle and attitude variables to generate appropriate marketing strategies for each market segment [17]. Market basket analysis (MBA) is a very powerful data mining technique which provides various types of information, like buying behavior of the customer, likes, dislikes, etc. to the retailer, all of which can help the retailer perform correct decision-making. It can be used in various fields, such as marketing, management, bioinformatics, the education field and many more. MBA is a very useful technique to find out interesting patterns from a large amount of data which can automatically track any type of changes in facts from previous data [18–20].

The authors used a *k*-means clustering algorithm to identify the customer segmentation in a supervised manner. The methodology could understand the complex relationships existing in the data attributes [21]. The authors analyzed the data of an e-commerce portal to understand the requirements of the customers so as to provide them better services in the future. A *k*-means clustering algorithm was used to analyze the data, considering customer segmentation as an important aspect of it [22–23]. The authors analyzed data of three online food chains and applied various clustering algorithms on the same. Though there is no fixed model of a particular algorithm which could show best results, *k*-means clustering has shown promising results on the data [24]. The authors suggested how customer segmentation can be implemented and can be useful to understand the customers. Customer segmentation can be a stepping-stone for identifying future prospective customers and specific marketing strategies can be formulated considering the customer segment [25]. The authors have stressed the use of machine learning algorithms for customer segmentation, since it can enhance productivity and profitability of an organization. *K*-means clustering was used for the study and it has shown very promising results for customer segmentation [26].

3. Research design and methodology (data collection method)

A. Sampling method. Secondary data, mall customer segmentation dataset from an internet source [4].

B. Sample size. The dataset which is used for the analysis contains people's purchasing attributes in the Malls. This dataset has five features – customerID, age, gender, credit score and income. There are data for about 200 transactions used for data analysis.

C. Research rationale. To devise a comprehensive model that can be used to classify customers based on spending score and annual income for developing appropriate marketing strategies.

D. Tools used. R Studio [27], Weka [28], MS Excel [29].

R Studio and Weka are freeware which are used for data analytics. These tools are mostly used for data analytics in the field of industry as well as academia. MS Excel is a sub-tool of Microsoft Office, which is generally used for data preparation and preprocessing.

E. Clustering algorithms used. The *k*-means clustering algorithm [24] is based on the Euclidian distance to figure out k clusters in the data. The clusters are homogeneous within them and represent similar types of data. *K*-means clustering is most suitable to handle big and hyper spherical data. It is best suited for market segmentation, social network analytics, image segmentation and so on.

4. Results and discussion

The dataset is preprocessed and cleansed in Excel [29] by using descriptive statistics. The dataset is balanced as the distribution of males and females are almost the same. The distribution of data is checked in Weka [28] and pair plots are generated to show the same.

From the pair plot in *Fig. 1*, we found that the last row is insightful since it gives an indication of hidden clusters in the data. There is cluster formation between the Spending score (1–100) vs. CustomerID, the Spending score vs. Age and the Spending score (1–100) vs. Annual income (in thousands of US dollars).

The joint-plots, distributions, correlation matrix, silhouette coefficient and *k*-means clusters are generated in R-Studio [27] by using its libraries. The joint-plot of spending score and age as shown in *Fig. 2a* shows that there are two bright core areas where density is very

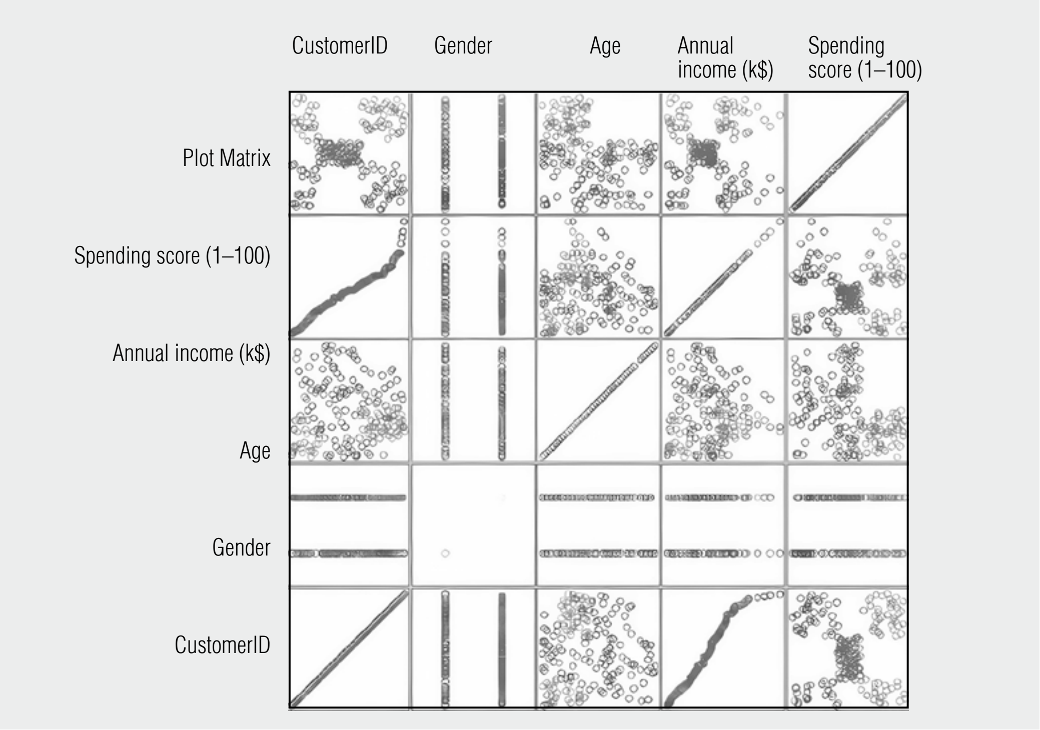

Fig. 1. Pair–plot of variables representing a pairwise relationship among CustomerID, Age, Annual Income (in thousands of US dollars) and Spending Score (1–100).

high. There are two different spending habits in different age groups represented in the plot.

The joint-plot of the spending score and annual income as shown in *Fig. 2b* shows that there is one area where density is very high i.e., in the middle. The other four areas show different patterns for the user. This may be possible because of the different purchasing power of the customers and their different spending habits. The five groups from the observations include Low income & High spending habits, Low income & Low spending habits, Moderate income & Moderate spending habits, High income & High spending habits and High income & Low spending habits. Joint-plot of Annual income and Age as shown in *Fig. 2c* and *Fig. 3* shows that the people in mid-30s have roughly a mean income of $80 000. If we compare the yearly income of more than $100 000, we find that males have more than $100 000 income in their early 30s and in the case of females it's around mid-40s. Perhaps, this is due to the disparity in pay.

The average value of the spending score of females is slightly more than that of the males. Notice the bulge of the graph in *Fig. 3*, which shows the mean value. *Figure 4* shows the Annual income of Males and Females. Thus, the average income of females (Female = 0) is less than that of males (Male = 1).

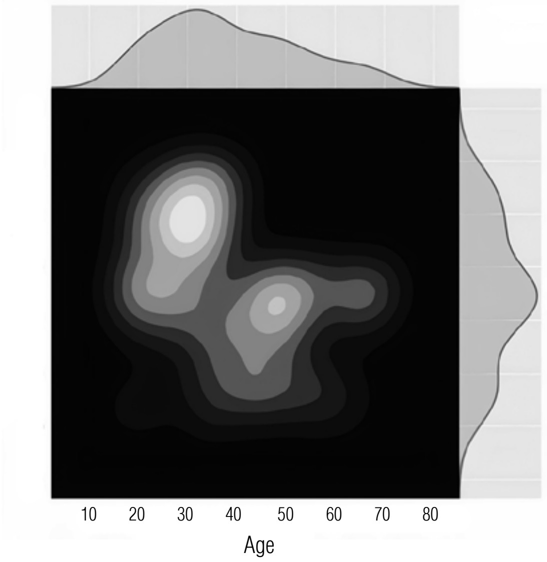

Fig. 2a. Joint plots for describing (Spending score and Age) distributions on the same plot.

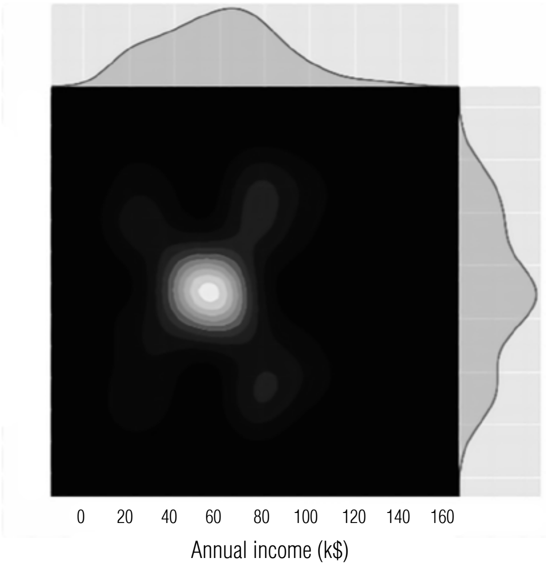

Fig 2b. Joint plots for describing (Spending score and Annual income) distributions on the same plot

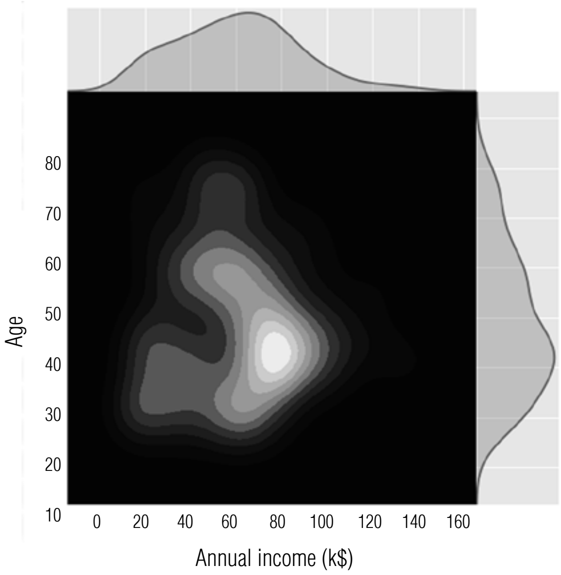

Fig 2c. Joint plots for describing (Annual income and Age) distributions on the same plot.

The distribution of the data showed that most of the people are less than 45 years of age. For

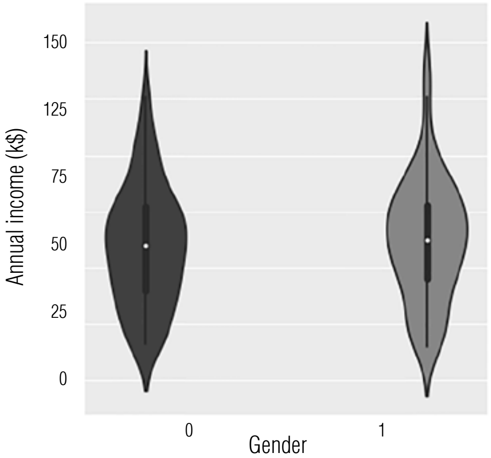

Fig. 3. Annual income (in thousands of USD) vs. Gender.

most of the customers, the spending score centers between 40 and 60. Annual income and gender have a positive correlation. There is a positive correlation between annual income and the spending score as shown in *Fig. 5.*

Customer Segmentation with *k*-means using the Silhouette Score. The Silhouette coefficient in *k*-means clustering is calculated using the mean intra-cluster distance "*a*" and the mean nearest-cluster distance "*b*" for each sample. The Silhouette coefficient for a sample is $(b - a) / \max(a, b)$ where *b* is the distance between a sample and the nearest cluster that the sample is not a part of. Note that the Silhouette coefficient is only defined if the number of labels is $2 \le n\_labels \le (n\_samples - 1)$. Trying to find clusters based on features like Annual income & Spending score. Since the Silhouette score is maximum for $k = 5$, it is a good idea to cluster the data into five subgroups.

The Silhouette coefficient has helped us to decide on the number of clusters to be taken for the study. In *k*-means clustering, selecting the value of *k* is of utmost importance. By using the Silhouette coefficient, selecting the value *k* has become very clear and simple. From the plot in *Fig. 6* there are five clusters: *cluster A, cluster B*, *cluster C*, *cluster D* and *cluster E.* Customers of the *cluster E* are misers as they have more purchasing power but they have a lower spending score. Customers of the *cluster A, cluster B* and *cluster C* are easy to handle. The customers of *cluster D* groups are a threat to the organization because they have less income but a larger spending score. They can be defaulters in the future. Therefore, by using *k*-means clustering and the Silhouette coefficient, customer classifications can be easily visualized. Hence, the marketing teams can easily identify potential customers.

Conclusion

Nowadays, it is very crucial to identify your potential customers in order to have a more data driven strategy to target customers. From the above data analysis, it is concluded that the distribution of males and females is nearly same. The pair-plot helps us to explain the permutation of the attributes, which helps in pattern, or cluster identification. With the help of the joint-plot between the spending score and annual income, purchasing capacity as well as spending habits of the customers can be analyzed. It is observed from the *k*-means clustering that customers can be classified into five groups such as "Low income & High spending habits," "Low income & Low spending habits," "Moderate

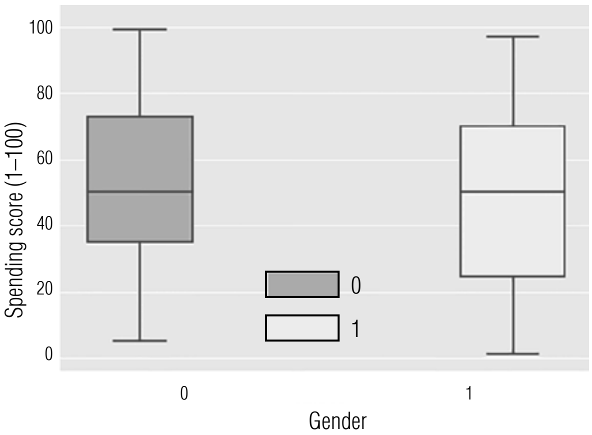

Fig. 4. Box–plot for Spending score (1–100) vs. Gender.

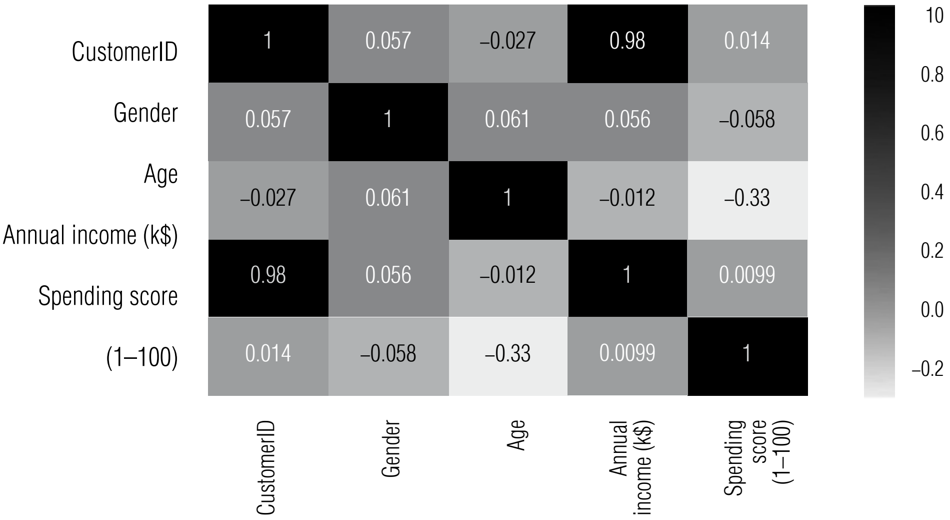

Fig. 5. Correlation matrix to show the correlation among variables such as customerID, Gender, Age, Annual income, Spending score.

income & Moderate spending habits," "High income & High spending habits" and "High income & Low spending habits." The box-plot for Spending score and Gender shows that the mean Spending scores of females are slightly more than that of the males whereas the violin plot shows that the average income of females is less than that of males. In this data, the majority of the people are less than 45 years of age and most of the customers, spending score centers between 40 and 60. Annual income and gender have positive correlation. There is a positive correlation annual income and spending score. The segmentation of the customers is done by a *k*-means algorithm. The value of the *k* is decided by the Silhouette score. We tried to find clusters based on features like Annual income & Spending score. The algorithm divided the data into five subgroups from which one could formulate marketing strategies in order to sell their products to the target audience.

In this study, the following problems have been resolved:

- to perform STP;
- to know the target customer and develop marketing strategies;
- to solve this problem, the dataset of a retail store was taken and we performed exploratory data analysis by using data visualization using various plots starting from the pair plot to the *k*-means plot.

The pair plot explains the relationships and patterns, which act as a stepping-stone for further analysis. There is cluster formation between Spending score (1-100) vs. CustomerID, Spending score vs. Age and Spending score (1-100) vs. Annual income (in thousands

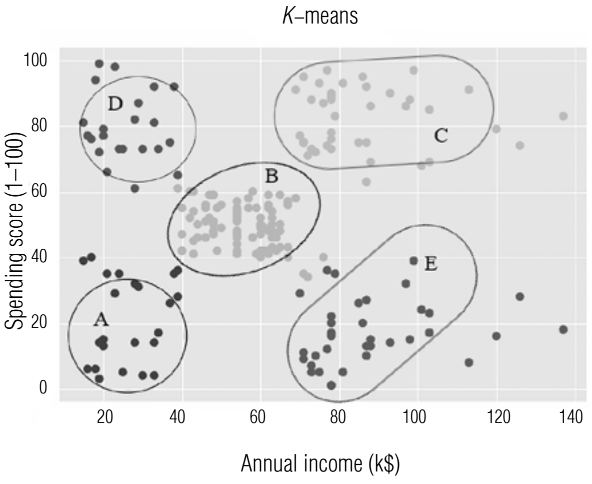

Fig. 6. K–means clusters for customer classification.

of USD) and these attributes were taken up for the study using joint plots. The joint-plot of Spending score and Annual income shows that there is one area where density is very high i.e., in the middle. The other four areas show different patterns for customers. This may be because of the different purchasing capacity of the customers and their different spending habits. The five groups from the observations includes Low income & High spending habits, Low income & Low spending habits, Moderate income & Moderate spending habits, High income & High spending habits and High income & Low spending habits.

The data visualization between Spending score and Annual income was presented. To know the accurate number of clusters, the Silhouette coefficient was used and its value came out to be five in this dataset, i.e. five clusters were considered for this study. Customers in *cluster E* have an annual income but they are spending very less from their income because their spending score is very low. This is a very important segment because they have the money to pay. A lot of marketing and promotional activities are required to influence this segment (group). A market basket analysis should be done to know the pattern associated with the purchase of goods. This can help us to know the better product placement of the product on the aisle. Marketing should be done in such a way that the consumers relate the product with their lifestyle; this will boost sales. A discount and offer can help the store to get more revenue. Customers who are in *cluster A, cluster B* and *cluster C* clusters have nearly the same annual income and spending score. These customers do not need more marketing and promotional activities. They can be tackled by the salesmen directly. Customers in *cluster D* are the ones who rely on the credit card because their annual income is very much less than their spending score. These customers have higher chances of becoming defaulters because their income is low but they are spendthrifts. The difference of outcomes in comparison with other competitive approaches and solutions is that our solutions are accurate because they have scientific and mathematical backup. The solution proposed in this paper is based on logic and data, not on gut feeling and experiences. The solutions are safer and more reliable than traditional methods of STP. These solutions can be applicable to SMEs and small business persons with low investment and help them to use their hard-earned money in a justified way. In this paper, an algorithm divided the data into five subgroups which could be used to formulate marketing strategies in order to sell products to the target audience. IT systems can be developed for the SMEs and small business for the management of the sales and marketing of the business so that they can allocate their limited resources and budget for maximum benefits. Data scientists have a crucial role in this field because it is very important to understand the data of a particular organization. The exponential generation of the data and the growth of artificial intelligence has given an opportunity to data scientists and marketing tycoons so they can come together and build an IT system at affordable rates which will enhance the business process. ■

References

1. Kotler P., Cunningham M.H., Keller K.L. (2008) *A framework for marketing management*. Toronto: Pearson Prentice Hall.
2. Rust R.T. (2020) Outside-in marketing: Why, when and how? I*ndustrial Marketing Management*, vol. 89, pp. 102–104. https://doi.org/10.1016/j.indmarman.2019.12.003
3. Gupta M.K., Chandra P. (2020) A comprehensive survey of data mining. *International Journal of Information Technology*, vol. 12, pp. 1243–1257. https://doi.org/10.1007/s41870-020-00427-7

4. Chaudhary V. (2018) *Mall customer segmentation data: Market basket analysis.* Data set. Available at: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python (accessed 2 March 2021).

5. Guney S., Peker S., Turhan C. (2020) A combined approach for customer profiling in video on demand services using clustering and association rule mining. *IEEE Access*, vol. 8, pp. 84326–84335. https://doi.org/10.1109/ACCESS.2020.2992064

6. Wang S., Zhang H., Chen H., Shi Q., Li Y. (2020) Association rule mining for precision marketing of power companies with user features extraction. *2020 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5, https://doi.org/10.1109/PESGM41954.2020.9281644

7. Ananda I., Salamah U. (2020) The application of product marketing strategy using an association rule mining apriori method. *International Journal of Information System and Computer Science*, vol. 4, no. 2, pp. 80–88. Available at: https://ojs.stmikpringsewu.ac.id/index.php/ijiscs/article/download/899/pdf (accessed 21 February 2021).

8. Ünvan Y.A. (2020) Market basket analysis with association rules. *Communications in Statistics – Theory and Methods*, vol. 50, no. 7, pp. 1615–1628. https://doi.org/10.1080/03610926.2020.1716255

9. Ranjan J., Bhatnagar V. (2008) Critical success factors for implementing CRM using data mining. *Interscience Management Review*, vol. 1, no. 1, article 7. https://doi.org/10.47893/IMR.2008.1006

10. Ngai E.W.T., Xiu L., Chau D.C.K. (2009) Application of data mining techniques in customerrelationship management: A literature review and classification. *Expert Systems with Applications*, vol. 36, no. 2, pp. 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

11. Hosseini S.M.H., Maleki A., Gholamian M.R. (2010) Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, vol. 37, no. 7, pp. 5259–5264. https://doi.org/10.1016/j.eswa.2009.12.070

12. Silva S., Cortez P., Mendes R., Pereira P.J., Matos L.M., Garcia L. (2018) A categorical clustering of publishers for mobile performance marketing. In: *The 13th International Conference on Soft Computing Models in Industrial and Environmental Applications*. Springer, Cham. https://doi.org/10.1007/978-3-319-94120-2_14

13. Beheshtian-Ardakani A., Fathian M., Gholamian M. (2018) A novel model for product bundling and direct marketing in e-commerce based on market segmentation. *Decision Science Letters*, vol. 7, no. 1, pp. 39–54. https://doi.org/10.5267/j.dsl.2017.4.005

14. Javadi S., Hashemy S.M., Mohammadi K., Howard K.W.F., Neshat A. (2017) Classification of aquifer vulnerability using *k*-means cluster analysis. *Journal of Hydrology*, vol. 549, pp. 27–37. https://doi.org/10.1016/j.jhydrol.2017.03.060

15. Rahmani B., Javadi S., Shahdany S.M.H. (2021) Evaluation of aquifer vulnerability using PCA technique and various clustering methods. *Geocarto International*, vol. 36, no. 18, pp. 2117–2140. https://doi.org/10.1080/10106049.2019.1690057

16. Jahwar A.F., Abdulazeez A.M. (2020) Meta-heuristic algorithms for k-means clustering: A review. *PalArch's Journal of Archaeology of Egypt / Egyptology*, vol. 17, no. 7, pp. 12002–12020. Available at: https://archives.palarch.nl/index.php/jae/article/view/4630 (accessed 21 February 2021).

17. Tleis M., Callieris R., Roma, R. (2017) Segmenting the organic food market in Lebanon: An application of k-means cluster analysis. *British Food Journal*, vol. 119, no. 7, pp. 1423–1441. https://doi.org/10.1108/BFJ-08-2016-0354

18. Kaur M., Kang S. (2016) Market basket analysis: Identify the changing trends of market data using association rule mining. *Procedia Computer Science*, vol. 85, pp. 78–85. https://doi.org/10.1016/j.procs.2016.05.180

19. Gupta S., Mamtora R. (2014) A survey on association rule mining in market basket analysis. *International Journal of Information and Computation Technology*, vol. 4, no. 4, pp. 409–414. Available at: http://ripublication.com/irph/ijict_spl/ijictv4n4spl_11.pdf (accessed 21 February 2021).

20. Aguinis H., Forcum L.E., Joo H. (2013) Using market basket analysis in management research. *Journal of Management*, vol. 39, no. 7, pp. 1799–1824. https://doi.org/10.1177/0149206312466147

21. Muhal H., Jain H. (2021) Two-stage customer segmentation using k-means clustering and artificial. *International Research Journal of Engineering and Technology*, vol. 8, no. 3, pp. 485–490.
22. Punhani R., Arora V.P.S., Sabitha S., Kumar Shukla V. (2021) Application of clustering algorithm for effective customer segmentation in E-Commerce. In: *Proceedings of the 2021 International Conference on Computational Intelligence and KnowledgeEconomy (ICCIKE), 17-18 March 2021, Dubai, United Arab Emirates*, pp. 149–154. https://doi.org/10.1109/ICCIKE51210.2021.9410713
23. Popović N., Savić A., Bjelobaba G., Veselinović R., Stefanović H., Ilić P.M. (2021) The implementation of hierarchical and nonhierarchical clustering for customer segmentation in one luxury goods company. In: *Proceedings of the 37th International Business Information Management Association (IBIMA), 30–31 May 2021, Cordoba, Spain*, pp. 8370–8381.
24. Aktaş A.A., Tunalı O., Bayrak A.T. (2021) Comparative unsupervised clustering approaches for customer segmentation. In: *Proceedings of the 2021 2nd International Conference on Computing and Data Science (CDS), 28-29 Jan. 2021, Stanford, CA, USA*, pp. 530–535. https://doi.org/10.1109/CDS52072.2021.00097
25. Suresh Y., Senthilkumar J., Mohanraj V., Kesavan S. (2021) Customer segmentation using machine learning in python. *Turkish Journal of Physiotherapy and Rehabilitation*, vol. 32, no. 3, pp. 4338–4342. Available at: https://turkjphysiotherrehabil.org/pub/pdf/321/32-1-521.pdf (accessed 21 February 2021).
26. Pradana M., Ha H. (2021) Maximizing strategy improvement in mall customer segmentation using *k*-means clustering. *Journal of Applied Data Sciences*, vol. 2, no. 1, pp. 19–25. https://doi.org/10.47738/jads.v2i1.18
27. RStudio Team (2020) *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA. Available at: http://www.rstudio.com/ (accessed 21 February 2021).
28. Ngo T. (2011) Data mining: practical machine learning tools and technique, Third Edition by Ian H. Witten, Eibe Frank, Mark A. Hell. *ACM SIGSOFT Software Engineering Notes*, vol. 36, no. 5, pp. 51–52. https://doi.org/10.1145/2020976.2021004
29. Microsoft Corporation (2018) *Microsoft Excel*. Available at: https://office.microsoft.com/excel (accessed 21 February 2021).

About the authors

Nidhi Gautam

PhD (Computer Science & Engineering);

Assistant Professor, Fellow Senate Panjab University, University Institute of Applied Management Sciences, Panjab University, Chandigarh, India;

E-mail: nidhi.uiams@pu.ac.in

ORCID: 0000-0002-9454-3625

Nitin Kumar

MBA (IT & Telecommunications);

Management Trainee, Housing Development Finance Corporation Limited, Pathankot, Punjab, India;

E-mail: nk2268913@gmail.com

ORCID: 0000-0002-8887-2350

DOI: 10.17323/2587-814X.2022.1.83.100

Centralized resource allocation based on energy saving and environmental pollution reduction using data envelopment analysis models

Sima Madadi [a]
E-mail: sima.madadi@gmail.com

Farhad Hosseinzadeh Lotfi [a] [*]
E-mail: farhad@hosseinzadeh.ir

Mehdi Fallah Jelodar [b]
E-mail: Mehdi.fallah_jelodar@yahoo.com

Mohsen Rostamy-Malkhalifeh [a]
E-mail: Mohsen_rostamy@yahoo.com

[a] Department of Mathematics, Science and Research Branch, Islamic Azad University
Address: Shahid Sattari Square, Tehran 14515/775, Iran

[b] Department of Mathematics, Ayatollah Amoli Branch, Islamic Azad University
Address: 5 Km of the old road from Amol to Babol, Amol 678, Iran

Abstract

Environmental pollution has caused governments to be concerned about energy saving and the reduction of environmental pollution. Some researchers have presented resource allocation models as multi-objective linear programming (MOLP) in order to pay more attention to energy saving and environmental pollution reduction. Energy saving affects both desirable and undesirable outputs. In this paper, we argue for the inapplicability of the existing models for reducing the undesirable outputs through energy saving. The purpose of this paper is to design a model based on data envelopment analysis (DEA) that would result in reduced pollution through energy saving. Moreover, since an undesirable output is

[*] Corresponding author

considered as a function of the total desirable outputs, if necessary, the changes should be applied to the total desirable outputs and there is no need to reduce each desirable output individually. Finally, the model proposed based on goal programming (GP) is used in 20 different regions in China. The results produced by this model indicate that the reduction proportion of total environmental pollution emissions per energy saving was larger than the reduction proportion of total desirable outputs.

Keywords: resource allocation, data envelopment analysis, energy saving, environmental pollution, desirable output, undesirable output

Citation: Madadi S., Hosseinzadeh Lotfi F., Fallah Jelodar M., Rostamy-Malkhalifeh M. (2022) Centralized resource allocation based on energy saving and environmental pollution reduction using data envelopment analysis models. *Business Informatics*, vol. 16, no. 1, pp. 83–100.
DOI: 10.17323/2587-814X.2022.1. 83.100

Introduction

One of the outputs of industrial development is environmental pollution. Carbon dioxide gas accounts for about 60 percent of all greenhouse gases, and about 81 percent of all greenhouse gas emissions come from fossil fuel consumption. Fossil fuels, on the one hand, are the most important sources of energy for cities, while being the main source of pollution. Iran is one of the highest carbon emission-intensive countries in the world. Total CO_2 emissions in 1990 were 201.8 million metric tons (MMT), which has increased rapidly at an average annual rate of 5.7% to 372 MMT by 2003 [1].

The amount of particulate matters in the atmosphere is one of the most important indicators of air pollution. Aerosols have a great impact on the climate and human environment. Tropospheric aerosols, known as particulate matter, have an adverse effect on human health [2]. One of the atmospheric pollutants, sulfur dioxide, causes acid rain and many other adverse environmental effects and health hazards [3].

The main reason for the increasing CO_2 emissions in industries is high levels of energy consumption. Based on reports from the International Energy Agency [4], the global industrial sector is responsible for about 40% of total energy consumption in the world. Due to the raised awareness about environmental issues and technological advancements that help reduce environmental damage, there has been a decline in CO_2 emissions from industries in developed countries; however, such emissions are greatly increasing in developing countries [5].

Data envelopment analysis (DEA) is novel area of study, as well as a necessary mathematical tool for evaluating the relative efficiency of a set of homogenous decision-making units (DMU). This method has attracted a lot of attention in various fields of management sciences [6]. DEA, which is a non-parametric method, has had many applications in solving the problem of resource allocation when all DMUs are under the control of a single centralized decision maker (DM). There is no prior functional form in DEA, and there is no need for the many assumptions that emerge by using statistical methods for function estimation. At the same time, DEA produces good results when used in resource allocation [5, 7]. Many papers have been presented based on DEA for allocating resources to a set of DMUs in which the purpose of the DM has been to minimize the total input consumption or to maximize the total output production of all DMUs instead of considering each one individually and set separate targets for each DMU. Centralized resource allocation was presented by [8] for the first time, which sought radical reduction in the total consumption of every input by all units. In [7], the authors designed a multi-objective model for resource allocation to find the maximum amount of production. They defined a transformation possibility set for each DMU with two assumptions; the first one was to assume that the unit's efficiency stays constant during the planning period, and the other assumption was that each unit could have a proportional scaling of changes in inputs and outputs. Later, [9] considered one of the models presented by [8] and modified it to adjust the inefficient units. The study [10] extended a simplified model of [8]. This model recognized more efficient units and was much simpler to execute than the previous models. The special characteristic of this model was that the centralized DM did not necessarily need to keep the original number of DMUs fixed. For other extensions of the model proposed by [8], one can refer to [11, 12]. Other proposed models for centralized resource allocation can be found in [13–15]. The study [14] proposed two ideas: one idea maximized the total efficiency, while the other one simultaneously maximized the output production and minimized the input consumption. In this method, the new efficiency of all DMUs becomes equal to one after production design; this efficiency improvement is not logical and feasible in practice. On the other hand, there is no guarantee that the inputs (outputs) will decrease (increase) significantly. Also, there is no logical connection between these changes and they may not be fair. In [16], the authors extended a method that implemented the demand and supply changes in a centralized decision-making environment under a predictable assumption. The study [17] presented a DEA model for centralized resource allocation with the assumption of adjustable and non-adjustable inputs and transferable and non-transferable outputs. Then, he analyzed the structural efficiency of the model using the structural efficiency analysis presented by [18]. The study [19] combined energy consumption reduction through resource allocation with DEA models with undesirable outputs, and proposed a multi-objective model for resource allocation under energy saving constraints. Since energy saving decreases both the desirable and undesirable outputs, the aim of their model was to make the reduction proportion of the desirable outputs would be less than the reduction proportion of the harmful outputs. This way, recommendations can be made regarding energy and environmental policies toward saving energy and reducing air pollution. They also studied the classification of natural resources in China and used an input-oriented slacks-based model for measuring the efficiency of provinces [20]; then, they proposed a DEA-based approach for allocating the total natural resources. Unlike conventional DEA models, it seems necessary to consider both desirable and undesirable outputs in environmental performance evaluation [21].

Many of the findings of DEA studies have been used for environmental performance measurement. The study [22] focused on the analysis of optimal energy allocation and environmental performance of China's three major urban agglomerations. In particular, that study first used a fixed-input DEA model to obtain the optimal allocation of energy input. Then, an evaluation model based on the optimal allocation of energy input was proposed to evaluate the environmental performance. In [23], the researchers constructed an evaluation indicator system based on three stages, namely economic production, wastewater treatment, and human health, and used the undesirable three-stage dynamic data envelopment analysis model to empirically evaluate the total efficiency, stage efficiency, and the efficiency of various indicators.

Goal Programming (GP) is a developed form of Linear Programming. GP tries to achieve several goals simultaneously and allows deviation from the goal. Therefore, it has flexibility in decision-making processes. The main approach of GP is to allocate a special target value to each objective function and then look for a solution that would minimize unwanted deviations from the intended goals [24]. GP was used as a method for solving multi-objective problems with the aim of minimizing unwanted deviations from the set goals. There exist two main algorithms for solving a GP problem: the weighted sum model and the lexicographic model. The studies [25, 26] proposed using GP for MCDEA models. The difficulty in solving a multi-objective problem is finding a solution that would optimize all the objectives simultaneously [27]. Since there is no such solution in most cases, a non-dominated solution set is needed. Paper [25] proposed using the lexicographic model to solve GP problems and allocating priority to the objective functions of MCDEA.

In [26] was proposed the weighted goal programming method (GPDEA). The studies [28, 29] addressed the connections between multi-objective problems and DEA. Furthermore, there have been some models that maximized the efficiency of all DMUs simultaneously (e.g., [30–33]). Industrial production is always associated with energy consumption and greenhouse gas emission (the most important is CO_2 emission). As energy consumption decreases, the desirable output will also decrease, but when industrial estates are required by governments to reduce and control pollution, if energy storage does not lead to a reduction in environmental pollution, the model is not valid in the eyes of the central manager. In the centralized resource allocation model proposed by [19], we show that undesirable output changes become zero by saving energy. In this paper, we modify their model so that with a reduction in energy consumption, a significant reduction in CO_2 emission is achieved, and show that if the centralized DM considers boundaries for changes in the inputs and outputs, the model may be infeasible, since choosing suitably and feasibly will be a difficult task for the centralized DM. Therefore, the model is modified through GP in a way that makes it feasible.

On the other hand, the reduction of individual desirable outputs due to reductions in the undesirable outputs has the weakness that some undesirable outputs may have been out of the acceptable standard range. In such cases, some undesirable outputs may be reduced without any reduction in the desirable outputs.

What this paper proposes is that since an undesirable output is considered to be a function of the total desirable outputs, if necessary, the changes should be applied to the total desirable outputs. According to the abovementioned, the innovations of this research are:

- Rectifying the infeasibility of the allocation model in cases where unsuitable boundaries are selected for the input/output changes, which are assigned by the DM.
- Modifying the pre-presented model and eliminating the weakness of the respective model in reducing the undesirable outputs.
- Presenting a new model that does not require the reduction of each and every desirable output in the units (production industries) in order to save energy and reduce pollution, since there could be a case where in a given region, some units have a large amount of undesirable outputs due to performance weaknesses, in which case the reduction of a portion of undesirable outputs in the entirety of units may not require a reduction of desirable outputs in all units.

The rest of the paper is organized as follows: In section 1 (Theoretical background), an introduction is provided to the conventional DEA model and the centralized resource allocation models, as well as the method of using GP to solve multi-objective problems. This section also discusses the defects of the previously mentioned model. In section 2 (Proposed model), we present our proposed model for centralized resource allocation with the aim of energy saving and reducing environmental pollution emissions. The advantages to the model are also included in this section. The application of GP in the proposed resource allocation model is illustrated through a numerical example in section 3 (Numerical example). Finally, some conclusions and remarks are provided.

1. Theoretical background

Data envelopment analysis (DEA) is a powerful tool for evaluating the relative efficiency of a set of DMUs that consume multiple inputs to produce multiple outputs. Suppose there are n DMUs that are in need of evaluation, and each one consumes m different inputs to produce s different outputs. Suppose $X_j = (x_{1j}, \ldots x_{mj})^T$ and $Y_j = (y_{1j}, \ldots y_{sj})^T$, $X_j \geq 0$, $Y_j \geq 0$ are the input and output vectors, respectively. The production possibility set T is defined as:

$$T = \{(x, y) \mid y \text{ can be produced from } x\}. \quad (1)$$

In [34] is defined the following PPS using the constant returns to scale (CRS) assumption.

$$T_{CCR} = \left\{ (x, y) \in R^{m+s}_{\geq 0} \left| \sum_{j=1}^{n} \lambda_j x_{ij} \leq x, \right. \right.$$
$$\left. \sum_{j=1}^{n} \lambda_j y_{rj} \geq y,\ \lambda_j \geq 0,\ j = 1,\ldots,n \right\}. \quad (2)$$

The input-oriented model for evaluating DMU_o, $o \in \{1, \ldots, n\}$ under the assumption of CRS can be achieved by solving the following ratio programing problem [34].

$$\max \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}}, \quad (3)$$

$$\text{s.t.} \quad \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \leq 1,$$

$$u_r, v_i \geq \varepsilon \qquad r = 1, \ldots, s, \qquad i = 1, \ldots, m.$$

Here $\varepsilon > 0$ is a non-Archimedean element defined to be smaller than any positive real number.

GP provides the means for attempting to achieve several objectives simultaneously. Many researchers, including [28, 29, 35], have investigated the relationships between DEA and MOP. Several methods have been developed to solve multi-objective problems (see: [36–38]), one of which is Goal programming [24, 39].

1.1. Resource allocation models

In recent years, various applications of DEA have been seen in most countries around the world for the purposes of evaluating the performance of organizations and other common activities in different areas. In the context of planning and resource allocation, a number of optimization techniques have been introduced, such as multi-objective programming. The purpose of a central unit is to design a reasonable resource allocation mechanism that can bring the greatest benefits for the central organization [7, 40, 41]. In many real-world scenarios, all of the DMUs may be under the influence of a central decision maker who can supervise the resource consumption of these units. The main purpose of resource allocation is to allocate resources in such a way that the general goals of the organization are achieved as far as possible. Unlike conventional DEA models, [42] considered undesirable factors as an important factor in efficiency evaluation. The studies [43–45] suggested an alternative approach in environmental technology in which the desirable outputs increased while the undesirable outputs decreased. The study [19] considered both desirable and undesirable outputs in their evaluation, as there are undesirable outputs in the production process. Their model helped the DM allocate future resources while taking energy saving into account. They combined the energy consumption reduction targets with resource allocation and proposed a multi-objective programming model that not only reduced the undesirable outputs but also decreased the desirable outputs in order to improve the undesirable output production.

They defined the transformation possibility set as follows:

$$F_j = \left\{ \begin{pmatrix} x_j - \Delta x_j \\ y_j^g - \Delta y_j^g \\ y_j^b - \Delta y_j^b \end{pmatrix} \middle| \Delta y_j^g \geq \delta_j y_j^g,\ \Delta y_j^b \leq \delta_j y_j^b \right\}, \quad (4)$$

$$\delta_j = \max\left\{ \frac{\Delta x_{ij}}{x_{ij}} \middle| i = 1,\ldots,m \right\}.$$

Their model, based on the CRS assumption, is formulated as follows:

$$\min \quad \Delta Y^g = \sum_{r=1}^{s_1}\sum_{j=1}^{n} \frac{\Delta y_{rj}^g}{\sum_{j=1}^{n} y_{rj}^g},$$

$$\max \quad \Delta X = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\Delta x_{ij}}{\sum_{j=1}^{n} x_{ij}}, \quad (5)$$

s.t.

$$\Delta y_j^g \geq \delta_j y_j^g,\ j = 1, \ldots, n, \quad (5\text{-}1)$$

$$\Delta y_j^b \leq \delta_j y_j^b,\ j = 1, \ldots, n, \quad (5\text{-}2)$$

$$\Delta x_j \leq \delta_j x_j,\ j = 1, \ldots, n, \quad (5\text{-}3)$$

$$y_j^g - \Delta y_j^g \leq Y^g \Lambda_j,\ j = 1, \ldots, n, \quad (5\text{-}4)$$

$$y_j^b - \Delta y_j^b \geq Y^b \Lambda_j,\ j = 1, \ldots, n, \quad (5\text{-}5)$$

$$x_j - \Delta x_j \geq X \Lambda_j,\ j = 1, \ldots, n, \quad (5\text{-}6)$$

$$A_j \leq \Delta y_j^g \leq B_j,\ j = 1, \ldots, n, \quad (5\text{-}7)$$

$$\Delta Y^b = \sum_{j=1}^{n} \Delta y_j^b \leq M. \quad (5\text{-}8)$$

Where the vectors

$$\begin{pmatrix} X_j - \Delta X_j \\ Y_j^g - \Delta Y_j^g \\ Y_j^b - \Delta Y_j^b \end{pmatrix} \in R_{\geq 0}^{m+s_1+s_2}$$

and the matrices X, Y^g, Y^b are defined as follows:

$$X = \left[x_1, x_2, \ldots, x_n\right] \in R^{m\times n},$$

$$Y^g = \left[y_1^g, y_2^g, \ldots, y_n^g\right] \in R^{s_1\times n},$$

$$Y^b = \left[y_1^b, y_2^b, \ldots, y_n^b\right] \in R^{s_2\times n},$$

$$X > 0,\ Y^g > 0,\ Y^b > 0.$$

ΔX_j represents the saving amount of inputs in DMU_j, any Δy_j^g, Δy_j^b denote the reduction amounts of desirable and undesirable outputs in DMU_j, respectively. F_j (transformation possibility set) represents the capacity of input and output changes for DMU_j. $[A_j, B_j]$ indicates the capacity of desirable output changes, and M is the maximum emission reduction, which is determined by the DM.

2. Proposed model: resource allocation models based on goal programming

In Model (5), independent of what the value of δ_j^* (positive or zero) is in the optimal solution, in the constraints (5-2) and (5-8), $\Delta y_j{}^b = 0$ is true. It has also already been established that $\Delta y_j{}^b = 0$ is true in constraint (5-5). Therefore, in all constraints, $\Delta y_j{}^b = 0$ ($j = 1,\ldots, n$) is a solution, and since it does not exist in the objective function, then $\Delta y_j{}^b = 0$ is always true, which indicates a defect in Model (5). Let us also assume that the manager considers the following goals:

$$A_j \leq \Delta x_j \leq B_j,\ C_j \leq \Delta y_j^g \leq D_j,$$

$$\sum_{j=1}^{n} \Delta y_j^b \geq M,\ A_j, B_j \in R_{\geq 0}^m,$$

$$C_j, D_j \in R_{\geq 0}^{s_1},\ M \in R_{\geq 0}^{s_2}.$$

If A_j, B_j, C_j, D_j, M for $j = 1,\ldots, n$ are not chosen pro-perly, Model (5) will be infeasible. In this paper, this model is modified using GP in a way that it becomes feasible and $\Delta y_j^b > 0$ is obtained.

Therefore, we define F_j as follows.

$$F_j = \left\{ \begin{pmatrix} x_j - \Delta x_j \\ y_j^g - \Delta y_j^g \\ y_j^b - \Delta y_j^b \end{pmatrix} \middle| \Delta y_j^g \ge \delta_j y_j^g \,,\ \Delta y_j^b \ge \delta_j y_j^b \right\}, \tag{6}$$

$$\delta_j = \max\left\{ \frac{\Delta x_{ij}}{x_{ij}} \middle| i = 1,\dots,m \right\}.$$

And assuming that the production possibility set remains unchanged in each step,

$$T = \left\{ \left(x, y^g, y^b\right) \in R_{\ge 0}^{m+s_1+s_2} \middle| \begin{array}{l} x \ge \sum_{j=1}^{n} \lambda_j x_j, \\ y^g \le \sum_{j=1}^{n} \lambda_j y_j^g, \\ y^b \ge \sum_{j=1}^{n} \lambda_j y_j^b, \\ \lambda_j \ge 0, j = 1,\dots,n \end{array} \right\}, \tag{7}$$

$$\Delta X = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\Delta x_{ij}}{\sum_{j=1}^{n} x_{ij}}, \quad \Delta Y^g = \sum_{r=1}^{s_1} \sum_{j=1}^{n} \frac{\Delta y_{rj}^g}{\sum_{j=1}^{n} y_{rj}^g}. \tag{8}$$

Now, the resource allocation model is presented using GP as follows:

$$\begin{aligned} \min \quad & Z_1 = \sum_{j=1}^{n} (n_j^1 + n_j^2) + \sum_{j=1}^{n} (p_j^1 + p_j^2) + L, \\ \min \quad & Z_2 = \Delta Y^g, \\ \min \quad & Z_3 = \Delta X. \end{aligned} \tag{9}$$

s.t.

$$x_{ij} - \Delta x_{ij} \ge \sum_{l=1}^{n} \lambda_{jl} x_{il},$$

$$j = 1, \dots, n, \ i = 1, \dots, m, \tag{9-1}$$

$$y_{rj}^g - \Delta y_{rj}^g \le \sum_{l=1}^{n} \lambda_{jl} y_{rl}^g,$$

$$j = 1, \dots, n, \ r = 1, \dots, s_1, \tag{9-2}$$

$$y_{pj}^b - \Delta y_{pj}^b \ge \sum_{l=1}^{n} \lambda_{jl} y_{pl}^b,$$

$$j = 1, \dots, n, \ p = 1, \dots, s_2, \tag{9-3}$$

$$\Delta x_{ij} \le \delta_j x_{ij}, j = 1, \dots, n, \ i = 1, \dots, m, \tag{9-4}$$

$$\Delta y_{rj}^g \ge \delta_j y_{rj}^g, j = 1, \dots, n, \ r = 1, \dots, s_1, \tag{9-5}$$

$$\Delta y_{pj}^b \ge \delta_j y_{pj}^b, j = 1, \dots, n, \ p = 1, \dots, s_2, \tag{9-6}$$

$$\Delta y_j^g \ge C_j - n_j^1, \ j = 1, \dots, n, \tag{9-7}$$

$$\Delta y_j^g \le D_j + n_j^2, \ j = 1, \dots, n, \tag{9-8}$$

$$\Delta x_j \ge A_j - p_j^1, \ j = 1, \dots, n, \tag{9-9}$$

$$\Delta x_j \le B_j - p_j^2, \ j = 1, \dots, n, \tag{9-10}$$

$$\sum_{j=1}^{n} \Delta y_j^b \ge M - L, \tag{9-11}$$

$$\Delta y_j^g \le y_j^g, j = 1, \dots, n, \tag{9-12}$$

$$\lambda_{jl} \ge 0, \ j = 1, \dots, n, \ l = 1, \dots, p,$$

where ΔY^g and ΔX are as defined in equation (8), and

$$\begin{gathered} n_j^1, n_j^2 \in R_{\ge 0}^{s_1}, \ p_j^1, p_j^2 \in R_{\ge 0}^{m}, \ L \in R_{\ge 0}^{s_2}, \\ \Delta x_j = [\Delta x_{1j}, \Delta x_{2j}, \dots \Delta x_{mj}], \\ \Delta y_j^g = \left[\Delta y_{1j}^g, \Delta y_{2j}^g, \dots, \Delta y_{s_1 j}^g\right], \\ \Delta y_j^b = \left[\Delta y_{1j}^b, \Delta y_{2j}^b, \dots, \Delta y_{s_2 j}^b\right]. \end{gathered}$$

Constraints (9-1) – (9-3) in model (9) indicate that the reduced outputs and inputs belong to the PPS. Constraints (9-4) – (9-6) ensure that the changed output and input values for each DMU belong to its own transformation possibility set. In constraints (9-7) and (9-8), if the management's expectation for Δy_j^g to fall within the interval of C_j, D_j is unattainable, the deviation variables n_j^1, n_j^2 will modify it and make the problem feasible. This is also true for constraints (9-9) – (9-11).

The optimal values of this model can be obtained in two steps. The first step is to obtain the minimum of the total deviation variables for the goal considered by the central manager, which is considered as the first priority for the problem to be feasible, and then obtain the optimal solution to this model using the lexicographic method. The second step is to obtain the weighted sum of the two next objective functions in order to minimize the desirable output reduction and maximize energy saving in the optimal solution obtained from the first step.

Theorem. $\Delta y_j^b > 0$ is true for all $j = 1,\ldots, n$ that have a positive Δx_j.

Proof: Since $x_j > 0$ and $\Delta x_j > 0$ in constraint (9-4), $\delta_j > 0$ is true. On the other hand, based on constraint (9-6) and the fact that $y_j^b > 0$ and $\delta_j > 0$, we arrive at $\Delta y_j^b > 0$.

On the other hand, since an undesirable output is a function of the total desirable outputs, if necessary, the required changes shall be applied to the totality of the desirable outputs. Therefore, reducing individual desirable outputs is not logical, as some of the undesirable outputs may have been out of the acceptable standard range. For example, the carbon monoxide gas produced in industrial plants in a geographical region would cause pollution in that region. However, reduced pollution may be achieved by a reduction in any one of the factories, so all factories are not necessarily forced to reduce their emissions. In this regard, the transformation possibility set defined in this paper will, in addition to energy saving, reduce the undesirable outputs following a minimum reduction in the total desirable outputs. Therefore, the set F is defined as a transformation possibility set for the total inputs and outputs as follows:

$$F=\left\{\left(\begin{array}{c}\sum_{j=1}^{n}(x_j-\Delta x_j)\\ \sum_{j=1}^{n}(y_j^g-\Delta y_j^g)\\ \sum_{j=1}^{n}(y_j^b-\Delta y_j^b)\end{array}\right)\left|\begin{array}{l}\sum_{j=1}^{n}\Delta y_{rj}^g \ge \delta \sum_{j=1}^{n} y_{rj}^g,\\ \sum_{j=1}^{n}\Delta y_{pj}^b \ge \delta \sum_{j=1}^{n} y_{pj}^b\\ r=1,\ldots,s_1,\ \ p=1,\ldots,s_2\end{array}\right.\right\}, \tag{10}$$

$$\delta_j = \max\left\{\frac{\Delta x_{ij}}{x_{ij}}\middle| i=1,\ldots,m\right\},$$

$$\delta = \min_j \delta_j.$$

Now, by considering a tradeoff of reductions in the desirable and undesirable outputs and using GP, the centralized resource allocation model is presented as follows:

$$\min \quad Z_1 = \sum_{j=1}^{n}(n_j^1+n_j^2)+\sum_{j=1}^{n}(p_j^1+p_j^2)+L, \tag{11}$$

$$\min \quad Z_2 = \Delta Y^g, \max \quad Z_3 = \Delta X,$$

s.t.

$$\sum_{j=1}^{n}(x_{ij}-\Delta x_{ij}) \ge \sum_{j=1}^{n}\sum_{l=1}^{n}\lambda_{jl}x_{il},\ i=1,\ldots,m, \tag{11-1}$$

$$\sum_{j=1}^{n}(y_{rj}^g-\Delta y_{rj}^g) \le \sum_{j=1}^{n}\sum_{l=1}^{n}\lambda_{jl}y_{rl}^g,\ r=1,\ldots,s_1, \tag{11-2}$$

$$\sum_{j=1}^{n}(y_{pj}^b-\Delta y_{pj}^b) \ge \sum_{j=1}^{n}\sum_{l=1}^{n}\lambda_{jl}y_{pl}^b,\ p=1,\ldots,s_2, \tag{11-3}$$

$$\Delta x_{ij} \le \delta_j x_{ij},\ j=1,\ldots,n,\ i=1,\ldots,m, \tag{11-4}$$

$$\sum_{j=1}^{n}\Delta y_{rj}^g \ge \delta \sum_{j=1}^{n} y_{ri}^g,\ r=1,\ldots,s_1, \tag{11-5}$$

$$\sum_{j=1}^{n}\Delta y_{pj}^b \ge \delta \sum_{j=1}^{n} y_{pj}^b,\ p=1,\ldots,s_2, \tag{11-6}$$

$$\delta \le \delta_j,\ j=1,\ldots,n, \tag{11-7}$$

$$\Delta y_j^g \ge A_j - n_j^1,\ j=1,\ldots,n, \tag{11-8}$$

$$\Delta y_j^g \le B_j + n_j^2,\ j=1,\ldots,n, \tag{11-9}$$

$$\Delta x_j \ge C_j - p_j^1,\ j=1,\ldots,n, \tag{11-10}$$

$$\Delta x_j \le D_j + p_j^2,\ j=1,\ldots,n, \tag*{(11-10)}$$

$$\sum_{j=1}^{n}\Delta y_j^b \ge M-L, \tag{11-12}$$

$$\begin{array}{c} x_j-\Delta x_j \ge 0,\ y_j^g-\Delta y_j^g \ge 0,\\ y_j^g-\Delta y_j^g \ge 0,\ \ j=1,\ldots,n,\end{array} \tag{11-13}$$

$$\lambda_{jl} \ge 0,\ j=1,\ldots,n,\ l=1,\ldots,p.$$

Constraints (11-1) – (11-3) in model (8) indicate that the total reduced outputs and inputs belong to the PPS. That is to say,

$$\begin{pmatrix} \sum_{j=1}^{n}(x_j - \Delta x_j) \\ \sum_{j=1}^{n}(y_j^g - \Delta y_j^g) \\ \sum_{j=1}^{n}(y_j^b - \Delta y_j^b) \end{pmatrix} \in T$$

and constraints (11-4) – (11-7) indicate that

$$\begin{pmatrix} \sum_{j=1}^{n}(x_j - \Delta x_j) \\ \sum_{j=1}^{n}(y_j^g - \Delta y_j^g) \\ \sum_{j=1}^{n}(y_j^b - \Delta y_j^b) \end{pmatrix} \in F.$$

In other words, it is guaranteed that the total changed values of inputs and outputs belong to the transformation possibility set for all inputs and outputs. Constraints (11-8) – (11-12) are conditions set by the central manager. The model above is converted to a model under variable returns to scale (VRS) assumption by adding $\sum_{l=1}^{n}\lambda_{jl} = 1$. The optimal values of model (11) can beobtained through prioritization.

Lemma 1. Models (9) and (11) are always feasible regardless of the goals set by the manager.

Proof: In Model (9), by choosing $\forall j, \forall l \quad \lambda_{jl} = 0$, $x_j = 0, \Delta y_j^{\,b} = y_j^{\,b}, \Delta y_j^{\,g} = y_j^{\,g}, n_j^{\,1} = A_j, n_j^2 = y_j^{\,g}, p_j^{\,1} = C_j$, $p_j^{\,2} = 0$, $L = M$, $\delta_j = 0.2$ we have a feasible solution to the model. Similarly, Model (11) is also feasible.

Lemma 2. If the first objective function receives a positive value in optimality, it means that the goals set by the manager are unreachable and deviation variables play an important role in the feasibility.

For the computational comparison of models (5) and (11), the conditions considered for changes in the inputs and outputs can be discarded, because in each problem, depending on the opinion of the central manager, these conditions may or may not apply. When the conditions imposed by the central manager are set aside in both models (constraints (5-7) and (5-8) in model (5), and constraints (11-8) and (11-13), then in model (11), we will have only the second and third objective functions, which are equivalent to both objective functions in model (5). As can be seen in the calculations *Table1*, model (11) has less computational volume than model (5).

Even if the conditions imposed by the central manager are considered the same in each mod-

Table1.

Comparison of the constraints of model (5) and model (11)

Model (5)	Number of constraints	Model (11)	Number of constraints
(5–1)	$s_1 \times n$	(11–1)	m
(5–2)	$s_2 \times n$	(11–2)	s_1
(5–3)	$m \times n$	(11–3)	s_2
(5–4)	$s_1 \times n$	(11–4)	$(m \times n)$
(5–5)	$s_2 \times n$	(11–5)	s_1
(5–6)	$m \times n$	(11–6)	s_2
		(11–7)	n
Total:	$(2s_1 + 2s_2 + 2m) \times n$	Total:	$2s_1 + 2s_2 + mn + m + n$

el, for model (5) to be always feasible, the first objective function of model (11) must be added to model (5), in which case since the model is solved by lexicography's prioritization method, the computational volume in both cases will be doubled, which again makes model (11) computationally eco-nomical, especially when the number of units is significant.

Advantages of model (11) compared with model (5):

1. While in model (5), $\Delta y^b = 0$ is obtained along with the reduction of energy, model (11) was changed so that Δy^b can receive a positive value (these have been proven at the beginning of part 2 and the theorem). That is, model (11) can reduce environmental pollution by reducing energy consumption, while model (5) cannot.

2. Even if the parameters are chosen inappropriately, the proposed model (11) is always feasible (due to the existence of deviation variables, while these variables do not exist in model (5)).

3. Since the undesirable outputs may not be within the acceptable standard range, Model (11) is not forced to reduce each desirable output individually. Thus, the required changes are applied to the totality of the desirable outputs and inputs.

4. The number of constraints is significantly reduced in model (11).

3. Numerical example

In this section, we apply Models (9) and (11) to a numerical example for the purposes of analysis. *Table 1* exhibits a simple data set for six DMUs that produce two outputs using one input (desirable and undesirable), which are under the supervision of a central management. We solve Model (9) and Model (11) under CRS and VRS assumptions by lexicography's prioritization method. The first objective function is considered as the first priority for the problem to be feasible. In other words, $Z_1^* > 0$ means that the deviation variable makes the problem feasible, and if we had not considered the problem as GP, then it would be infeasible. The second step is to obtain the sum of the next two weighted objective functions in order to minimize desirable output reduction and maximize input saving in the optimal solution, which is obtained from the first step. *Table 2* shows the input and output data for the 6 DMUs. The following *Tables 3* and *4* provide the results of solving the model (9) using Gams software under CRS assumption and entering the parameters as $A_j = 0$, $B_j = 0.6x_j$, $c_j = 0$, $D_j = 0.3y_j^g$, $M = 0.8\sum_{j=1}^{n} y_j^b$.

By solving the model (9), the optimal value obtained for the first objective function is $Z_1^* = 3.16$; this means that the deviation variables have played an important role in making the problem feasible, and if we did not consider the problem as a GP, then it would be infeasible. *Table 3* shows the reduced values of inputs and outputs, as well as the reduction proportion of each one. In general, the reduction proportion of inputs is 0.19, the reduction proportion of desirable outputs is 0.38, and the reduction proportion of undesirable outputs is 0.75, which shows that overall, the reduction proportion of undesirable outputs is larger than the reduction proportion of desirable outputs. *Table 4* presents the values allocated to the inputs and outputs (desirable and undesirable) after energy

Table 2.

Input and output data for illustrating the proposed models

Unit	x	y^g	y^b
A	3.00	2.00	2.00
B	4.20	3.00	7.10
C	2.70	4.00	5.00
D	5.00	6.00	4.50
E	6.00	4.00	2.00
F	3.80	2.00	5.00
Total	24.7	21	25.6

Table 3.

Reduction amounts of inputs and outputs under CRS assumption in model (9)

Unit	Δx	Δy^g	Δy^b	Reduction proportion		
				x	y^g	y^b
A	0.90	0.60	1.30	0.30	0.30	0.65
B	1.05	0.90	6.05	0.25	0.30	0.01
C	0.00	2.20	4.10	0.00	0.55	0.82
D	0.00	2.66	2.83	0.00	0.44	0.62
E	1.80	1.20	0.60	0.30	0.30	0.30
F	1.14	0.60	4.30	0.30	0.30	0.86
Total	4.89	8.16	19.18	0.19	0.38	0.75

saving and reducing environmental pollutions for the purposes of providing recommendations to the central decision maker. Furthermore, the amount of reduction and the reduction proportion of inputs and outputs under VRS assumption model (9) are shown in *Table 5.*

As can be observed, in some DMUs, the reduction proportion of desirable outputs exceeds the proportion that was considered, and the reduction proportion of undesirable outputs is less than the lower bound that was set. This is due to the existence of deviation variables that make the problem feasible. To compare model (9) with model (5), the results obtained by placing the above parameters in model (5) are given in *Table 6* (we even set the conditions for Δx to be the same conditions for both models). As can be seen from the results of *Table 6*, the amount of reduction in the undesirable outputs for each unit is zero. This is the weakness of the respective model, which does not allow the reduction of undesirable outputs by reducing the desirable outputs and desirable outputs of the model.

Table 4.

Allocated values for inputs and outputs under CRS assumption in model (9)

Unit	$x - \Delta x$	$y^g - \Delta y^g$	$y^b - \Delta y^b$
A	2.10	1.40	0.70
B	3.15	2.10	1.05
C	2.70	1.80	0.90
D	5.00	3.33	1.66
E	4.20	2.80	1.40
F	2.66	1.40	0.70

By solving Model (11), the optimal value obtained for the first objective function is $Z_1^* = 2.23$. In general, the reduction proportion of inputs is 0.10, the reduction proportion of desirable outputs is 0.30, and the reduction proportion of undesirable outputs is 0.71 (according to *Table 7*).

Table 8 provides the values allocated to the inputs and outputs (desirable and undesirable) after energy saving and reducing environmental pollution by considering a tradeoff of reductions in the inputs and outputs. Now we analyze our model through a real example of 20 Chinese regions. The values regarding China's total fossil fuel energy consumption (i.e., raw coal, clean coal, briquettes, coke, coke oven gas, crude oil, gasoline, kerosene, fuel oil, diesel oil, refinery gas, liquefied petroleum gas and natural gas), non-fossil fuel consumption, CO_2 emissions and regional GDP were collected from [46]. These values are listed in *Table 9.*

Table 5.

Reduction amounts of inputs and outputs with VRS assumption model (9)

Unit	Δx	Δy^g	Δy^b	Reduction proportion		
				x	y^g	y^b
A	0.00	0.00	0.00	0.00	0.00	0.00
B	1.05	0.90	5.10	0.25	0.30	0.72
C	0.00	0.00	0.00	0.00	0.00	0.00
D	0.00	1.80	1.68	0.00	0.30	0.37
E	0.00	0.00	0.00	0.30	0.00	0.00
F	0.80	0.42	3.00	0.21	0.21	0.60
Total	1.85	3.12	9.78	0.07	0.14	0.38

Table 6.

Reduction amounts of inputs and outputs under CRS assumption in model (5)

Unit	Δx	Δy^b	Δy^g
A	0.90	0.60	0
B	1.26	0.90	0
C	0.00	0.00	0
D	0.00	0.00	0
E	1.80	1.20	0
F	1.14	0.60	0
Total	5.10	3.30	0

As can be observed in *Table 10*, Model (11) has changed a number of inputs and outputs, not necessarily all of them. In general, for a 0.44 reduction in total energy consumption and a 0.04 reduction in non-fossil fuel consumption, we will have a 0.62 reduction in environmental pollution, whereas the desirable outputs are reduced by 0.30. These results provide important information to the decision maker, namely to reduce CO_2 emission by about 0.62 through saving energy in fossil fuel energy consumption by 0.44 and saving energy in Non-fossil fuel energy consumption by 0.04. This reduces the desirable output (GDP) by 0.30.

Table 7.

Reduction amounts of inputs and outputs with CRS assumption model (11)

Unit	Δx	Δy^g	Δy^b	Reduction proportion		
				x	y^g	y^b
A	1.80	0.60	2.00	0.60	0.30	1.00
B	0.85	0.90	7.10	0.20	0.30	1.00
C	0.00	1.20	5.00	0.00	0.30	1.00
D	0.00	1.80	4.15	0.00	0.30	0.92
E	0.00	1.20	0.00	0.00	0.30	0.00
F	0.00	0.60	0.00	0.00	0.30	0.00
Total	2.65	6.30	18.25	0.10	0.30	0.71

Table 8.

Results of Allocated value for inputs and outputs with CRS assumption model (11)

Unit	$x - \Delta x$	$y^g - \Delta y^g$	$y^b - \Delta y^b$
A	1.20	1.40	0.00
B	3.35	2.10	0.00
C	2.70	2.80	0.00
D	5.00	4.20	0.35
E	6.00	2.80	2.00
F	3.80	1.40	5.00

Conclusion

Controlling the pollution from manufacturing industries in developed and developing countries has become a common concern among researchers and governments. The use of DEA-based models as a powerful tool in problems of pollution reduction and energy consumption has attracted the attention of researchers. This also relates to the allocation of resources in organizations that have a central decision maker, such as the Ministry of Health, the Ministry of Education, and the World Health Organization, which are able

Table 9.

The data set are compiled from 30 regions of China in 2005 [51]

Unit	Total fossil fuel Energy consumption (million tce)	Non-fossil fuel consumption (million tce)	GDP (billion RMB at 2005) (million tce)	CO_2 emissions (million tone)
1	55.2	2.6	697.0	110.5
2	41.2	0.5	390.6	99.3
3	197.5	3.7	1001.2	507.1
4	123.1	1.8	423.1	307.1
5	96.4	1.2	390.5	266.5
6	146.9	4.3	804.7	334.2
7	59.6	3.8	362.0	162.7
8	80.3	2.7	551.4	172.2
9	80.7	1.4	924.8	179.7
10	169.0	2.1	1859.9	425.0
11	120.3	14.1	1341.8	254.4
12	65.2	1.0	535.0	162.7
13	61.6	10.3	655.5	133.4
14	42.9	3.5	405.7	104.1
15	236.1	2.6	1836.7	579.3
16	146.3	3.0	1058.7	337.2
17	98.5	11.3	659.0	197.2
18	91.1	10.9	659.6	191.6
19	177.7	19.5	2255.7	352.8
20	49.8	8.5	398.4	112.1
Total	2139.4	108.8	17211.3	4989.1

Table 10.

Results of Allocated value for inputs and outputs with model (11)

Unit	Allocation value				Reduction proportion			
	Total fossil fuel Energy consumption (million tce)	Non-fossil fuel consumption (million tce)	GDP (billion RMB at 2005) (million tce)	CO_2 emissions (million tone)	x_1	x_2	y^g	y^b
1	22	1.04	490	0.00	0.59	0.60	0.30	1.00
2	16	0.35	270	0.00	0.60	0.29	0.30	1.00
3	79	1.48	700	0.00	0.60	0.60	0.30	1.00
4	49	1.80	300	0.00	0.60	0.00	0.30	1.00
5	39	0.48	270	0.00	0.60	0.60	0.30	1.00
6	59	4.30	560	0.59	0.59	0.00	0.29	1.00
7	24	3.80	250	0.00	0.60	0.00	0.30	1.00
8	32	2.70	390	0.00	0.59	0.00	0.30	1.00
9	32	1.40	650	0.00	0.59	0.00	0.30	1.00
10	68	2.10	130	0.00	0.59	0.00	0.30	1.00
11	48	14.1	940	0.00	0.59	0.00	0.29	1.00
12	26	1.0	370	0.00	0.59	0.00	0.30	1.00
13	25	10.3	655.5	0.00	0.60	0.00	0.30	0.89
14	17	3.50	460	104.1	0.60	0.00	0.29	0.00
15	190	2.60	280	579.3	0.20	0.00	0.30	0.00
16	59	3.00	130	337.2	0.60	0.00	0.30	0.00
17	39	11.30	740	197.2	0.59	0.00	0.30	0.00
18	36	10.90	460	191.6	0.60	0.00	0.30	0.00
19	71	19.50	1600	352.8	0.62	0.00	0.30	0.00
20	20	8.50	280	112.1	0.60	0.00	0.30	0.00
Total					0.44	0.04	0.30	0.60

to implement policies for their subdivisions. In these systems, the central manager is interested in evaluating all units individually at the same time, so that total input consumption is minimized or total desirable output production is maximized, or to achieve two or more goals as multi-objective functions. When energy consumption is reduced, it will affect both the desirable and undesirable outputs. Regarding environmental pollution control policies, if energy storage does not lead to a reduction in environmental pollution, this indicates that the model has a weakness and needs to be modified. A model had already been proposed that did not reduce environmental pollution by reducing energy consumption, and hence, we modified the model to reduce environmental pollution. Depending on the decision of the central manager to adopt a policy based on energy saving and reduced environmental pollution emissions, in this paper we developed two new general centralized resource allocation models that the manager can choose from. The first model is modified such that Δy^b can receive a positive value and become feasible. The second model is defined based on the idea that the required changes should be applied to the totality of the desirable outputs. It is not logical to reduce individual desirable outputs, as the reduction of undesirable outputs may not be within the acceptable standard range. In each of the presented models, the undesirable outputs are changed by a larger proportion than the desirable outputs. We added goal programming to the problem so as to prevent the infeasibility of the problem. We also analyzed our model through a real example of 20 Chinese regions. The results showed that the proposed methods significantly reduced the CO_2 emissions compared with the competing model. These models can be effective in preventing energy waste and protecting the environment. The second EU (European Union) clean air outlook report looks at the prospects for EU member states' air quality up to 2050. According to the European Commission targets, by 2030, the amount of greenhouse gases in EU member states will be reduced by 55% compared with 1990 [47]. To achieve this target, manufacturing industries in the EU must purchase permits to produce a certain amount of greenhouse gases. Any industrial unit that produces less harmful gas than its allowed amount can sell its remaining permits to other units and benefit from it. Any plant that produces more harmful gas than its allowed amount will have to buy more permits. In other words, there is a trade-off between industrial units. Therefore, the authors suggest the model presented in this paper to reduce pollution in industrial units under the supervision of the EU. The total amount of permits issued can be considered as the amount obtained after reallocation for environmental pollution in model (11). This means that the allowable amount of pollution considered for all industrial units should be equal to the allocated amount of undesirable outputs from model (11), and the same number of permits should be issued. Furthermore, the proposed models are applicable to any similar system to reduce pollution and save energy. ■

References

1. Sabetghadam M. (2006) Energy and sustainable development in Iran. In: *Sustainable Energy watch. HELIO International*. Available at: https://sustainabledevelopment.un.org/content/documents/854Iran-EN.pdf
2. Wang Z., Chen L., Tao J., Zhang L., Su L. (2010) Satellite-based Estimation of regional particulate matter (PM) in Beijing using vertical-and-RH correcting method. *Remote of Environment*, vol. 114, pp. 50–63. https://doi.org/10.1016/j.rse.2009.08.009
3. Song X.-D., Wang S., Hao C., Qiu J. (2014) Investigation of SO_2 gas adsorption in metal-organic frameworks by molecular simulation. *Inorganic Chemistry Communications*, vol. 46, pp. 277–281. https://doi.org/10.1016/j.inoche.2014.06.003
4. *CO_2 emissions from fuel combustion: Overview 2020* (2020) International Energy Agency (IEA). Available at: https://enerji.mmo.org.tr/wp-content/uploads/2020/08/IEA-CO_2_Emissions_from_Fuel_Combustion_Overview_2020_edition.pdf
5. Wu J., Zhu Q., Liang L. (2016) CO_2 emissions and energy intensity reduction allocation over provincial industrial sectors in China. *Applied Energy,* vol. 166, pp. 282–291. https://doi.org/10.1016/j.apenergy.2016.01.008
6. Emrouznejad A., Yang G.L. (2018) A survey and analysis of the first 40 years of scholarly literature in DEA: 1978–2016. *Socio-Economic Planning Sciences,* vol. 61, pp. 4–8. https://doi.org/10.1016/j.seps.2017.01.008

7. Korhonen P., Syrjanen M. (2004) Resource allocation based on efficiency analysis. *Management Science*, vol. 50, pp. 1134–1144. https://doi.org/10.1287/mnsc.1040.0244

8. Lozano S., Villa G. (2004) Centralized resource allocation using data envelopment analysis. *Journal of Productivity Analysis*, vol. 22, pp. 143–161. https://doi.org/10.1023/B:PROD.0000034748.22820.33

9. Asmild M., Paradi J.C., Pastor J.T. (2009) Centralized resource allocation BCC models. *Omega*, vol. 37, pp. 40–49. https://doi.org/10.1016/j.omega.2006.07.006

10. Mar-Molinero C., Prior D., Segovia M.M. et al. (2009) On centralized resource utilization and its reallocation by using DEA. *Annals of Operations Research*, vol. 221, pp. 273–283. https://doi.org/10.1007/s10479-012-1083-8

11. Lozano S., Villa G., Canca D. (2011) Application of centralized DEA approach to capital budgeting in Spanish ports. *Computers & Industrial Engineering*, vol. 60, pp. 455–465. https://doi.org/10.1016/j.cie.2010.07.029

12. Lotfi F.H., Noora A.A., Jahanshahloo G.R., Geramia J., Mozaffari M.R. (2010) Centralized resource allocation for enhanced Russell models. *Journal of Computational and Applied Mathematics*, vol. 235, no. 1, pp. 1–10. https://doi.org/10.1016/j.cam.2010.05.029

13. Fang L., Zhang C.Q. (2008) Resource allocation based on the DEA model. *Journal of the Operational Research Society*, vol. 59, no. 8, pp. 1136–1141.

14. Du J., Liang L., Chen Y., Bi G. (2010) DEA-based production planning. *Omega*, vol. 38, pp. 105–112. https://doi.org/10.1016/j.omega.2009.07.001

15. Du J., Cook W.D., Liang L., Zhu J. (2014) Fixed cost and resource allocation based on DEA cross-efficiency. *European Journal of Operational Research*, vol. 235, no. 1, pp. 206–214. https://doi.org/10.1016/j.ejor.2013.10.002

16. Amirteimoori A., Kordrostami S. (2012) Production planning in data envelopment analysis. *International Journal of Production Economics*, vol. 140, no. 1, pp. 212–218. https://doi.org/10.1016/j.ijpe.2011.09.025

17. Fang L. (2013) A generalized DEA model for centralized resource allocation. *European Journal of Operational Research*, vol. 228, no. 2, pp. 405–412. https://doi.org/10.1016/j.ejor.2013.01.049

18. Li S.-k., Cheng Y.-s. (2007) Solving the puzzle of structural efficiency. *European Journal of Operational Research*, vol. 180, no. 2, pp. 713–722. https://doi.org/10.1016/j.ejor.2006.05.010

19. Hong L., Yang W., Zhou Z., Huang C. (2013) Resource allocation model's construction for the reduction of undesirable outputs based on DEA methods. *Mathematical and Computer Modelling*, vol. 58, nos. 5–6, pp. 913–926. https://doi.org/10.1016/j.mcm.2012.10.026

20. Zhu Q., Wu J., Li X., Xiong B. (2017) China's regional natural resource allocation and utilization a DEA-based approach in a big data environment. *Journal of Cleaner Production*, vol. 142, no. 2, pp. 809–818. https://doi.org/10.1016/j.jclepro.2016.02.100

21. Wang H., Bi J., Wheeler D., Wang J., Cao D., Lu G., Wang Y. (2004) Environmental performance rating and disclosure: China's Green Watch program. *Journal of Environmental Management*, vol. 71, no. 2, pp. 123–133. https://doi.org/10.1016/j.jenvman.2004.01.007

22. Sun J., Wang Z., Zhu Q. (2020) Analysis of resource allocation and environmental performance in China's three major urban agglomerations. *Environmental Science and Pollution Research*, vol. 27, pp. 34289–34299. https://doi.org/10.1007/s11356-020-09665-5

23. Shi Z., Huang H., Chiu Yh., et al. (2021) Linkage analysis of water resources, wastewater pollution, and health for regional sustainable development – using undesirable three-stage dynamic data envelopment analysis. *Environmental Science and Pollution Research*, vol. 28, pp. 19325–19350. https://doi.org/10.1007/s11356-020-12067-2

24. Tamiz M., Jones D., Romero C. (1998) Goal programming for decision making: An overview of the current state-of-the-art. *European Journal of Operational Research*, vol. 111, pp. 569–581. https://doi.org/10.1016/S0377-2217(97)00317-2

25. Bal H., Örkcü H.H. (2007) A goal programming approach to weight dispersion in data envelopment analysis. *Gazi University Journal of Science*, vol. 20, pp. 117–125. Available at: https://dergipark.org.tr/en/pub/gujs/issue/7399/96859

26. Bal H., Örkcü H.H., Celebioglu S. (2010) Improving the discrimination power and weights dispersion in the data envelopment analysis. *Computers and Operations Research*, vol. 37, pp. 99–107. https://doi.org/10.1016/j.cor.2009.03.028

27. Ghasemi M.-R., Ignatius J., Emrouznejad A. (2014) A bi-objective weighted model for improving the discrimination power in MCDEA. *European Journal of Operational Research*, vol. 233, pp. 640–650. https://doi.org/10.1016/j.ejor.2013.08.041

28. Yang J.B., Wong B.Y.H., Xu, D.-L., Stewart T.J. (2009) Integrating DEA-oriented performance assessment and target setting using interactive MOLP methods. *European Journal of Operational Research*, vol. 195, pp. 205–222. https://doi.org/10.1016/j.ejor.2008.01.013

29. Lotfi F.H., Jahanshaloo G.R., Ebrahimnejad A., Soltanifar M., Manosourzadeh S.M. (2010) Target setting in the general combined-oriented CCR model using an interactive MOLP method. *Journal of Computational and Applied Mathematics*, vol. 234, pp. 1–9. https://doi.org/10.1016/j.cam.2009.11.045

30. Jahanshahloo G.R., Lotfi F.H., Moradi M. (2005) A DEA approach for fair allocation of common revenue. *Applied Mathematics and Computation*, vol. 160, pp. 719–724. https://doi.org/10.1016/j.amc.2003.11.027

31. Chiang C.I., Hwang M.J., Liu Y.H. (2011) Determining a common set of weights in DEA problem using a separation vector. *Mathematical and computer modeling*, vol. 54, pp. 2464–2470. https://doi.org/10.1016/j.mcm.2011.06.002

32. Lotfi F.H., Hatami-Marbini A., Agrell P.J., Aghayi N., Gholami K. (2013) Allocating fixed resources and setting targets using a common-weights DEA approach. *Computers & Industrial Engineering*, vol. 64, no. 2, pp. 631–640. https://doi.org/10.1016/j.cie.2012.12.006

33. Fang L., Li H. (2015) Multi-criteria decision analysis for efficient location-allocation problem combining DEA and goal programming. *RAIRO-Operations Research*, vol. 49, pp. 753–772. https://doi.org/10.1051/ro/2015003

34. Charnes A., Cooper W.W., Rhodes E. (1978) Measuring the efficiency of decision making units. *European Journal of Operational Research*, vol. 2, pp. 429–444.

35. Golany B. (1988) An interactive MOLP procedure for the extension of DEA to effectiveness analysis. *The Journal of the Operational Research Society*, vol. 39, no. 8, pp. 725–734. https://doi.org/10.1057/jors.1988.127

36. Hwang C.L., Masud A.S.M. (1979) *Multiple objective decision making – Methods and applications: A State-of-the-art survey*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-45511-7

37. Steuer R.E. (1986) *Multiple criteria optimization: Theory, computation, and application*. New York: Wiley.

38. Hatami-Marbini A., et al. (2015) A common weights data envelopment analysis for centralized resource reduction and target setting. *Computers & Industrial Engineering*, vol. 79, pp. 195–203. https://doi.org/10.1016/j.cie.2014.10.024

39. Davoodi A., Zhiani Rezai H. (2012) Common set of weights in data envelopment analysis: a linear programming problem. *Central European Journal of Operational Research*, vol. 20, no. 2, pp. 355–365. https://doi.org/10.1007/s10100-011-0195-6

40. Athanassopoulos D. (1995) Goal programming & data envelopment analysis (GODEA) for target-based multi-level planning: Allocating central grants to the Greek local authorities. *European Journal of Operational Research*, vol. 87, no. 3, pp. 535–550. https://doi.org/10.1016/0377-2217(95)00228-6

41. Golany B., Tamir E. (1995) Evaluating efficiency-effectiveness-equality trade-offs: A data envelopment analysis approach. *Management Science*, vol. 41, no. 7, pp. 1172–1184. Available at: http://www.jstor.org/stable/2632774

42. Seiford L.M., Zhu J. (2002) Modeling undesirable factors in efficiency evaluation. *European Journal of Operational Research*, vol. 142, no. 1, pp. 16–20. https://doi.org/10.1016/S0377-2217(01)00293-4

43. Färe, R., Grosskopf, S. and Hernandez-Sancho, F. (2004) Environmental performance: an index number approach. *Resource and Energy Economics*, vol. 26, no. 4, pp. 343–352. https://doi.org/10.1016/j.reseneeco.2003.10.003

44. Zhou P., Ang B.W., Poh K.L. (2008) Measuring environmental performance under different environmental DEA technologies. *Energy Economics*, vol. 30, no. 1, pp. 1–14. https://doi.org/10.1016/j.eneco.2006.05.001

45. Wang J., Wang S., Li S., Cai Q., Gao S. (2019) Evaluating the energy-environment efficiency and its determinants in Guangdong using a slack-based measure with environmental undesirable outputs and panel data model. *Science of the Total Environment*, vol. 663, pp. 878–888. https://doi.org/10.1016/j.scitotenv.2019.01.413

46. Wang K., Zhang X., Wei Y.-M., Yu S. (2013) Regional allocation of CO_2 emissions allowance over provinces in china by 2020. *Energy policy*, vol. 54, pp. 214–229. https://doi.org/10.1016/j.enpol.2012.11.030

47. Jager-Waldau A., Kougias I., Taylor N., Thiel Ch. (2020) How photovoltaic can contribute to GHG emission reductions of 55% in the EU by 2030. *Renewable and Sustainable Energy Reviews*, vol. 126, Article ID 109836. https://doi.org/10.1016/j.rser.2020.109836

About the authors

Sima Madadi

Doctoral Student, Department of Mathematics, Science and Research Branch, Islamic Azad University, Shahid Sattari Square, Tehran 14515/775, Iran;

E-mail: sima.madadi@gmail.com

ORCID: 0000-0002-2558-7501

Farhad Hosseinzadeh Lotfi

Ph.D.

Professor, Department of Mathematics, Science and Research Branch, Islamic Azad University, Shahid Sattari Square, Tehran 14515/775, Iran;

Editor-in-Chief, Journal of New Research in Mathematics, Science and Research Branch, Islamic Azad University Square, Shohadaye Hesarak Boulevard, North Sattari Highway, Tehran, Iran;

E-mail: farhad@hosseinzadeh.ir

ORCID: 0000-0001-5022-553X

Mehdi Fallah Jelodar

Ph.D.

Associate Professor, Department of Mathematics, Ayatollah Amoli Branch, Islamic Azad University, 5 Km of the old road from Amol to Babol, Amol 678, Iran;

E-mail: Mehdi.fallah_jelodar@yahoo.com

ORCID: 0000-0002-8473-564X

Mohsen Rostamy-Malkhalifeh

Ph.D.

Associate Professor, Department of Mathematics, Science and Research Branch, Islamic Azad University, Shahid Sattari Square, Tehran 14515/775, Iran;

Editor-in-Chief, International Journal of Data Envelopment Analysis (IJDEA), Science and Research Branch, Daneshgah Blvd, Simon Bulivar Blvd, Tehran, Iran;

E-mail: Mohsen_rostamy@yahoo.com

ORCID: 0000-0001-6105-7674