ORIGINAL SCIENTIFIC PAPER
UDK: 336.225(497.11)
005.591.6:004]:336.1/.5(497.11)
DOI: 10.5937/EKOPRE1808115A
Date of Receipt: December 7, 2018

Jasna Atanasijević*
University of Novi Sad
Faculty of Sciences
Department of Mathematics and Informatics

Dušan Jakovetić
University of Novi Sad
Faculty of Sciences
Department of Mathematics and Informatics

Nataša Krejić
University of Novi Sad
Faculty of Sciences
Department of Mathematics and Informatics

Nataša Krklec-Jerinkić
University of Novi Sad
Faculty of Sciences
Department of Mathematics and Informatics

Dragana Marković
Tax Administration of the Republic of Serbia
Belgrade

USING BIG DATA ANALYTICS TO IMPROVE EFFICIENCY OF TAX COLLECTION IN THE TAX ADMINISTRATION OF THE REPUBLIC OF SERBIA**

Upotreba analize masovnih podataka za unapređenje efikasnosti naplate poreza u Poreskoj upravi Republike Srbije***

Abstract

Tax evasion poses a major problem for the overall business environment in every economy - it endangers competition, reduces resources for budget-funded public goods and services, and public policies enforcement as well. Moreover, fundamental human rights are denied in an informal labour market. The reform of the Tax Administration of the Republic of Serbia, which has been implemented since 2015, has already contributed to the increase in efficiency of tax collection over the past years; however, there is still significant room for improvement of tax collection, especially individual income taxes, social security contributions, and value added tax. One of the pillars of the Tax Administration Reform refers to the improvement of the analytic function regarding risk management aiming to make the tax control function more efficient, and raise the awareness of voluntary tax declaration. The paper presents the first results of a joint scientific-research project between the Tax Administration and the Faculty of Sciences of the University of Novi Sad, aiming to develop the algorithms for detecting the risk of tax evasion by using advanced methods in big data analytics and the development of artificial intelligence with the help of machine learning. The presented indicator is based on the weighted norm distance between income distribution in a legal entity and the average income distribution in the business sector which it operates in. The results show the sound performance of the developed indicator. In addition to improving the efficiency of field control, the approach is also going to enable an affirmative approach to those taxpayers who are classified in a low-risk category in terms of tax evasion. Furthermore, additional positive effects are expected based on higher self-reporting of risky categories due to higher probability of being a subject of a field control and detecting tax evasion.

Keywords: *big data, tax evasion, risk management in tax collection, machine learning.*

Sažetak

Poreska evazija predstavlja krupan problem za sveukupno poslovno okruženje u svakoj privredi – ugrožava konkurenciju, smanjuje izvore za budžetski finansirana javna dobra i usluge i za sprovođenje javnih politika, a na neformlanom tržištu rada uskraćuje ljudima elementarna prava. Reforma Poreske uprave Republike Srbije koja se sprovodi od 2015. godine već je doprinela povećanju efikasnosti i naplati poreza u prethodnim godinama, ali i dalje postoji značajan prostor za unapređenje naplate, naročito poreza i doprinosa koji se plaćaju na dohodak fizičkih lica i poreza na dodatu vrednost. Jedan od stubova reforme Poreske uprave odnosi se na unapređenje analitičke funkcije za potrebe upravljanja rizicima kako bi se efikasnije koristili kapaciteti kontrole, ali i jačala svest o dobrovoljnom prijavljivanju poreza. U ovom radu predstavljeni su prvi rezultati zajedničkog naučno-istraživačkog projekta Poreske uprave i Prirodno-matematičkog fakulteta Univerziteta u Novom Sadu čiji je

* The authors would like to thank Olivera Pavlović, Branislav Ivošev and Vladimir Jokić for assisting in data processing and creating technical preconditions for data processing; Nataša Arsić and all the employees at the Department of Strategic Risk Analysis; Rade Šević and Miško Marković from the Transformation Sector in Tax Administration for their useful suggestions and comments regarding the approach and the results of the research. Any remaining flaws are the sole responsibility of the authors.

** The research is supported by the Ministry of Science, Education and Technological Development of the Republic of Serbia, grant no. 174030.

*** Ovo istraživanje je podržano od strane Ministarstva prosvete, nauke i tehnološkog razvoja, projekat broj 174030.

cilj razvoj algoritama za detekciju rizika od poreske evazije upotrebom naprednih metoda analize masovnih podataka (big data) i razvoj veštačke inteligencije uz pomoć mašinskog učenja. Predstavljeni pokazatelj zasniva se na merenju ponderisane distance raspodele zarada u pojedinačnom privrednom subjektu od prosečne raspodele zarada u oblasti delatnosti u kojoj posluje. Rezultati ukazuju na dobre performanse razvijenog indikatora. Pored efekta na unapređenje efikasnosti terenske kontrole, ovaj pristup omogućiće i afirmativni pristup prema onim poreskim obveznicima koji spadaju u ocenjenu niskorizičnu kategoriju po pitanju poreske evazije. Takođe, očekuju se i dodatni pozitivni efekti po osnovu samoprijavljivanja rizičnih kategorija zbog veće verovatnoće da će biti predmet terenske kontrole i otkrivanja poreskog prekršaja.

Ključne reči: *masovni podaci, poreska evazija, upravljanje rizicima u naplati poreza, mašinsko učenje.*

Introduction

Tax evasion is one of the major obstacles to increasing the competitiveness of an economy. It directly and negatively affects the conditions for business activities in the market for the companies that legally declare and pay taxes, making their production costs, and, consequently, the price of their products and services higher compared to the prices of competitors who do not pay taxes and contributions. Moreover, tax evasion, due to relatively low budget revenues, directly erodes the space for improving the level and quality of public goods and services and citizens' satisfaction. From the perspective of individuals working in the informal sector, it generates the conditions of insecurity and denial of fundamental human rights, that is, the features of a modern society (health, retirement, disability insurance, etc.).

Tax evasion is directly related to the level of grey economy in a country. Empirically proved, grey economy is higher in the countries with a lower level of income per capita [1]. Additionally, the level of tax evasion is higher in the tax systems where the structure of taxation implies more the taxation of production factors than the taxation of consumption. The reason for this is that tax evasion regarding production factors, especially income, is much simpler.

Although grey economy has been significantly reduced in the period between two consecutive assessments, its level in Serbia remains relatively high causing very high losses in terms of unpaid budget revenues. Namely, according to the empirical assessment of grey economy in Serbia, especially product turnover and salary payments of registered legal entities, there was a decrease in its level from 21.2% in 2012 to 15.4% of GDP in 2017 when two consecutive assessments were conducted [2]. The same study estimated, based on a survey on the perception of registered legal entities, that there was still a high share of unregistered legal entities that only operate in the black market (17.2%).

According to the standard model of tax evasion Allingham-Sandmo [3], [4] and empirical researches [4], [5], the degree of tax collection depends on the level of tax rates, the level of penalties and most of all, on the probability of detecting tax evasion.

Within the framework of the Tax Administration Reform which is being implemented in the period 2015-2020 based on the Programme for Transformation of Tax Administration, adopted in 2015, the main objectives are to increase the efficiency of tax collection, improve the service quality, reduce the costs of fulfilling tax obligations and improve institutional capacities through a better organization, information system and adequate human resources [6].

Over the past three years, the performance of the Tax Administration has been significantly improved through a number of measures, and the most important among them are: the introduction of electronic reporting of all tax obligations, strengthening the advisory role of employees, raising the awareness of voluntary compliance with the regulations and reporting taxes, identifying secondary activities and then focusing on the core business, reducing the number of branches which deal with the core activities (from 78 to 37), strengthening the advisory role of inspectors, and improving data quality through the development of registers and data exchange with other institutions.

One of the most important ways of improving Tax Administration efficiency is based on strengthening the analytical function through the work of the Strategic Risk Department in order to improve risk detection and optimize the performance of tax control tasks by concentrating on the riskiest categories of taxpayers.

Big data are considered to be a new type of resource, i.e., assets in business operation and the largest source for

productivity growth in the upcoming period, as stated in a reference report by the McKinsey Global Institute in 2011 [7]. Theoretical and empirical papers on income distribution are available in the literature. Since then big data has been used to improve business processes and increase efficiency and productivity, mostly in the private sector, primarily telecommunication data and the data from the Internet and social networks. The aforementioned McKinsey Institute report estimates that along with the financial sector, the largest area for increasing productivity by using big data – is in the public sector, that is, state administration.

Since 2018, the most advanced big data analytics has being introduced and used in the risk management system in the Tax Administration, which includes the development of algorithms based on the so-called machine learning, enabling the introduction of artificial intelligence as well. The advanced stage of the research will include deep learning methods according to which an analysis of the behaviour of legal entities by using big data is going to be conducted. In order to provide sustainable work on the introduction of these methods in the activities of the Tax Administration as well as reliance on appropriate expertise, in 2018 the Tax Administration signed a Business Cooperation Agreement on Scientific Research with the Faculty of Sciences – University of Novi Sad (Department of Mathematics and Informatics) under the project "Detecting the risk of evasion of paying individual income taxes based on the appropriate methods by using artificial intelligence". The aim of the project is to develop a series of risk indicators for tax evasion by using the methods for analysing big data on depersonalized data and applying them to the data in the Tax Administration for better direction of the activities and more efficient tax collection.

The aim of the paper is to present the results of the half-year research, conducted in cooperation between the researchers from the Department of Mathematics and Informatics at the Faculty of Sciences in Novi Sad and the Tax Administration of the Republic of Serbia. The research aims to improve the methodology of risk management, which will be used for the assessment and support to risk management of tax evasion in the Tax Administration, continuing to be developed in the following period.

In the second part, the database is presented based on which risk indicators are developed by using mathematical models. It is a tax return base for all types of income earned by citizens. The third part presents a methodological approach to developing risk indicators and an example of risk indicators for tax evasion based on an appropriate metrics that compares the characteristics of distribution of the amount of net income within a single legal entity with the income distribution for all employees in the line of business which the legal entity operates in. Finally, the results are discussed and, in the concluding remarks, there is a brief description of a future research based on the methods of machine learning that will be applied to big data aiming to improve the methods for detection and assessment of the risk levels of tax evasion among single legal entities.

Description of the database

The development of the algorithms for risk indicators and their testing is conducted by using depersonalized individual data (excluding personal data of income recipients and income payers) based on the tax returns from the unified tax collection database. To develop and test these risk indicators, the data from the tax returns for the period from April 2014 to May 2018 were previously prepared in an appropriate way in accordance with the regulations and principles on personal data protection. The so-called data anonymization was performed. Personal identification numbers of individuals and tax identification numbers of legal entities, after having added basic data to each legal entity from the register of tax identification numbers (business activity code, municipality, year of establishment and a form of organization), were encrypted by the authorized persons in the Tax Administration. Personal data from different sources were not combined by using these personal identifiers. All researchers who participated in the research were employed at the University and signed appropriate data confidentiality agreements with the Tax Administration and committed to the exclusive use of the data for scientific purposes in accordance with the agreed project goals.

In the entire database used for the research, there were 6,141,812 income recipients, 234,310 income payers

(unique tax identification numbers, hereinafter: TINs) and 201,635,126 different combinations of income recipients, income payers, tax returns and types of income.

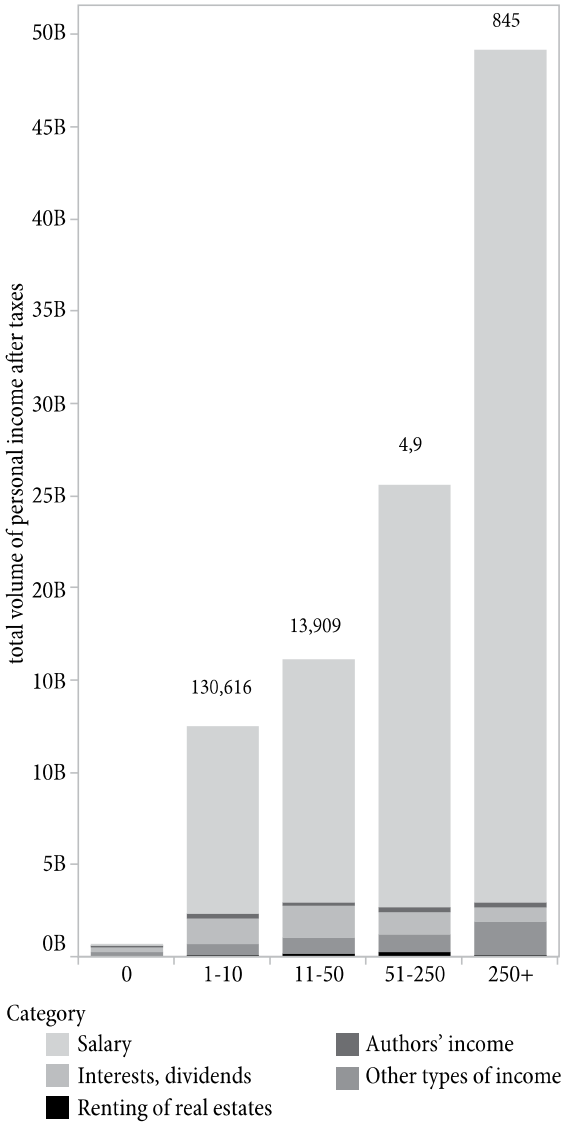
During the whole analysed period, for 36,011 income payers there was a submitted at least one tax return by tax control based on field control findings of tax breach. It refers to approximately 15% of all income payers (unique TINs) that reported their tax obligations in the analysed four-year period.

Only in one whole year, for example in 2017, tax control submitted tax returns for 16,440 employers based on the field control reports out of 157,163 different TINs that reported their income tax obligations.

Table 1 represents the breakdown of income payers (TINs) by type, on the example of March 2017: about 50% are legal entities not financed from the budget, similar to the number of entrepreneurs; 5,174 are legal entities financed from the budget, and only 427 are foreign representatives. In the total number of tax returns the prevailing ones are those related to salary, followed by other incomes (performance based contract and temporary jobs), then sick leaves compensation and capital income (interests and dividends).

When we look at the structure of the total reported net income for which tax obligations were paid in one month (March 2017) according to the type of income, the size of a legal entity measured by the number of employees (Figure 1), the largest volume of income was realized based on salaries and mainly from large legal entities with over 250 employees. There is also a disproportionately higher ratio of capital income (interests and dividends) in relation to salary payments in small legal entities compared to larger ones (this is probably a tax arbitrage because of a lower burden of profit due to salary income). A lot of legal entities with few employees (0 or 1-10) can be noted as well.

Figure 1: The total net personal income after tax, by income categories and the size of the income payer measured by the number of employees, in March 2017, in dinars



Source: Authors` calculation; Tax administration
Note: above each bar there is a number of TINs of a matching size category measured by the number of employees.

Table 1: Structure of tax returns according to the type of income and the type of income payer in March 2017

Type of income payers	Number of income payers (TINs)	Number of tax declarations by type of personal income						
		Salary	Sick leaves compensa-tions	Authors' income	Interests and dividends	Renting of real estate	Other types of income	In total
Legal entities not financed from the budget	76,625	1,608,166	67,492	13,488	75,532	11,989	125,946	1,902,613
Legal entities financed from the budget	5,174	959,926	28,688	12,972	614	1,791	119,334	1,123,325
Foreign representation offices	427	3,829	133	419	262	29	371	5,043
Entrepreneurs	74,933	200,504	17,211	329	14,304	1,875	7,817	242,040
Other	4	58,062	1,251	84	3	0	363	59,763
In total	157,163	2,830,487	114,775	27,292	90,715	15,684	253,831	3,332,784

Methodological approach

The aim of the research on improving the efficiency of tax collection is the development of a number of risk indicators that can be applied to each income payer or income recipient or a particular group of income payers or income recipients. Each risk indicator is based on a specific algorithm that is developed and tested on the data described in section 2. Once developed and tested, the indicators are applied in the Tax Administration on personalized data and used in the business activities of the relevant Tax Administration services. Tax compliance risk includes the risk of non-declaration of income, the risk of not reporting the full amount of income and the risk of error in reporting tax obligations.

Generally speaking, risk detection algorithms are based on establishing a significant deviation of the "behaviour" of a taxpayer from the expected/average behaviour based on theoretical and/or empirically determined patterns. Therefore, the algorithms applied on depersonalized big data owned by the Tax Administration identify the "risk" deviations of the features of a taxpayer from the expected values. During the analytical procedure of identifying the risk indicators, time and longitudinal components are taken into account, i.e., the expected values of certain features of taxpayers over certain period of time are analysed, as well as the distribution of the value of the corresponding features related to taxpayers (income recipients or payers) at a specific time.

Once established, the algorithms for measuring the risk indicators on the personalized data owned by the Tax Administration from the unified tax database, the values of one or a combination of several risk indicators, are used to prepare the plan for control activity – both for deep analysis of the behaviour of specific "risky" taxpayers and field control. This contributes to the efficiency of tax inspectors' performance that regrettably have limited capacity.

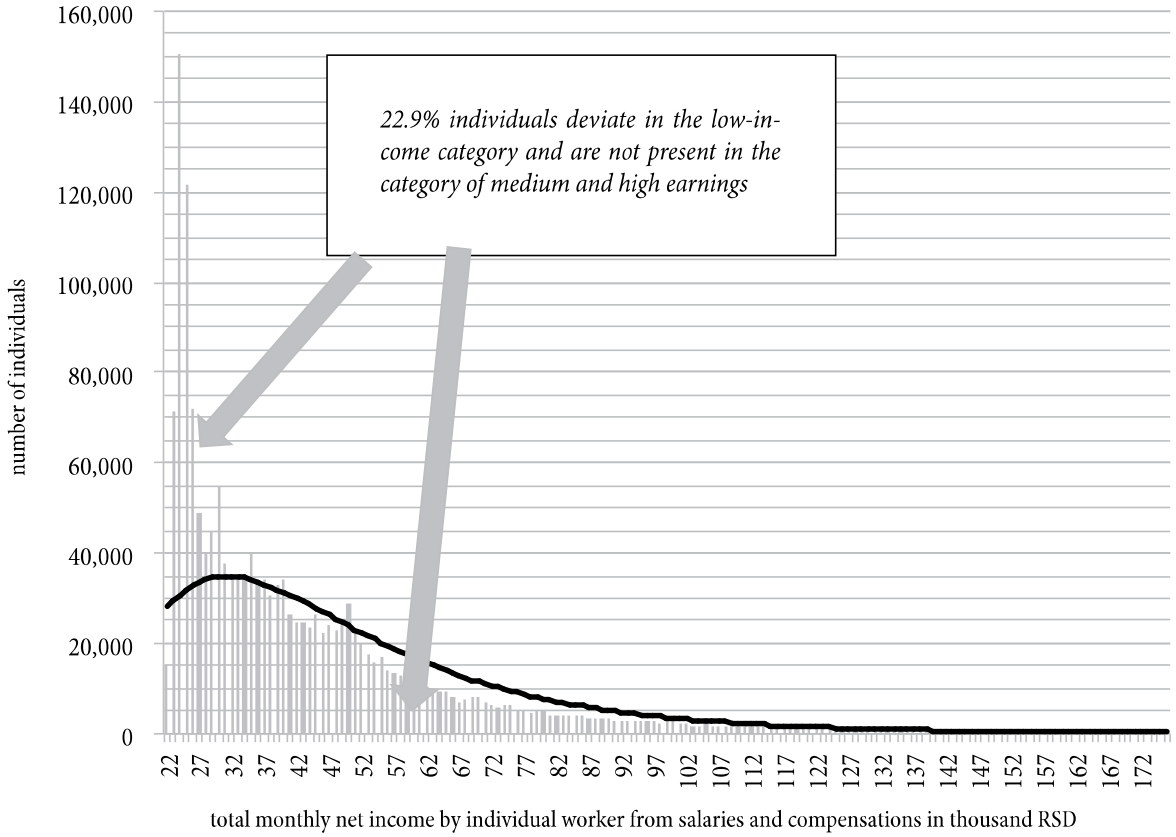
The first type of indicators developed and presented in the paper uses the data on monthly net individual income and is based on the assumption that they follow certain patterns and therefore, that the deviations from the patterns can indicate the risks of tax evasion.

Theoretical and empirical papers on income distribution are available in the literature. An overview of the prevailing theories can be found in Bjerke [8], and other various models have also been analysed in Neal and Rosen [9]. On of the theories of wealth and income distribution among individuals is given in Stiglitz [10]. This theory is based on the idea of identifying the economic factors that lead to income and wealth equalization and the economic factors that lead to the dispersion of income and wealth. Alternative distribution factors are considered, such as consumption function, heterogeneity of working skills, inheritance policy and differences in natural growth depending on income category.

Statistically, income distribution is characterized by skewness and a long right tail, the arithmetic mean of income is higher than median income, and salaries in several right percentiles take a disproportionately large part of total income. Different theories explain these empirical facts. It is clear that income dispersion is the main motivation for investing in personal education and training. A generally accepted theoretical stochastic model involving empirically observed characteristics is that it is a stochastic process of Brown type, that is, the income distribution has the prevailing features of a lognormal distribution.

The distribution of the total monthly net income during one month for the entire Serbian economy indicates some anomalies, as depicted in Figure 2. Regarding the amount of net income ranging from the minimum income up to approximately 25,000 dinars, there is large "density" of employees (i.e., a large share of this category of employees in the total number), whereas in the range of 35-90,000 dinars there is a plunge in the distribution, i.e., rather low frequency of this income category compared to the expected frequency that we would have had if the income had been distributed according to the theoretical (lognormal) distribution that is empirically noticed and frequently seen in literature (more details in section 3). This observation indicates a large number of employees who "receive a minimum salary" or a little higher amount, while their actual salary should have been in this second income category with insufficiently recorded frequency (the difference is probably paid in cash, not into a bank account). This deviation, obtained as the difference between the empirical distribution of total net income

Figure 2: Individual net income (from all income payers) below 175,000 dinars, in March 2017



per employee and the lognormal distribution based on the parameters calculated on the basis of empirical data for the entire population, formally registered as being employed in that month, amounts to 22.9% of the total number of registered employees, that is, approximately 390 thousand employees (income recipients) in March 2017.

Results: development of risk indicators based on the deviation of the distribution in the reported income from the average income distribution in a relevant business sector

The risk indicator presented here is based on the typical income distribution in economy in the specific area of business activity where the legal entity belongs to, and in the legal entity itself.

If we look at the income distribution (as in Figure 2) for individual legal entities, whose size measured by a number of employees exceeds a certain critical size (e.g. 10 employees), certain "diversity" of individual income should also be expected. This "diversity", that is, a fact that there is a certain range of salaries and the structure of all employees neatly divided into different sub-categories within this range, is largely based on the corporate compensation policy and the type of a business activity, i.e., the need to employ certain categories of workers according to the type of work and qualification, which, in turn, reflects on the amount of an individual income and on the appropriate distribution of all salaries in a legal entity. On the other hand, the impact of market forces i.e., the competition in the labour market, leads to a relative equilibrium of wages for specific types of jobs within legal entities operating in a similar line of business.

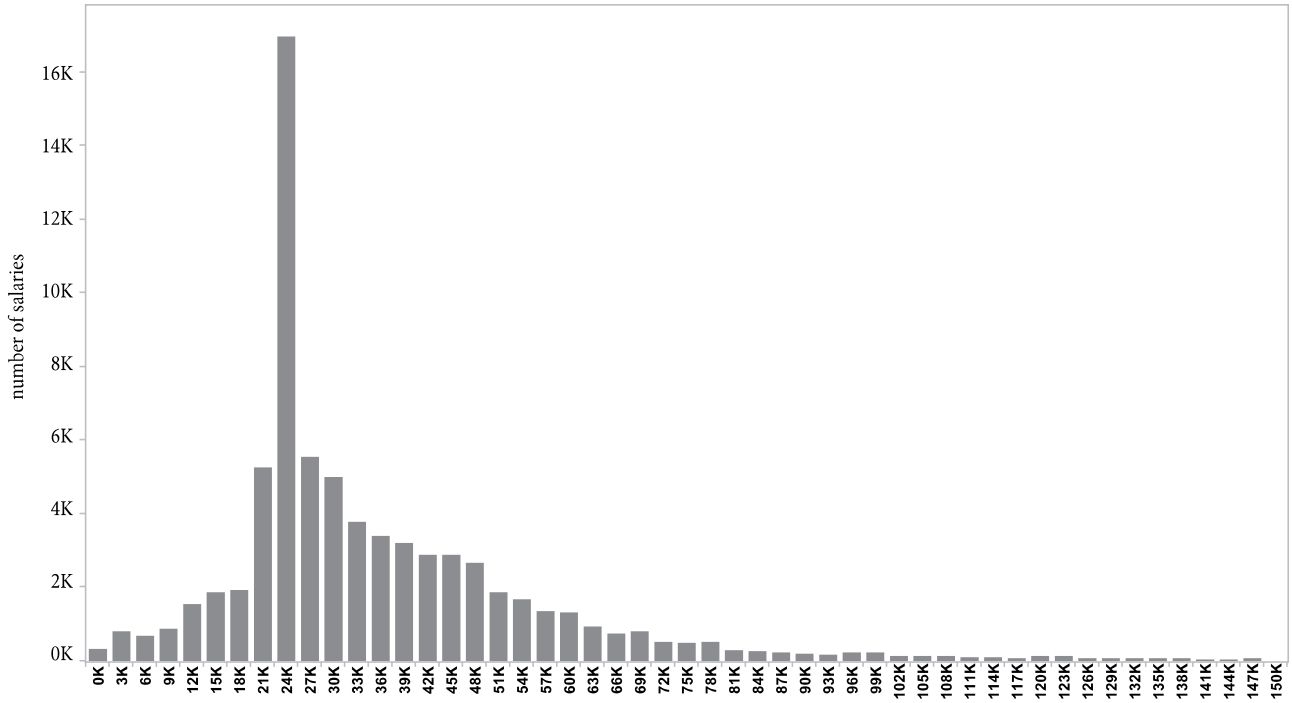
The differences in the distribution of net income per employee in different sectors of economy and in a few areas of business activities are presented in Table 2, as well as in Figures 3 to 7. The value from p10 column (10th percentile) indicates that 10% of employees in this area of business activity belong to the category of income lower than the values shown in the column. The column with the value of p50 implies that a half of the employees in this area of business activity are employed

Table 2: Overview of the statistics for the net income per employee based on salaries and compensations by sector, for March 2017, in dinars

Sector	Number of employers	Average salary	p10	p50	p90	Standard deviation
Agriculture, Forestry and Fishing	2,658	39,704	23,975	34,294	59,662	55,170
Mining	256	158,195	36,434	70,099	191,999	833,054
Manufacturing Industry	24,844	44,324	23,920	32,447	70,162	92,401
Electricity, Gas, Steam and Air Conditioning Supply	328	125,485	47,327	136,267	181,261	59,693
Water Supply; Wastewater Management, Control of Waste Removal Processes	785	41,052	26,820	36,944	58,655	26,718
Construction	8,833	42,225	19,000	30,175	64,307	225,799
Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles	49,807	39,588	22,293	26,849	58,196	116,794
Traffic and Storage	8,020	43,182	23,962	35,000	60,659	50,780
Accommodation and Food Services	11,803	28,172	17,342	25,019	39,725	36,177
Information and Communication	4,196	90,524	25,019	53,071	169,124	247,343
Financial and Insurance Activities	1,772	94,029	24,683	63,982	164,130	202,697
Real Estate	1,046	52,951	22,000	31,752	83,222	331,602
Professional, Scientific, Innovation and Technical Activities	16,465	57,729	22,209	33,145	100,383	154,696
Administrative and Support Service	4,458	39,414	13,304	28,645	58,974	91,994
State Administration and Defence; Compulsory	1,243	54,440	30,568	50,905	85,709	25,487
Social Security Education	3,732	43,755	20,883	45,580	63,407	25,393
Health and Social Care	3,941	44,405	25,264	39,157	72,992	34,409
Art; Entertainment and Recreation	2,885	38,623	24,081	31,424	60,000	31,524
Other services	10,007	37,098	18,003	26,121	65,354	36,148

Source: Authors` calculation; Tax Administration.

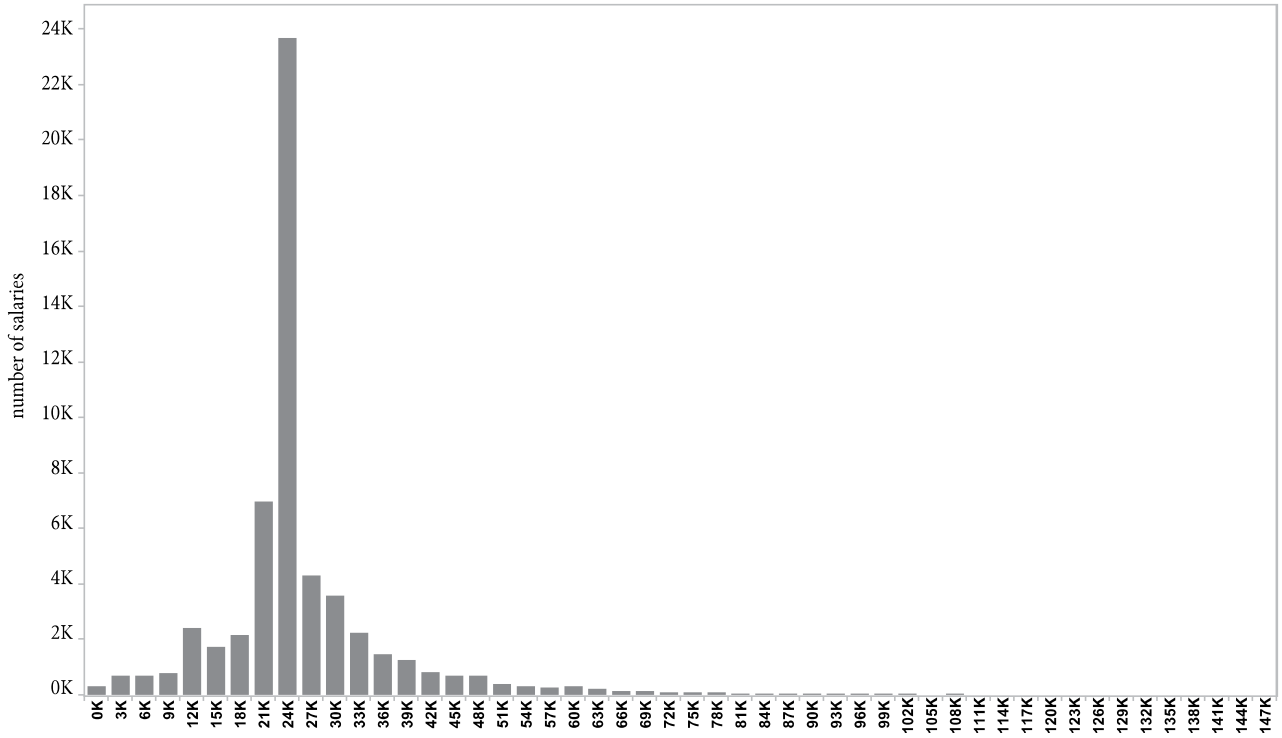
Figure 3: Net income distribution in the Construction Sector, March 2017, in dinars



Source: Authors` calculation; Tax Administration.
Note: Only net income from salaries and compensations of salaries below 150,000 dinars are presented on the histogram.

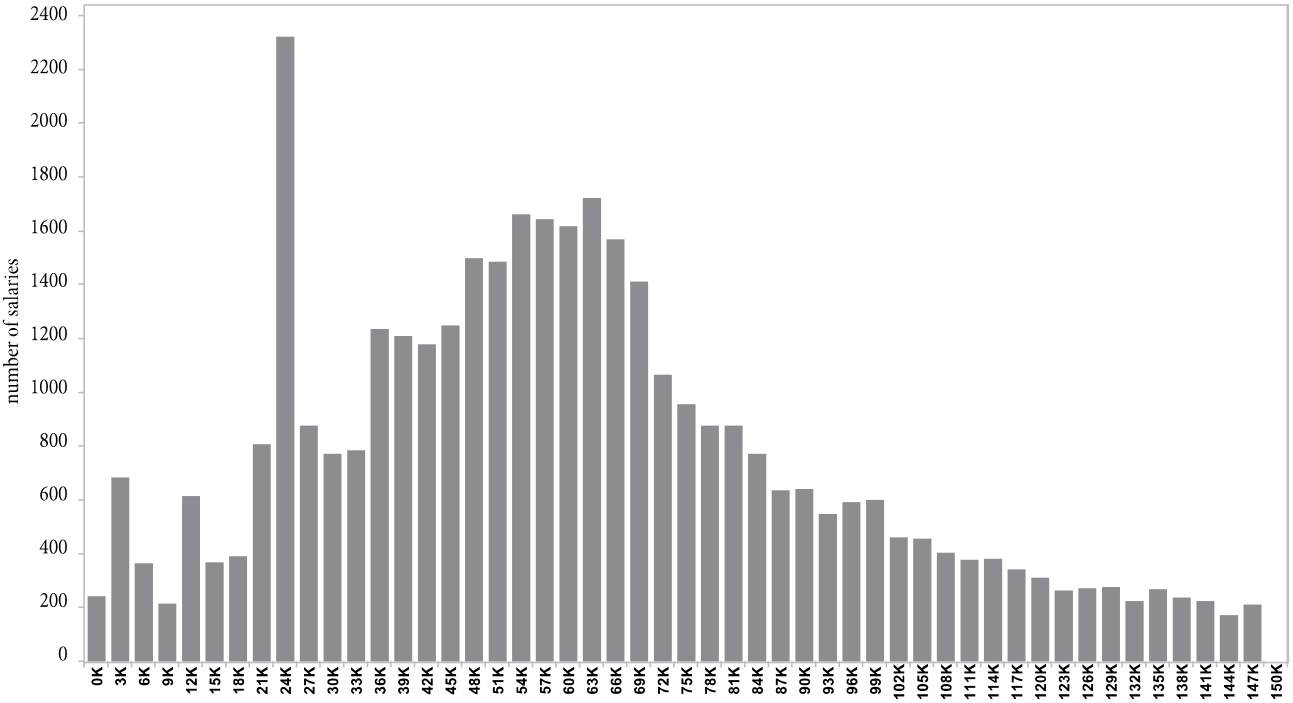
according to the reported net income of lower or equal value related to the corresponding column. The last column – a standard deviation is the way to measure the "dispersion" of distribution. It is calculated as the square root of the sum of the squares of deviations of individual net income from the average net income in the corresponding business activity. Cross-sectoral differences can be clearly noticed.

Figure 4: Net income distribution in the Accommodation and Food Services Sector, March 2017, in dinars



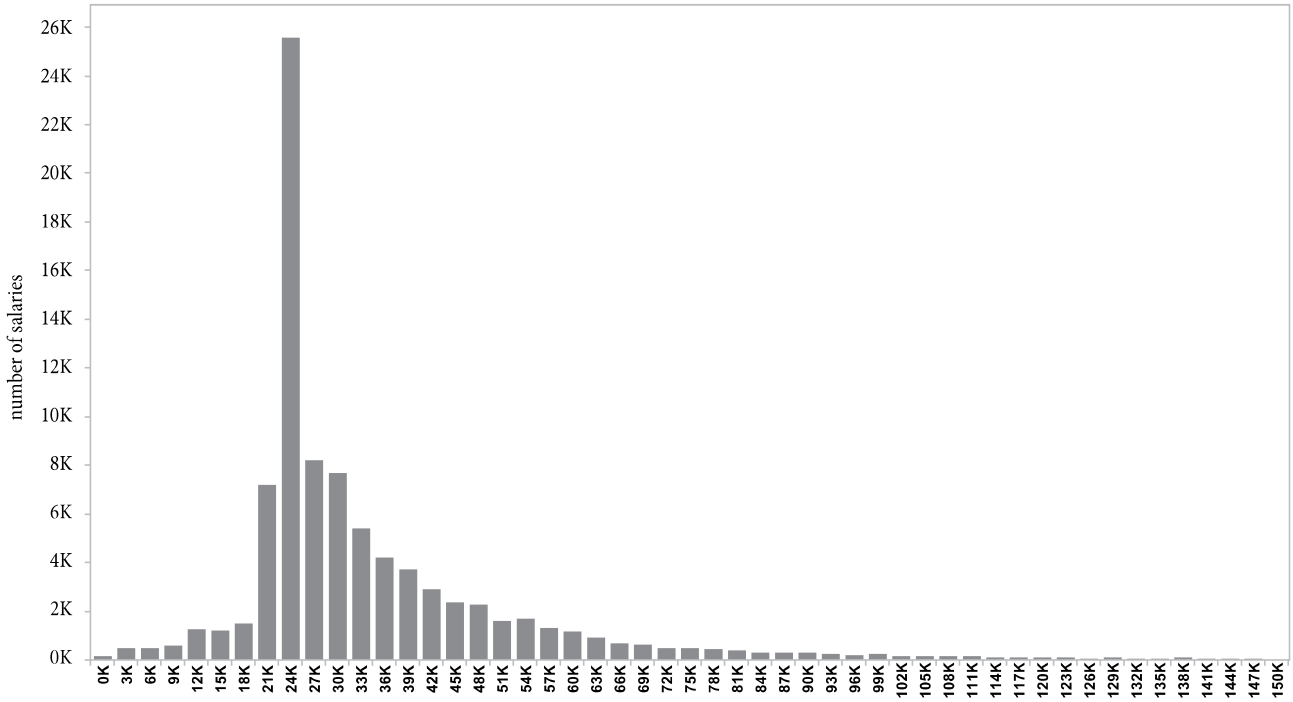
Source: Authors` calculation; Tax Administration.
Note: Only net income from salaries and compensations of salaries below 150,000 dinars are presented on the histogram.

Figure 5: Net income distribution in the Financial and Insurance Sector, March 2017, in dinars



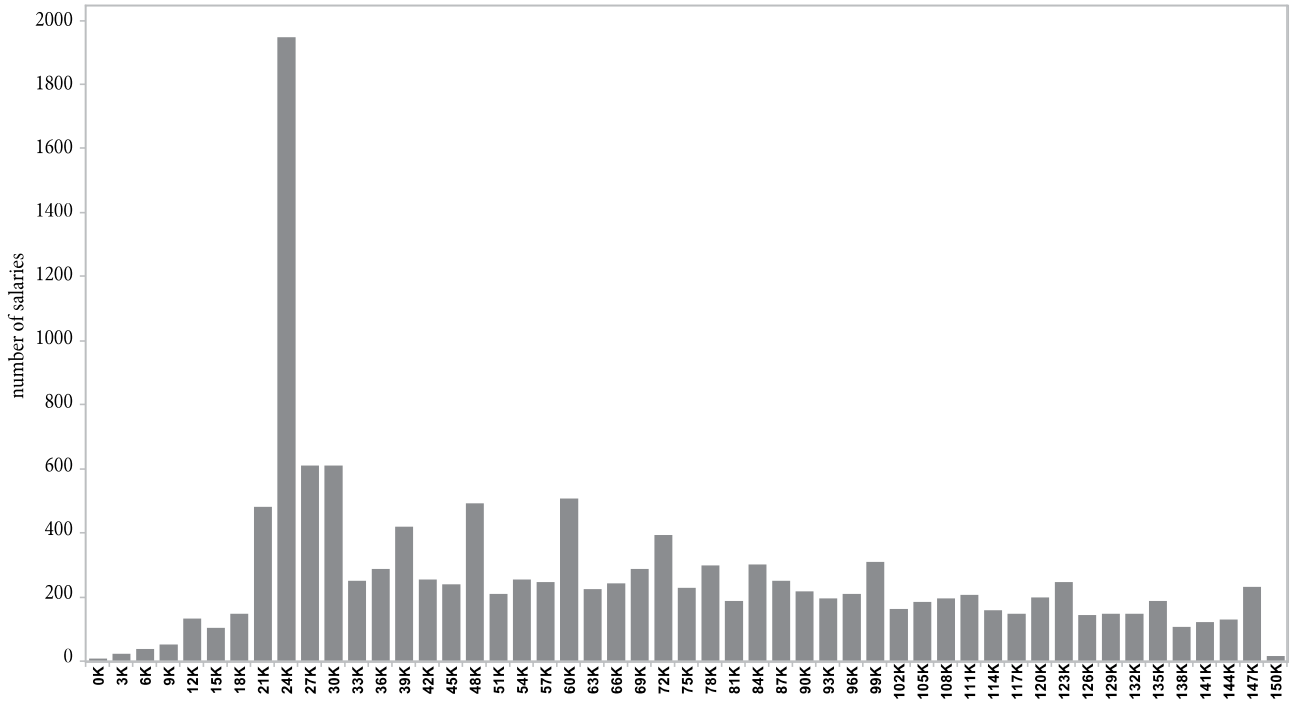
Source: Authors` calculation; Tax Administration.
Note: Only net income from salaries and compensations of salaries below 150,000 dinars are presented on the histogram.

Figure 6: Net income distribution in Beverage and Food Production, March 2017, in dinars



Source: Authors` calculation; Tax Administration.
Note: Only net income from salaries and compensations of salaries below 150,000 dinars are presented on the histogram.

Figure 7: Net income distribution in Computer Programming, March 2017



Source: Authors` calculation; Tax Administration.
Note: Only net income from salaries and compensations of salaries below 150,000 dinars are presented on the histogram.

The initial assumption in the presented analysis is that the income distribution in legal entities corresponds, to a lesser or greater extent, to the income distribution in the relevant economic sector and in the specific area of business activity. The assumption is based on the aforementioned facts – most importantly, the overall

economic environment, the seasonality typical of the sector and the operating of the labour market that leads to approximately equal salaries in the sector.

In accordance with these statements, two risk indicators for tax evasion have been developed, ρ_1 and ρ_2. The indicator ρ1 shows the deviation of a company's income distribution from the distribution of the whole business industry for a month, whereas the indicator ρ_2 has a time dimension and monitors the deviation of the income distribution for the analysed legal entity from the income distribution in its business industry over a certain period of time.

The indicator ρ1 is calculated in the following way. The net income range, with a minimum salary of net amount amounting 15,000 dinars, is divided into intervals (bins) with an increase of 10% so that the first interval includes the net income of 15,000 to 16,500 dinars, the next bin includes salaries from 16,500 to 18,150 dinars etc. The last of 27 bins includes net income exceeding 196,650 dinars. We are going to mark these intervals as (E_i, E_{i+1}). For any chosen month, a vector with 27 components is calculated (corresponding to a histogram) for the entire business activity in the following way. All companies with 10 or more employees were selected. The number of the elements for every bin was calculated by simple counting, $D_1, ..., D_{27}$. Then the vector $d = (d_1,...,d_{27})$ was formed with the components

$$d_i = \frac{D_i}{\sum_{i=1}^{27} D_i}$$

The defined components of the vector *d* have the value in the interval [0,1]. Their sum is 1 and they represent the approximation of the corresponding probabilities (the probability of income being in the *i-th* bin).

Then, for a company with 10 or more employees, a vector corresponding to the net income distribution in that company is formed in the same way. If we mark the number of salaries (or compensations) in *i-th* bin with P_i and define

$$p_i = \frac{P_i}{\sum_{i=1}^{27} P_i}$$

we get a vector $p = (p_1,..., p_{27})$. The risk measure ρ1 for the selected legal entity is defined as the weighted norm distance between vectors d and p in the following way. Starting from the assumption that the risk of tax evasion is higher if there is a significantly higher mass in bins with low salary in the income distribution within the selected legal entity, we defined the weight coefficient $\frac{1}{E_i^2}$ which penalizes a large portion of salaries in low earning intervals, the most in the first bin, and the least in the last bin. Therefore, the risk indicator is calculated as

$$\rho_1 = \sum_{i=1}^{27} \frac{|p_i - d_i|}{E_i^2}$$

The risk indicator ρ_1 obtained in this way shows the deviation of the income distribution in the selected legal entity in relation to the whole line of business. The high value of the indicator ρ_1 indicates a high risk of the existence of tax evasion, as it shows that, compared to the whole line of business, the disproportionate part of low salaries is paid out. On the other hand, the existence of the weight coefficient $\frac{1}{E_i^2}$ for $i = 27, 26,...$ – for sufficiently high salaries, reduces the impact of the difference that is caused if we have an excess of higher salaries in the analysed legal entity.

By monitoring the behaviour of ρ_1 over certain period, we can also define the risk indicator ρ_2 that has a time dimension. By assuming that seasonality is a feature of the whole industry, the change of indicator ρ_1 over certain period of time indicates that the change in the income distribution does not result from the oscillations due to natural seasonality or general market trends, but comes as a consequence of the changes in the policy of paying salaries of the analysed legal entity. We mark with $\rho_1(j)$ the risk indicator ρ1 for the selected legal entity in a certain month. Then, we suppose that we have available data for j = 1, ..., m months. Then we can form the coefficients $\rho_1(1),...,\rho_1(m)$ for the legal entity and define

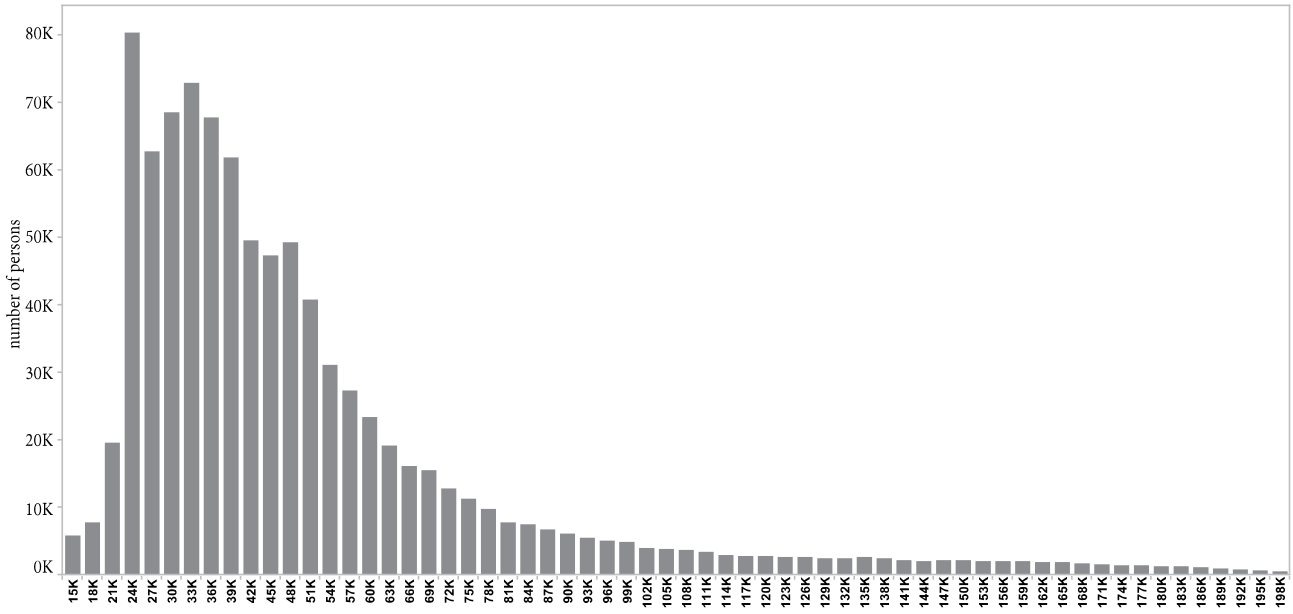
$$\rho_2 = \frac{1}{m} \sum_{j=1}^{m} \rho_1(j)$$

as an average deviation from the expected distribution depending on the line of business. Analogously, we can also consider the variation of the sequence $\rho_1(j)$, as well as the variation coefficient. The high values of these indicators also indicate an increased risk of tax evasion in the payment of salaries.

Descriptive statistics of the risk indicator ρ_1

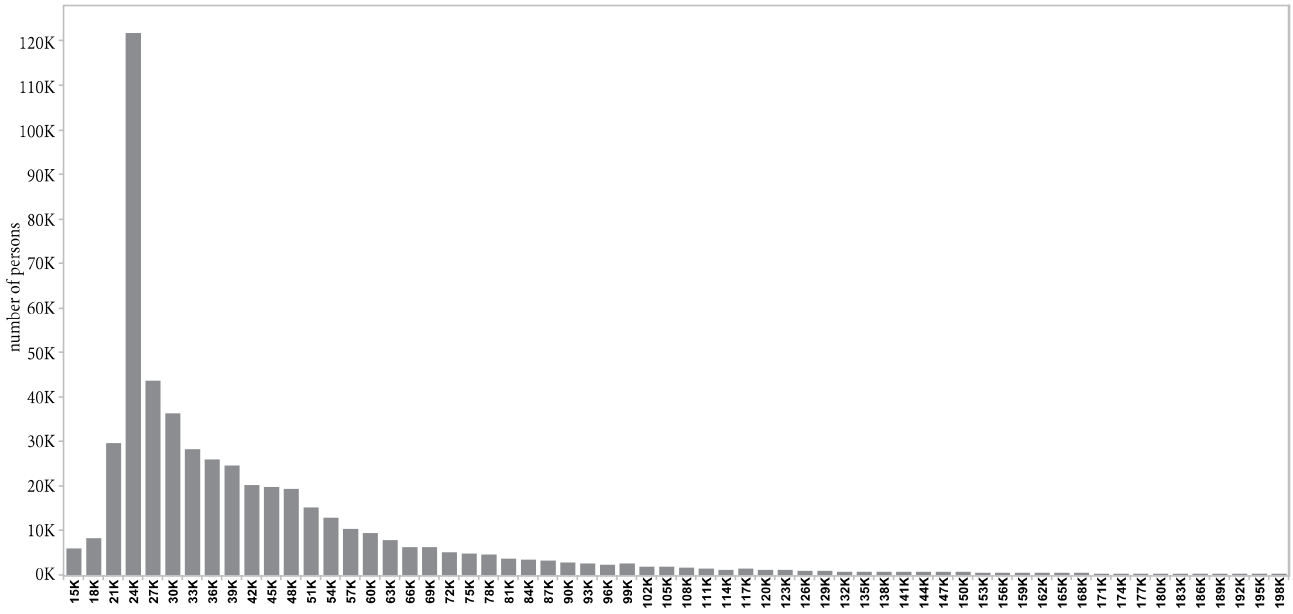
By using the presented methodology for the calculation of the risk indicator ρ_1 on the specific data for the selected accounting period (March 2017), the obtained values are normalized by converting them into a corresponding percentile of distribution so that they are in the range 1-100. Figures 8-10 show individual net income (for net incomes above 15,000 dinars) distribution within legal entities (TINs) that are classified as low, medium and high risk entities regarding the risk of tax evasion. It can be clearly seen that the share of earnings around the official minimal wage is higher with the increase in the value of the risk indicator ρ_1, and that in less "risky" entities the income distribution is much corresponding to the log normal distribution, which is most often cited in theoretical and empirical literature as a referent one.

Figure 8: Individual net income distribution in March 2017 for the income payers with the risk indicator ranging from 0 to 33 (33% of the "low-risk" ones)



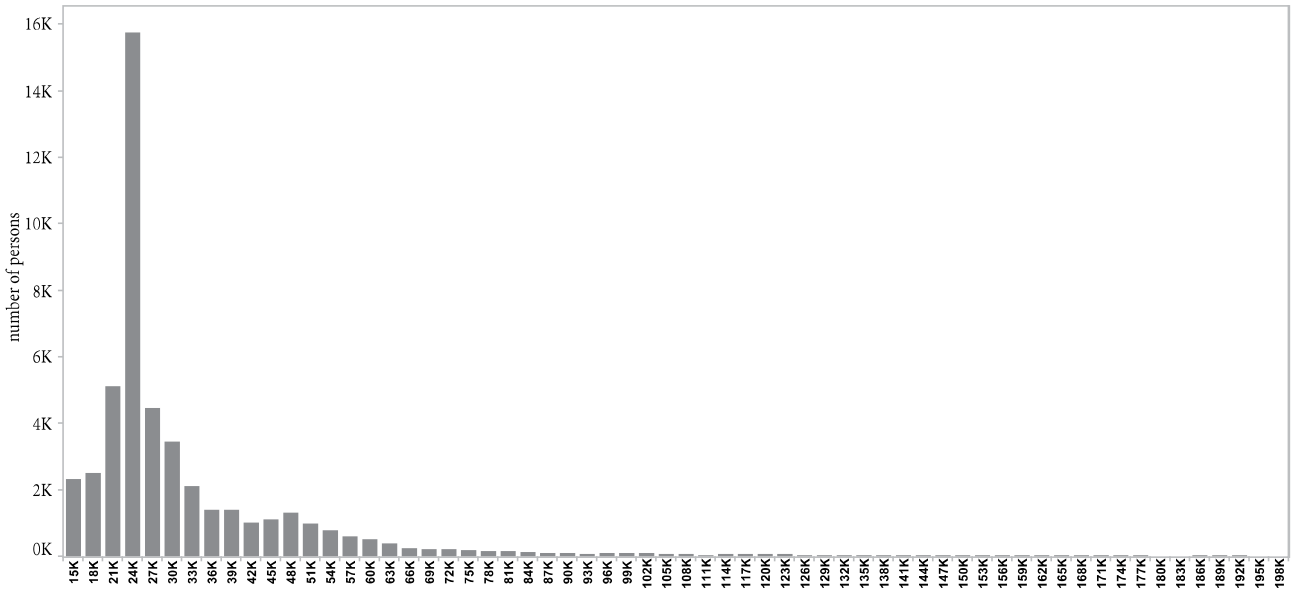
Source: The author.

Figure 9: Net income distribution in March 2017 for the income payers with the risk indicator ranging from 33 to 90 ("medium risk")



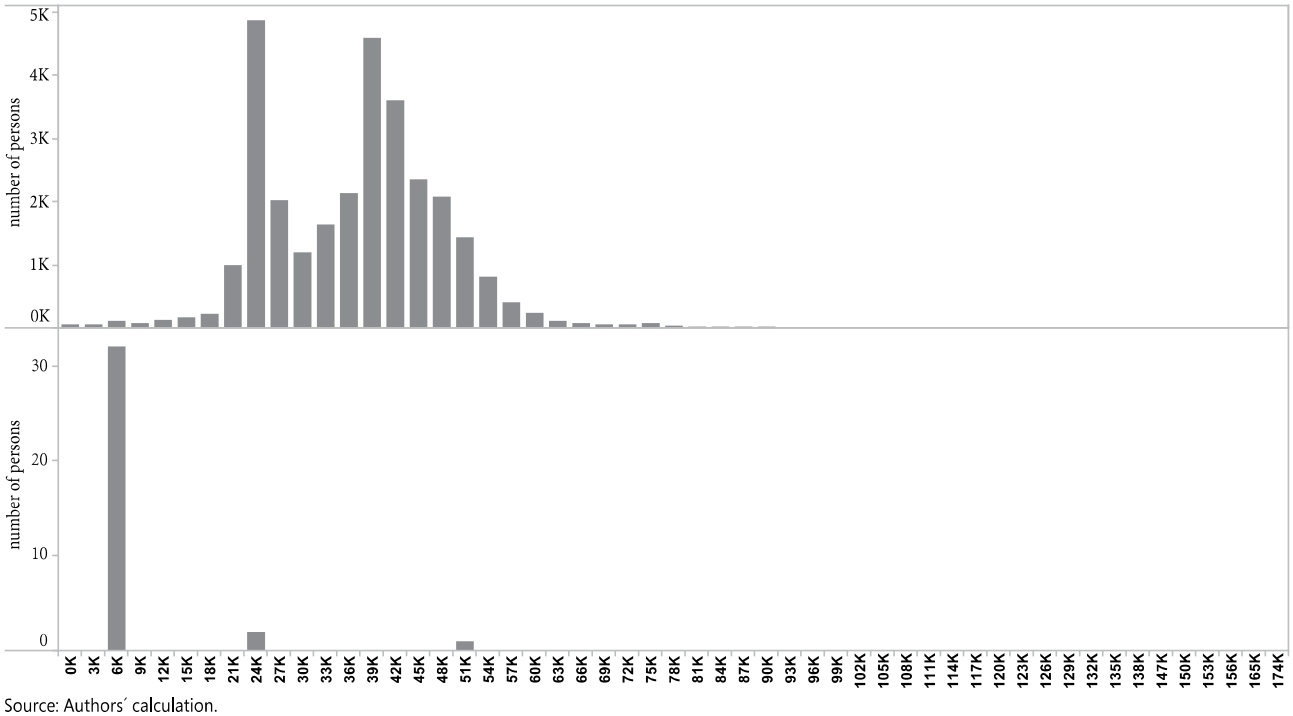
Source: Authors´ calculation

Figure 10: Net income distribution in March 2017 for income payers with the risk indicator higher than 90 (10% of the "high-risk" ones)



Source: Authors´ calculation.

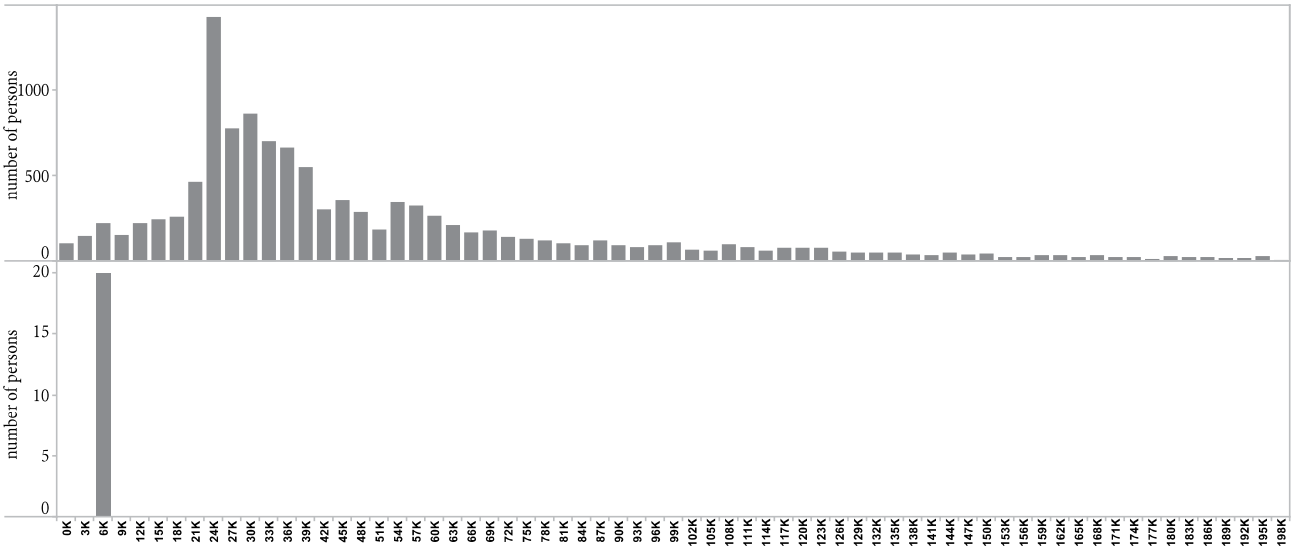
Figure 11: Net income distribution in March 2017 for the area of business activity "Security protection and investigation activities" and the income payer (TIN) with the value indicator ρ1 =99.83



Source: Authors´ calculation.

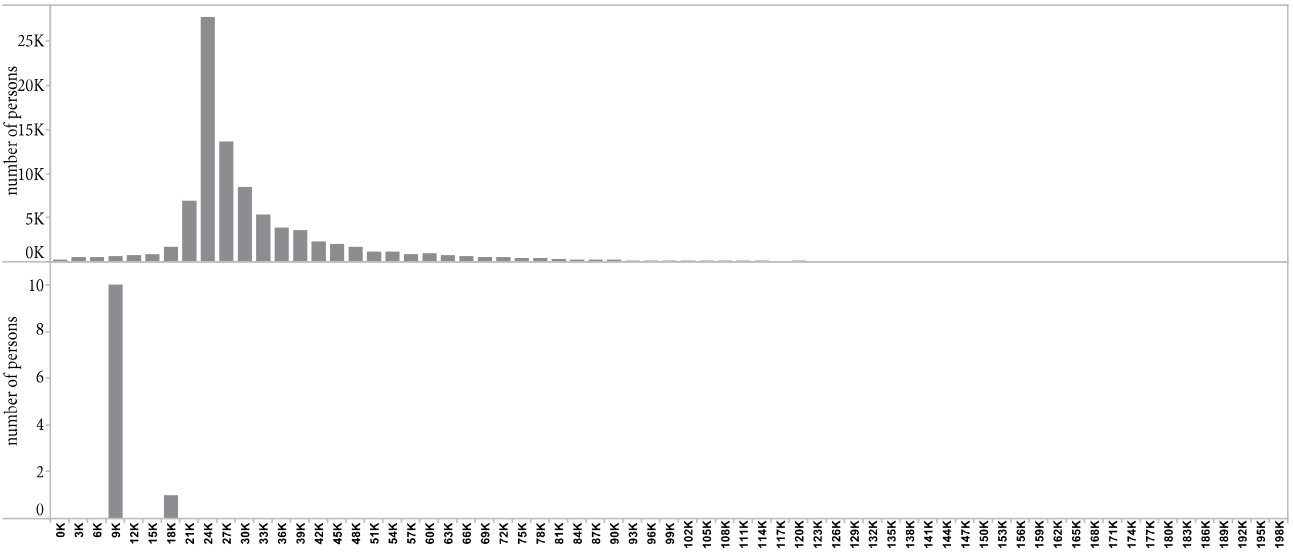
On the example of legal entities with very high values of risk indicator ρ1 and the line of business they operate in, the risk detection method can be even better illustrated, as shown in Figures 11-14. For example, in a company whose net income distribution is shown in the lower part of Figure 11, out of 35 employees, 32 received net income ranging from 6,000 to 9,000 dinars (probably part-time jobs), while two employees were registred to a minimum salary and one employee to a salary within a range from 51,000 to 54,000 dinars. Judging by all this, due to the lack of disparity in salaries and a wide range of different categories of employees with different levels of earnings

Figure 12: Net income distribution in March 2017 for the area of business activity "Management Activities; Management Consultancy" (totalling 251 different TINs) and the income payer (TIN) with the value of indicator ρ1 =99.88



Source: Authors´ calculation.

Figure 13: Net income distribution in March 2017 for the area of business activity "Retail Trade, except for motor vehicles and motorcycles" (totalling 1,720 different TINs) and the income payer (TIN) with the value of indicator ρ1 =99.86



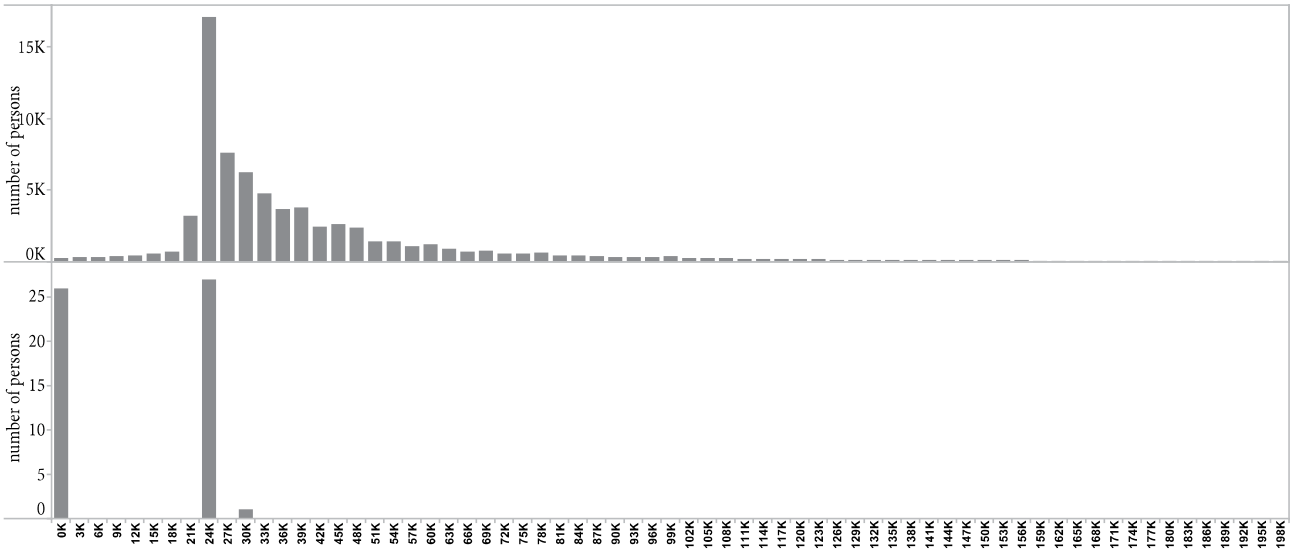
Source: Authors´ calculation.

(as illustrated in the upper part of the Figure referring to the whole line of business), there is a high possibility that some part of the employees' income is paid in cash and not into a bank account. Similarly, within a legal entity in Management activities; Management Consultancy, depicted in Figure 12, with a very high net income dispersion at higher intervals as well, a very high value of risk indicator ρ1 is connected to the fact that all 20 employees in this company were registered on the amount of net earnings ranging from 6,000 to 9,000 dinars

Concluding remarks

Like the presented risk indicator, other risk indicators are going to be developed and used in future. The new ones will be developed according to a similar methodology relying on some other or additional information as risk factors. For example, a deviation in terms of the expected ratio of hiring employees according to different types of contracts (employment contracts, service contracts, author contracts, temporary contracts), a deviation from the

Figure 14: Net income distribution in March 2017 for the area of business activity "Wholesale Trade, except for motor vehicles and motorcycles" (totalling 2,096 different TINs) and the income payer (TIN) with the value of indicator ρ1 =98.79



Source: Authors´ calculation.

expected ratio of received income from interest and from salary by individual person, a deviation from the expected seasonality in earnings and the number of employees in a business entity, a deviation from the expected ratio of paid income to different production factors by business entity - labour and capital share in paid income. Different risk indicators may be combined in some cases.

Further development of risk indicators presented in this paper relies on the development of artificial intelligence methods based on the methods of machine learning. The machine learning will be employed to find good predictors for the delays in the payment of obligations, the probability of irregularities based on historical tax control findings, as well as predictors for the amount of individual net income based on certain attributes (age, gender, line of business, headquarters, tenunre, previous salaries and other sources of income) while relying on some of the models in literature (e.g. Mincer equation [11]). The development of these predictors, in particular the one for probabilities of irregularities, is a very difficult challenge due to incomplete information, i.e., tax evasion may have occurred in companies in which there was no field control. In the case of individual income predictor, it might be possible to use the model on the income payers which are, according to the indicator presented in this paper, considered to be less risky, and "test" it on the rest of taxpayers in order to predict salaries and other income.

Unknown features, as well as that the fact that the data changes over time pose a great problem, as well as the fact that it is necessary to define a model that minimizes the objective function that changes to a certain extent over time, i.e. the objective function is not fixed in time. The first two problems can be considered as the problems of binary classification – with a zero-one result (no delays/there are delays and no violations/there are some violations). Such classification cannot be quite precise, so k-classification will be considered, which can provide better estimates regarding tax evasion. Considering the size of the database, unknown features and incomplete information, this is a very challenging task – both mathematically and economically, and it is necessary to apply customized methods of numerical optimization.

The use of the indicators developed this way will lead to more efficient detection of tax evasion, intended or as a result of reporting mistakes, by better management of tax control activities whose capacities by nature are limited to the control of only certain, small part of taxpayers in a specific period of time.

Apart from direct effects on more effective detection of tax evasion, it is possible that this approach will also produce additional – indirect effects on the increase in tax

collection due to greater self-reporting of tax obligations by taxpayers, aware that "deviations" in their behaviour will be taken into account when prioritizing control activities by the Tax administration.

The paper presents the application of big data analytics on a single database relating to individual income tax returns. An additional space for identifying the risk of tax evasion by using big data analytics is cross-referring to the data in different tax registers. For example, very good risk indicators can be obtained by cross-referring data on value added tax returns with the data on income tax returns of the same legal entity.

Another benefit of the development of risk indicators for individual taxpayers is the possibility to use these indicators to rank all taxpayers according to the level of risk, i.e., the degree of compliance with the regulations and the fulfilment of tax obligations. This ranking enables a positive approach by implicitly "privileging" regular tax payers. The last is not possible if tax administration relying exclusively on the factual verification of the tax compliance by tax control.

Finally, applying big data methods in the analysis in tax administration also provides valuable insights that could be translated into recommendations for tax policy improvements. For example, the evidence can serve to support reducing the space for tax arbitration or improving the regulated procedures related to tax declaration in order to eliminate some frequently perceived risks.

References

1. Schneider F. (2005). Shadow economies around the world: What do we really know?. *European Journal of Political Economy, 21*(3), 598-642.
2. Krstić G., & Radulović, B. (2018). *Siva ekonomija u Srbiji 2017.* Beograd: NALED.
3. Allingham M., & Sandmo, A. (1972). Income tax evasion: a theoretical analysis. *Journal of Political Economics*, 323-338.
4. Alm et al. (1992). Why people pay taxes?. *Journal of Public Economics*, *48*, 21-38.
5. Alm, J.R. (2012). Measuring , Explaining , and Controlling Tax Evasion : Lessons from Theory , Experiments , and Field Studies.
6. Ministarstvo finansija, Poreska uprava Republike Srbije. (2015). *Program transformacije poreske uprave u periodu 2015-2020.*
7. McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition and productivity.*
8. Bjerke K. (1970). Income and wage distributions. *The Review of Income and Wealth, 16*(3), 211-278.
9. Neal D., & Rosen, S. (1998). Theories of the distribution of labor earnings, *NBER Working Papers 6378*, National Bureau of Economic Research, Inc.
10. Stiglitz J. (1969). Distribution of income and wealth among individuals. *Econometrica, 37*(3), 382-397.
11. Mincer, J. (1970). The distribution of labor incomes: A survey with special reference to the human capital approach. *Journal of Economic Literature, 8*(1), 126.

Jasna Atanasijević

(1979) is Assistant Professor at the Department of Mathematics and Informatics at the Faculty of Sciences, University of Novi Sad, where she teaches Finance and Mathematical Introduction to Economics. She is coordinator of Erasmus plus KA2 project aiming to develop interdisciplinary short-cycle programs in policy-making and policy analysis at three largest universities in Serbia in partnership with renowned EU partner universities, few think tanks and relevant Government institutions. She is currently running a research project in the area of big data analytics and ML, and working as a consultant in the field of competitiveness and SME financing. She has a PhD in Applied Economics from the Paris 1 Panthéon-Sorbonne University, master degree in Finance from the Toulouse 1 University, and bachelor degree from the Faculty of Economics, University of Belgrade. From 2014 to 2018 she served as Director of the Public Policy Secretariat of the Republic of Serbia. She was in charge of establishing a new institution in the centre of government, promoting new practices of planning in public administration including result-based management, evidence-based policy-making and regulatory impact assessment. From 2009 to 2014, she was the Chief Economist in Hypo Alpe-Adria Bank in Serbia. She also worked as a researcher in a think tank, and as a consultant in the field of finance and economic policy. She authored several studies and reports in applied economics and finance, some of them published in journals, such as Comparative Economic Studies journal and Eastern European Economics journal. She is a member of the Presidency of the Serbian Association of Economists (SAE), editorial board of the SAE Journal of Business Economics and Management, Advisory Board of the Foundation for the Advancement of Economics, Advisory Board of the Centre for Applied Statistics at the University of Novi Sad and the Serbian Chapter of the Club of Rome.

Nataša Krejić

University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics, specialized in numerical optimization. She is Vice President of the European Consortium for Mathematics in Industry, ECMI, member of National Board for Mathematics, Mechanics and Computer Science, Chair of Numerical Mathematics Group and Vice Dean for Finance at the Faculty of Sciences. She supervised 7 PhD theses and over 20 MSc theses. She leads the project Numerical Methods, Simulations and Applications, coordinates H2020 MCS EID BIGMATH and IPA CRO-SRB project RealForAll, projects on behalf of the Faculty of Sciences, and serves as Associate Editor at the Journal of Computational and Applied Mathematics, Springer. Prof. Krejić received the 2015 Charles Broyden Prize.

Dušan Jakovetić

obtained a bachelor degree in engineering from the School of Electrical Engineering, University of Belgrade (August 2007), and PhD degree in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, PA, and Instituto de Sistemas e Robótica (ISR), Instituto Superior Tecnico (IST), Lisbon, Portugal (May 2013). He is currently Assistant Professor at the Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Serbia. From October 2013 to September 2015, he was a research fellow at the BioSense Institute, Novi Sad, Serbia, and a postdoctoral researcher at IST from June to September 2013. His research interests include distributed inference and distributed optimization.

Nataša Krklec Jerinkić

in September 2007 obtained a bachelor degree in mathematics from the Faculty of Sciences, University of Novi Sad, and a PhD degree in Numerical Mathematics in January 2014. Since July 2014 she has been Assistant Professor at the Department of Mathematics and Informatics at the same Faculty. Her scientific field is numerical optimization and her research interests include stochastic and distributed optimization. She has participated in several national and international projects including two H2020 projects.

Dragana Marković

graduated from the Faculty of Economics, University of Belgrade. She is a professional who has acquired a wide range of knowledge in various fields and has a long-time experience at leading positions both in private and state sectors. On 8 June 2015 the Government of the Republic of Serbia appointed Dragana Marković the Director of the Tax Administration for a five-year mandate. She was appointed this position after being the Deputy Head of Logistics of the Security Information Agency. Before that, she was a State Secretary in charge of finances in the Ministry of Natural Resources, Mining and Spatial Planning. She worked in the Bankruptcy Supervision Agency in the capacity of General Supervisor to Bankruptcy Administrators and has an official license for Bankruptcy Supervisors. During the course of her career, as an expert in financial and commercial transactions in major enterprises, she worked in national and foreign companies where she was engaged in managing domestic and foreign investments. She was the Financial Director in the Joint Stock Tourist Company "Putnik", and a Manager in an Austrian company which managed import and export of liquefied petroleum gas. Previously, she was the head of a department in "NIS Jugopetrol", where she worked from 1992 until 2005. She also gained the experience as the Head of Finances and Accounting in a domestic enterprise which imported and distributed food products. Her career started as a teacher at the High School of Economics "Nada Dimić" in Zemun. She passed the professional examination for members of security sector, examination for managers of financial management and control, as well as for administering the system of public procurement in budget organization. She is married and has a daughter and a granddaughter.