STATISTIKA

STATISTICS
AND ECONOMY
JOURNAL

VOL. **102** (2) 2022

CZECH
STATISTICAL
OFFICE

EDITOR-IN-CHIEF

EDITORIAL BOARD

EXECUTIVE BOARD

MANAGING EDITOR

CONTENTS

ANALYSES

INFORMATION

About Statistika

The journal of Statistika has been published by the Czech Statistical Office since 1964. Its aim is to create a platform enabling national statistical and research institutions to present the progress and results of complex analyses in the economic, environmental, and social spheres. Its mission is to promote the official statistics as a tool supporting the decision making at the level of international organizations, central and local authorities, as well as businesses. We contribute to the world debate and efforts in strengthening the bridge between theory and practice of the official statistics. Statistika is professional double-blind peer reviewed open access journal included in the citation database of peer-reviewed literature **Scopus** (since 2015), in the **Web of Science** *Emerging Sources Citation Index* (since 2016), and also in other international databases of scientific journals. Since 2011, Statistika has been published quarterly in English only.

Publisher

The Czech Statistical Office is an official national statistical institution of the Czech Republic. The Office's main goal, as the coordinator of the State Statistical Service, consists in the acquisition of data and the subsequent production of statistical information on social, economic, demographic, and environmental development of the state. Based on the data acquired, the Czech Statistical Office produces a reliable and consistent image of the current society and its developments satisfying various needs of potential users.

Contact us

Journal of Statistika | Czech Statistical Office | Na padesátém 81 | 100 82 Prague 10 | Czech Republic
E-mail: statistika.journal@czso.cz | Web: www.czso.cz/statistika_journal

The Global Pension Index of Slovakia

Jolana Gubalová[1] | *Matej Bel University in Banská Bystrica, Banská Bystrica, Slovakia*
Petra Medveďová[2] | *Matej Bel University in Banská Bystrica, Banská Bystrica, Slovakia*
Jana Špirková[3] | *Matej Bel University in Banská Bystrica, Banská Bystrica, Slovakia*

Received 3.11.2021 (revision received 24.2.2022), **Accepted** (reviewed) 24.2.2022, **Published** 17.6.2022

Abstract

In every corner of quality of the world, the issue of pension system is being addressed. One of the most important documents that has offered its evaluation is the Mercer consulting firm and the CFA Institute, in cooperation with the Monash Center for Financial Studies. Since Slovakia is not included among the countries that evaluate these companies in their study, this paper offers the calculation of the Global Pension Index for Slovakia in the year 2020. Based on the data obtained and the grade from A to E, Slovakia is one of the countries that are rated by C+ with a total score of 65 points out of 100 as a country with a pension system "that has some good features, but also includes major risks and/or shortcomings that should be addressed. Without these improvements, its efficacy and/or long-term sustainability can be questioned." The problems that affect the pension index of Slovakia are very low pensions for low-income groups, the level of pension assets as a percentage of GDP at the level of 14.35%, the participation in the labour rate at the level of 4.5% for the age 65 and over, and low real economic growth.

Keywords	***DOI***	***JEL code***
Global pension index, adequacy, sustainability, integrity, pensions, Slovakia	*https://doi.org/10.54694/stat.2021.38*	*H55, J21, J26, Q56*

INTRODUCTION

The pension system in the Slovak Republic is based on three pillars. The first one is a pay-as-you-go pillar, which is defined benefit and is regulated by Act 461/2003 Coll. on social insurance and is administered by the Social Insurance Agency. In the first pillar, policyholders only pay contributions to the Social Insurance Agency, and at old age, the Social Insurance Agency will provide them with income according to the number of years worked, income during working life, and current pension value, which is determined on the basis of the growth of the average wage. The second pillar is partially voluntary and is a capitalization scheme representing appreciation in the funds of pension management companies under Act 43/2004 Coll. on old-age pension savings. The third pillar is entirely established on a voluntary

[1] Faculty of Economics, Matej Bel University in Banská Bystrica, Tajovského 10, 975 90 Banská Bystrica, Slovakia.
[2] Faculty of Economics, Matej Bel University in Banská Bystrica, Tajovského 10, 975 90 Banská Bystrica, Slovakia.
[3] Faculty of Economics, Matej Bel University in Banská Bystrica, Tajovského 10, 975 90 Banská Bystrica, Slovakia. Corresponding author: e-mail: jana.spirkova@umb.sk, phone: (+421)908901464.

basis and the conditions of its operation are regulated by Act no. 650/2004 Coll. on supplementary pension savings. Its main advantage is the acquisition of a supplementary pension for the employee and the tax advantage of the employer, who pays a certain amount of contributions for the employee.

The ever-increasing life expectancy, but also the declining birth rate, can significantly affect the stability of pension systems. Changes in capital markets make us to think about investment strategies in both the public sector and private pension funds. Therefore, it is essential to follow up strategies for the development of pension systems and their quality worldwide. The Organisation for Economic Cooperation and Development (OECD) offers every year a detailed overview of the development of pension funds in the Global Pension Statistics, (Global Pension Statistics, 2021). The project was launched in 2002 by the OECD Working Party on Private Pensions and its Task Force on Pension Statistics. The project provides a valuable means to measure and monitor the pension industry. It allows intercountry comparisons of current statistics and indicators on key aspects of retirement systems across OECD and non-OECD countries.

The Global Pension Index 2020 document of the Mercer company (Mercer, 2020) undoubtedly plays a very important role in this area. In the document (formerly known as the Melbourne Mercer Global Pension Index) renowned experts have been evaluating adequacy, sustainability, and integrity of retirement incomes. Since its inception in 2009, it has expanded to cover 39 systems, which represents almost $^{2}/_{3}$ of the world's population, using more than 50 indicators, divided into three sub-indices – adequacy, sustainability, and integrity. However, Slovakia is not included among these countries. Therefore, we offer the determination of the global pension index for Slovakia in compliance with all requirements, which are set out in Mercer (2020).

According to the document the overall index value for each system represents the weighted average of the three sub-indices as follows: 40% for the adequacy sub-index, 35% for the sustainability sub-index, and 25% for the integrity sub-index. The weights determined in this way have been used since 2009, when the index was introduced.

The Adequacy sub-index is the most important way to compare different pension systems. The basic concept is the net replacement rate. According to Mercer (2020), this sub-index takes into account the basic level of income provided by each scheme as well as the net replacement rate at income levels ranging from 50% to 150% of the average wage. In 2020, the net replacement rate in Slovakia was 69%, but the forecast for the future is that it will gradually decrease (OECD, 2020c).

The Sustainability sub-index involves the old-age dependency ratio, pension age, real economic growth over the long-term, level of government debt and public pension expenditure, saving rates, and investment returns.

The Integrity sub-index considers the integrity of the overall pension system but with a focus on funded schemes, which are normally found in the private sector system. This sub-index includes the quality of pension plans and meaningful amount of costs that are associated with determining the amount of pensions and with their payment in the long term.

The Mercer company appears to have no competition in setting the Global Pension Index. Determining this index has been their domain for almost 13 years. However, investigation of pension savings and insurance is presented in several papers. Berstein and Morales (2021) investigate the role of a longevity insurance for defined-contribution pension systems. Hinrichs (2021) offers an overview of pension reforms in Europe. Jakubík et al. (2009) analyse the sensitivity for a dynamic stochastic accumulation model for optimal pension savings management. Špirková et al. (2019) provide a detailed analysis of a payout phase product of the old-age pension savings scheme in Slovakia. Kaščáková et al. (2015) analyze the social and economic situation of the older generation in Slovakia.

The paper is organized as follows. Section 1 – *Preliminaries* provides basic information on the calculation of the global pension index 2020. Individual tables for better orientation in the text show the maximum

score values and the corresponding weights of individual indicators set out in the original document. Section 2 – *Determination of score values of individual sub-indices* offers the calculation of the score of individual indicators and subindices for Slovakia for year 2020. Subsection 2.1 gives score value of the *adequacy sub-index*, subsection 2.2 score value of the *sustainability sub-index*, and subsection 2.3 score value of the *integrity sub-index*. Section 3 – *Global pension index of Slovakia* offers a complete summary of results, score values of individual sub-indices and overall global pension index for Slovakia for 2020. Final section summarizes the obtained results.

1 PRELIMINARIES

The individual indicators that enter into the calculation of the Global Pension Index are marked A1–A11 for the adequacy sub-index, S1–S9 for the sustainability sub-index and R1–R5, P1–P7 and Costs for the integrity sub-index. In addition, some indicators sometimes have two or more questions, and their answers are in some cases assigned different weights. These questions are marked with the letters a, b, or a, b, c, d. The indicators are rated on a scale from 0 to 10, some from 0 to 1, 0 to 2, 0 to 5, or 0 to 18, respectively. Those indicators that are rated on a scale from 0 to 10 are listed in Tables 1–3 as 10. Those that are rated on a scale from 0 to 1, 0 to 2, 0 to 5 or 0 to 18, are finally converted to a scale of 0 to 10 and are written in the tables in the form 10(1), 10(2), 10(2+1), 10(5) and 10(18), respectively. Based on (Mercer, 2020), we present clear tables with individual weights of indicators in a standardized form, i.e.,. Mercer (2020) states the weights in the sum of 100%. All relevant scales are situated in Tables 1–3, where there is a total score of 150 for the adequacy sub-index, 150 for the sustainability sub-index, and 260 for the integrity sub-index.

In the rows marked with the letter x, we present the weighted scores of individual indicators on a 10-point scale.

Table 1 Indicators of the adequacy sub-index

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
a	10	10	10	10(2)	10(2+1)	10	10(2)	10(2)	10	10	10(2)
b	10(2)		10	10(2)							10(2)
Normalized weights											
a	0.15	0.25	0.05	0.02	0.1	0.1	0.075	0.03	0.05	0.05	0.01
b	0.025		0.05	0.03							0.01
x	1.75	2.5	1.0	0.5	1.0	1.0	0.75	0.3	0.5	0.5	0.2
Total score = 150											
Sum of normalized weights = 1								(Mercer, 2020) Sum of weights = 100%			
Sub-index value = $10 \times \sum x = 100$								(Mercer, 2020) Sub-index value = 100			

Source: Author's work according to Mercer (2020)

Table 2 Indicators of the sustainability sub-index

	S1	S2	S3	S4	S5	S6	S7	S8	S9
a	10	10	10	10	10	10	10(2)	10	10(2)
b			10		10	10	10(2)		
c			10						
d			10						
Normalized weights									
a	0.2	0.15	0.05	0.1	0.08	0.05	0.04	0.09	0.01
b			0.05		0.02	0.05	0.01		
c			0.05						
d			0.05						
x	2.0	1.5	2.0	1.0	1.0	1.0	0.5	0.9	0.1

Total score = 150

Sum of normalized weights = 1	(Mercer, 2020) Sum of weights = 100%
Sub-index value = $10 \times \sum x = 100$	(Mercer, 2020) Sub-index value = 100

Source: Author's work according to Mercer (2020)

Table 3 Indicators of the integrity sub-index

	R1	R2	R3	R4	R5	P1	P2	P3	P4	P5	P6	P7	Costs
a	10(2)	10(2)	10(2)	10(2)	10(15)	10	10(2)	10(2)	10(2)	10(2)	10(2)	10(2)	10
b	10(2)	10(2)	10(2)	10(2)				10(2)		10(2)	10(2)		10(5)
c		10(5)	10(2)										
d			10(2)										
Normalized weights													
a	0.025	0.04	0.04	0.025	0.15	0.1	0.05	0.025	0.05	0.025	0.05	0.025	0.05
b	0.05	0.02	0.04	0.025				0.025		0.025	0.025		0.05
c		0.04	0.025										
d			0.02										
x	0.75	1.0	1.25	0.5	1.5	1.0	0.5	0.5	0.5	0.5	0.75	0.25	1.0

Total score = 240

Sum of normalized weights = 1	(Mercer, 2020) Sum of weights = 100%
Sub-index value = $10 \times \sum x = 100$	(Mercer, 2020) Sub-index value = 100

Source: Author's work according to Mercer (2020)

Section 2 provides answers to the individual questions that were set out in the original document, our comments on the individual indicators and their score regarding the conditions of Slovakia for 2020.

2 DETERMINATION OF SCORE VALUES OF INDIVIDUAL SUB-INDICES

In this section, we list all the questions and indicators. Moreover, we answer individual questions and assign them score. Finally, they are summarized in Section 3, Tables 6–8.

2.1 The adequacy sub-index

"The adequacy sub-index considers the benefits provided to the poor and a range of income earners as well as several design features and characteristics which enhance the efficacy of the overall retirement income system. The net household saving rate, the level of household debt and the home ownership rate are also included representing non-pension savings and, as such, important indicators of financial security during retirement" (Mercer, 2020: 47).

Question A1

A1a What is the minimum pension, as a percentage of the average wage, that a single-aged person will receive?

A1b How is the minimum pension increased or adjusted over time? Are these increases or adjustments made on a regular basis?

Publication (Mercer, 2020: 48) states that "an important objective of any retirement income system is to provide a minimum pension to the aged poor. In terms of the World Bank's recommended multi-pillar system, it represents the non-contributory basic pension or Pillar 0, which provides a minimum level of income for all aged citizens. Eligibility for this minimum pension requires no period in the paid workforce, but will often require a minimum period of residency. This question also considers how the minimum pension is increased or adjusted over time".

"For the first part of this question (A1a) "a minimum pension below 30% will score less than the maximum value of 10, with a zero score if the pension is 10% or less of the average earnings, since such a pension offers very limited income provision" (Mercer, 2020: 48).

If a person in the SR is not insured at all in the Social Insurance Agency (or in any other EU country or other countries with which the Slovak Republic has concluded relevant social security contracts) during his active working life until he reaches retirement age, he will not be entitled to any pension from the Social Insurance Agency. In such rare cases, the citizen may profit from another state social benefit, which is not within the competence of the Social Insurance Agency and therefore cannot be considered a pension. A senior who has not been entitled to the payment of a retirement pension and has no other income is entitled to a benefit in material need according to § 10 par. 2 letter a) of the Act on Assistance in Material Needs. The highest amount of assistance in material needs for an individual if the conditions stipulated by law are met is 198.60 € per month. Based on this fact, we can give 0 points to the sub-question A1a.

The valorisation of pensions in Slovakia takes place regularly in accordance with § 82 of the Increment Pension Benefits Act 461/2003 Coll. on Social Insurance (Social Insurance Act., 2021a), therefore, we give the sub-question A1b 1.5 points for increases granted on a regular basis related to price inflation.

Question A2

What is the net pension replacement rate for a range of income earners?

"The OECD (2012) calculates on page 161 net pension replacement rates for a single person at a range of income levels (revalued with earnings growth) throughout his/her working career. These calculations assume no promotion of the individual throughout his/her career, in other words, the individual earns a particular percentage of average earnings throughout. To recognise that a range of income levels exist

in practice, we have used the net replacement rates at three income levels: namely 50%, 100% and 150% of average earnings. Net replacement rates at these three levels are given weightings of 30%, 60% and 10%, respectively, which recognises that there are more individuals who earn less than the average wage than above it" (Mercer, 2020: 49).

On the basis of the OECD (2019: 157) the data for Slovakia are published in Table 4.

Table 4 Net pension replacement rates by earnings (in %)

	Half of the average wage	The average wage	1.5 times the average wage	Weighted average
SR 2019	71.7	65.1	63.3	66.9

Source: Author's work according to OECD (2019)

"The maximum score for this indicator is obtained for any country with a result between 70% and 100%. Any score outside this range scores less than the maximum with a zero score being obtained for a result of less than 20%" (Mercer, 2020: 50).

The net pension replacement rate by earnings for Slovakia is 66.9 % and its score corresponds to 9.38 points.

Question A3

A3a What is the net household saving rate in the country?

A3b What is the level of household debt in the country, expressed as a percentage of GDP?

"The living standards of the elderly will depend on the benefits arising from the total pension system, as well as the level of household savings outside the pension system" (Mercer, 2020: 51).

Based on data from the OECD (2020b), we found that the household savings rate in Slovakia was 5.16%.

"A maximum score is obtained for any country with a saving rate of 20% or higher, and a zero score for any country with a saving rate of less than minus 5%" (Mercer, 2020: 51). For 5.16% we can give Slovakia 4.06 points.

The level of household debt represents the financial liabilities that households must pay back in the future. According to OECD (2020a) the level of household debt in Slovakia represented almost 49.49% of GDP in 2020. "A maximum score is obtained for any country with zero household debt, and a zero score for any country with household debt of 130 percent of GDP or higher" (Mercer, 2020: 51). This means that we award to Slovakia 6.19 points.

Question A4

A4a Are voluntary member contributions made by a median- income earner to a funded pension plan treated by the tax system more favourably than similar savings in a bank account?

A4b Is the investment income earned by pension plans exempt from tax in the pre-retirement and/or post-retirement periods?

The amount of total retirement benefits that a retired individual receives depends on both compulsory contributions and voluntary contributions, the amount of which is often significantly affected by the existence of a taxation incentives. Investment returns are considered a critical aspect of the adequacy sub-index, as they often have a more significant impact on the final amount of benefits than the contributions themselves.

In Slovakia, it is possible to voluntarily save for retirement through pension companies within the second pillar (5 companies) and third pillar (4 companies). There is a potential tax relief, which depends on the amount of the contribution. Participation in the third pillar allows participants to reduce their tax base by 180 €. The management fee is slightly lower here, i.e., 0.3% for the second and 0.9%

for the third pillar (MLSAaF SR, 2021a). For these reasons, we give the sub-question A4a 2 points. All contributions paid are included in the payment of pensions and the difference, i.e., yield, is taxed at a 19% interest rate of tax according to § 43 par. 3 letter e) of Act 595/2003 Coll. Income Tax (2021). Although there are benefits to savings in pension companies, in particular, a tax concession, the yield is not exempt from tax. Therefore, the score for the second sub-question A4b is 0 points.

Question A5

A5a Is there a minimum access age to receive benefits from private pension plans (except for death, invalidity, and/or cases of significant financial hardship)?

A5b If so, what is the current age?

The free availability of invested funds in the stages before retirement age reduces the effectiveness of these funds, as it leads to a reduction in their assets and can thus adversely affect the amount of pension income paid by pension companies.

Two points are awarded for answering the first question in the affirmative. If earlier availability is only allowed in certain situations, 1 point is awarded. In other cases, the number of points is zero.

The second part of the question then concerns only those countries that received two points from the previous question. Zero to one point is added to countries where the age limit is between 55 and 60 years. The maximum score, 1 point, is reached if the minimum age for participants in the availability of funds is 60 years (Mercer, 2020: 53).

We obtained the required information from the website of the Social Insurance Agency, (2021a). Under the second pillar, the old-age pension can be paid to the saver at the earliest from the first day of the calendar month in which the saver has reached retirement age. Early selection is not possible, so we awarded 2 points for question A5a. For a full pay-out, the participant must be 62 years old, so we awarded 1 point for question A5b. Finally, we award these questions a total of 10 points.

Question A6

A6a What proportion, if any, of the retirement benefit from the private pension arrangements is required to be taken as an income stream?

A6b Are there any tax incentives that exist, or favourable conversion rates, to encourage the taking up of income streams?

"The primary objective of a private pension system should be to provide income during retirement. This indicator focuses on whether there are any requirements in the system for at least part of the benefit to be taken as an income stream, or if there are any tax incentives to encourage the take-up of income streams. For the first question, a maximum score is achieved where between 60% and 80% of the benefit is required to be converted into an income stream. A percentage above 80% reduces the flexibility that many retirees need, whilst an answer below 60% is not converting a sufficient proportion of the benefit into an income stream. A percentage below 30% results in a score of zero. For the second question, where there is no requirement for an income stream, half the maximum score could be achieved where significant tax incentives exist to encourage income streams" (Mercer, 2020: 53).

The total amount of contributions for old-age insurance in Slovakia, i.e., until the first and the second pillar, is 18% of the assessment base. Of which in 2020, the contribution to the first pillar represented 13% and to the second pillar 5% (MLSAaF SR, 2021d).

The forms of pension payment from the second pillar are: life annuity, temporary pension or programmed withdrawal. Programmed withdrawal from the 2nd pillar can also be transferred by a one-off withdrawal of the total amount saved. The only condition for the programmed withdrawal is that the sum of the amounts of pension benefits paid to the saver will be higher than the reference amount. In 2020, the reference amount was set at 464.60 € (The Social Insurance Agency, 2021d). So, according to Mercer (2020: 53) the sub-question A6a achieved a rating of 0 points.

"If it is not necessary for some part of the benefit to be covered in the form of an annuity, but are offered, for example, incentive tax relief for annuity income, these countries can receive up to 5 points" (Mercer, 2020: 53). Based on data published by MLSAaF SR (2021b) and the Social Insurance Agency (2021a) we found that tax incentives or favourable conversion rates for income streams, are not available in Slovakia. Therefore, we awarded the A6b sub-question 0 points.

Question A7

A7a On resignation from employment, are plan members normally entitled to the full vesting of their accrued benefit?

A7b After resignation, is the value of the member's accrued benefit normally maintained in real terms (either by inflation- linked indexation or through market investment returns)?

A7c Can a member's benefit entitlements normally be transferred to another private pension plan on the member's resignation from an employer?

"Each question was evaluated with a score of 2 for "yes", 0 for "no" and between 0.5 and 1.5 if it was applied in some cases. The actual score depended on the actual circumstances" (Mercer, 2020: 54).

The authors of the index assumed the existence of Occupational funds in the pension systems of the surveyed countries. Although there are no employee funds in Slovakia, and the given indicator does not fit into the Slovak environment, the questions are set in a sufficiently general way that they can be applied to all pension systems.

We found the answers to the research (MLSAaF SR, 2021d) dedicated to the third pension pillar.

Within the framework of old-age pension savings, it is not possible to cancel the contract and withdraw money early. Therefore, we give the A7a sub-indicator a rating of 2 points.

After resignation, the value of the member's accrued benefit is normally maintained in real terms. Based on this, sub-indicator A7b received a rating of 2 points.

Since 2013, when the amendment to the Supplementary Pension Savings Act was approved, the participant is allowed a free transfer between individual pension companies after one year from the conclusion of the contract, therefore the A7c sub-indicator received a rating of 2 points.

Question A8

Upon a couple's divorce or separation, are the individuals' accrued pension assets normally taken into account in the overall division of assets?

"The adequacy of an individual's retirement income can be disrupted by a divorce or separation. It is desirable that, upon a divorce or separation, the pension benefits that have accrued during the marriage be considered as part of the overall division of assets. The question was assessed on a three-point scale with a score of 2 for 'yes', 1 if it was applied in some cases and 0 for 'no'" (Mercer, 2020: 55).

According to §150 of the Civil Code 40/1964 Coll. each spouse is entitled to demand that he be reimbursed for what has been spent on the partner's other property. Therefore, if the payments to the pillars were made from common funds, the other partner has the right to settle these investments according to the cited legal resolution. We can state that the cash invested in the funds are the subject of divorce proceedings and we assign a rating of 2 points to the indicator A8.

Question A9

What is the level of home ownership in the country?

"In addition to regular income, home ownership represents an important factor affecting financial security during retirement. A maximum feasible level is considered to be 90%. Hence a home ownership level of 90% or more scores maximum results whilst a level of 20% or less scores zero" (Mercer, 2020: 55).

According to the Eurostat database (2020), the share of Slovak citizens owning dwellings was 92.3%, which represents a score of 10 points.

Question A10
What is the proportion of total pension assets invested in growth assets?

Hinz et al. (2010) state that an international comparison of pension fund returns may not be of sufficient informative value, there is no single and available way of allocating active income that is suitable for all fund's participants.

"A zero percentage in growth assets highlights the benefit of security for members, but without the benefits of diversification and the potential for higher returns. No exposure to growth assets scores 2.5 out of 10. This score increases to the maximum score of 10 as the proportion in growth assets increases to 45% of all assets. If the proportion in growth assets exceeds 65% the score is reduced to reflect the higher level of risk and volatility" (Mercer, 2020: 56).

In 2020, the share of equity funds was 13.7% and index funds 16.4% of total assets managed by pension management companies (MLSAaF SR, 2021c). Their common – 30.1% share represents a rating of 7.02 points on a 10-point scale.

Question A11
A11a Is it a requirement that an individual continues to accrue their retirement benefit in a private pension plan when they receive income support such as a disability pension or paid maternity leave?
A11b Does your system provide any additional contributions or benefits for parents who are caring for young children whilst the parent is not in the paid workforce?

"These questions were assessed on a three-point scale with a score of 2 for 'yes', 1 if contributions are paid in some cases and 0 for 'no'" (Mercer, 2020: 57).

An unemployed person can use voluntary pension insurance in the Social Insurance Agency. The period of voluntary insurance, when the voluntarily insured person pays the premiums properly and on time, is valued as the period of insurance when assessing entitlement to benefits from the Social Insurance Agency (2021d). Under the second pillar, the insured person does not have to continue to contribute to the fund in the event of loss of employment. In the case of supplementary pension savings, the insured has the right to interrupt, without any sanctions, because there is no state contribution. In the case of the first sub-question A11a, the score is equal to one.

During both maternity and parental leave, the state pays pension insurance premiums for the parents (Social Insurance Agency, 2021b). For this reason, the second sub-question A11b is rated 2 points.

2.2 The sustainability sub-index

The sustainability sub-index includes several indicators that should ensure the long-term stability of pension systems. These indicators focus on the economic importance of private pension funds, life expectancy in retirement now and in the future, the workforce of the elderly population, the current state of public expenditure on pensions and government debt, and real economic growth.

Question S1
What proportion of the working-age population are members of private pension plans?

Saving in private pension companies ensures the stability of the pension system, as it relieves state spending. Therefore, it is important that the proportion of the working-age population on private schemes is as high as possible.

The publication (Pensions at a Glance, 2019: 116) states that overall participation in funded pensions by type of plan, 2017 or latest available year as a percentage of the working-age population is approximately 38%. However, in Slovakia, this value has risen sharply in recent years. MLSAaF SR (2021d) states, that the number of savers as of 31 December 2020 was in the second pillar of 1 626 177 savers and in the third pillar 861 344 savers. The number of economically active inhabitants in 2020 was 2 712 700 (Working

age population, 2021). Thus, the proportion of the working age population for the second pillar is 60% and for the third pillar 32%. Tatra Bank (2021) states that about 60 to 68% of Slovaks save or declare that they save for retirement. Since some savers may be in the second, but also in the third pillar, we assume that the proportion of the working-age population is 64%. According to Mercer (2020: 59), based on the conversion from the interval (15–80)%, we allocate to the S1-Coverage Indicator a value of 7.54.

Question S2
What is the level of pension assets, expressed as a percentage of GDP, held in private pension arrangements, public pension reserve funds, protected book reserves, and pension insurance contracts?

The indicator S2 – Level of Assets is important for the stable payment of pensions in the future.

By using data from the Gross Domestic Product (2021) and Private pension assets Slovakia (2020), we obtain the level of pension assets as a percentage of GDP at the level of:

$$\frac{13\,137.18\,\text{mil €}}{91\,555.5\,\text{mil €}} \times 100\ \% = 14.35\ \%.$$

According to Mercer (2020: 60) conversion from interval (0–175)% to a 10-point scale gives to indicator S2 a value of 0.82.

Question S3
S3a What is the current life expectancy at the state pension age?
S3b What is the projected life expectancy at the expected state pension age in 2050?
S3c What is the projected old age dependency ratio in 2050?
S3d What is the estimated Total Fertility Rate (TFR) for 2015–2020?

We gradually answer individual questions to determine the indicator S3 – Life Expectancy at State Pension Age.

Life expectancy, according to (Statistical Office SR, 2021a), for a 63-year-old person (unisex) is 18.4 years and thus, according to Mercer (2020: 61) based on the scale from (28–18) years to 10-point scale, a value of 10.

Based on data from (Eurostat, 2021b) the projected life expectancy for 65-year-old person is 21.55 years. The projected life expectancy for men is 19.7 and for women 23.4 year. In our calculation, we used a simple arithmetic mean of these values for simplicity. Within the same scale, a value of 6.45 is assigned for S3b.

The old age dependency ratio is given by:

$$\text{Old age dependency ratio} = \frac{\text{population age 65+}}{\text{population 15–64}} \times 100\ \%.$$

The Eurostat database indicates the value of old age dependency ratio 51.4% and thus we assign, within the same scale, a score of 4.65 to S3c.

We determined the total fertility rate for Slovakia as the average value for the period 2015–2020. It is at the level of 1.50, so we will assign using the same scale the value 3.34 to S3d.

Question S4
What is the level of mandatory contributions that are set aside for retirement benefits (i.e., funded), expressed as a percentage of wages?

These include mandatory employer and/or employee contributions towards funded public benefits (i.e., social security) and/or private retirement benefits.

The Act 43/2004 Coll. in the old age pension scheme (2021), Article 22, item (e), states: "The rate of mandatory contributions shall be in 2020, 5% of the assessment base". Thus, according to Mercer (2020: 63) by transforming the scale (0–12)% to a 10-point scale, the value 4.17 is assigned to indicator S4 – Funded Mandatory Contributions.

Question S5

S5a What is the labour force participation rate for those aged 55-64?

S5b What is the labour force participation rate for those aged 65 or over?

Based on data from the International Labour Organization (2020) the labour force participation rate for ages 55–64 is 61.3%. According to Mercer (2020: 64) by scaling from (40–80)% to 10 points, we assign for indicator S5a – Labour Force Participation Rate aged 55–64, score 5.33.

For ages 65 and over, the participation in the labour rate is at the level of 4.5% and the assigned score for S5b, when applying the same scale, is 1.5.

Question S6

S6a What is the level of adjusted government debt (being the gross public debt reduced by the size of any sovereign wealth funds that are not set aside for future pension liabilities), expressed as a percentage of GDP?

S6b What is the level of public expenditures on pensions expressed as a percentage of GDP, averaged over the latest available figure and the projected figure for 2050?

According to Eurostat (2021b) the level of adjusted government debt as a percentage of GDP is on the level of 60.7%. Mercer (2020: 65) states: "A maximum score was achieved for countries with a zero or negative level of adjusted government debt, with a zero score for countries with an adjusted government debt of 150% of GDP or higher." Based on the scale of GDP to 10-point scale, we give a value 5.98 for indicator S6a – Adjusted Government Debt.

The International Labour Organization (2020a) states that the level of public expenditures on pensions is for year 2020 on the level of 8.3%, and for 2050 on the level of 8.8% of GDP. Hence, the average is 8.56%. Based on Mercer (2020: 65): "A maximum score was achieved for systems with public pension costs of 2% of GDP or less, with a zero score for systems with costs of 16% of GDP or higher." Therefore, by scaling from (16–2)% of GDP to a 10-point scale, score 5.32 is assigned to indicator S6b – Public cost of pensions.

Question S7

S7a In respect of private pension arrangements, are older employees able to access part of their retirement savings or pension and continue working (e.g., part time)?

S7b If yes, can employees continue to contribute and accrue benefits at an appropriate rate?

Older employees can receive a pension and can also work in addition to receiving a pension, so we assign a score of 2 to sub-indicator S7a, and thus assign a value of 10.

A pensioner can work and also receive a pension from the first pillar, but he can still have his funds valued in the second and third pillar. This means that we assign a maximum score of 2 to the sub-indicator S7b and we give a value of score 10.

Question S8

What is the real economic growth rate averaged over seven years (namely the last four years and projected for the next three years)?

The real economic growth rate shows the rate of change in a country's GDP, typically from one year to the next. The real GDP growth rate is a more useful measure than the nominal GDP growth rate because it considers the effect of inflation on economic data.

The average of the real economic growth rate averaged over seven years (2017–2023), International Monetary Fund (2021b) is on the level of 2.22%, i.e., and based on the scale from (–1.0–5.0)% to 10-point scale, we set indicator S8 – Real Economic Growth to a value of score 5.37.

Question S9
Is it a requirement for the pension plan's trustees/executives/fiduciaries to consider Environmental, Social and Governance (ESG) issues in developing their investment policies or strategies?

According to Directive (EU) 2016/2341 of the European Parliament and of the Council of 14 December 2016 on the activities and supervision of institutions for occupational retirement provision (2021), member states should require institutions for occupational retirement provision to disclose information on whether environmental, social, and management factors are taken into account in investment decisions and how they form part of their risk management system. The countries of the European Union must transpose these new rules into their national law by 13 January 2019 (European Commission, 2021). Therefore, we assign 2 points from the maximum score to this question.

2.3 The integrity sub-index

"The integrity sub-index considers three broad areas of the pension system: regulation and governance, protection and communication for members and operating costs" (Mercer, 2020: 69).

2.3.1 Regulation and governance

Question R1
R1a Do private sector pension plans need regulatory approval or supervision to operate?
R1b Is a private pension plan required to be a separate legal entity from the employer?

Act 43/2004 Coll. on old-age pension savings, §47 states that a pension fund management company is a joint-stock company with its registered office in the Slovak Republic, the subject of which is the creation and administration of pension funds for the implementation of old-age pension savings. The National Bank of Slovakia grants the permission to establish and operate a pension fund management company is granted by the National Bank of Slovakia, which subsequently supervises pursuant to §113.

Act 650/2004 Coll. in the Supplementary Pension Scheme, §22 states that a supplementary pension company is a joint stock company with its registered office in the Slovak Republic, the subject of which is the creation and administration of supplementary pension funds for the purpose of supplementary pension savings. The National Bank of Slovakia grants the permission to establish and operate a supplementary pension company, which subsequently supervises pursuant to §69.

Neither pension management companies nor supplementary pension companies can be dependent on the employer. We give a maximum score of 2 points for each question.

Question R2
R2a Are private sector pension plans required to submit a written report in a prescribed format to a regulator each year?
R2b Does the regulator make industry data available from the submitted forms on a regular basis?
R2c How actively does the regulator discharge its supervisory responsibilities? Please rank on a scale of 1 to 5.

In Act 43/2004 Coll. in old-age pension savings, Article 109 states that the pension fund management company is obliged to submit to the National Bank of Slovakia no later than three months after the end of the accounting period an annual report on equity management for the previous calendar year, alsoannual reports on the management of assets in managed pension funds for the previous calendar year. In Act 650/2004 Coll., Section 67 stipulates the same information obligations of a supplementary

pension company towards the National Bank of Slovakia. The National Bank of Slovakia requires regular sending of information from pension companies, their annual, half-yearly, and quarterly activity reports, collects data on the volume of assets and investments, regularly publishes summary statistics and information on the development of individual funds, performs remote and on-site supervision. We give a maximum score of 2 points for the first two questions. We give a score of 4 points to the last question according to supervisory responsibilities scaling system (Mercer, 2020: 71).

Question R3

R3a Where assets exist, are the private pension plan's trustees/executives/fiduciaries required to prepare an investment policy?

R3b Are the private pension plan's trustees/executives/ fiduciaries required to prepare a risk management policy?

R3c Are the private pension plan's trustees/executives/ fiduciaries required to prepare a conflicts of interest policy?

R3d Are the private pension plan's trustees/executives/ fiduciaries required to have:

– one or more independent members included in the governing body?

– equal member and employer representation on the governing body?

The second chapter of the sixth part of the Act 43/2004 Coll. deals with the investment of assets in pension funds and defines, inter alia, in §81 assets in a pension fund, in §82 the rules for limiting and distributing risk for a pension fund, in §85 the strategic location of pension fund investments. The second chapter of the fifth part of the Act deals with the conditions of operation of a pension management company, §53 sets out the rules of prudential business of a pension management company, §55a defines risk management and measurement, §58 conflicts of interest. The answers to all four questions are explained in the same way in the individual paragraphs in Act 650/2004 Coll. We give a maximum score of 2 points for all four questions.

Question R4

R4a Do the private pension plan's trustees/executives/ fiduciaries have to satisfy any personal requirements set by the regulator?

R4b Are the financial accounts of private pension plans (or equivalent) required to be audited annually by a recognised professional?

The information obligation of pension management companies to the National Bank of Slovakia is enshrined in Act 43/2004 Coll., §109. Regular audits are enshrined in §56, which concerns bookkeeping. It is stated that the financial statements of the pension fund management company and the pension fund are stated to be audited by an auditor or an audit company and approved by the general meeting of the pension management company. The pension fund management company is required to notify the name of the auditor in writing to the National Bank of Slovakia, which is entitled to reject the auditor within the specified period. Subsequently, the pension fund management company is obliged to notify the new auditor in writing, and the National Bank of Slovakia is entitled to reject the auditor, and then determine which auditor verifies the financial statements of the pension fund management company and the pension fund. The answers to both questions are explained in Act 650/2004 Coll. in §30 and §67. We give a maximum score of 2 points for each question.

Question R5

R5a What is the government's capacity to effectively formulate and implement sound policies and to promote private sector development?

R5b What respect do citizens and the state have for the institutions that govern economic and social interactions among them?

R5c How free are the country's citizens to express their views? What is the likelihood of political instability or politically motivated violence?

Every year, the World Bank compiles a Global Governance Indicator, which consists of 6 parts: Government Effectiveness, Regulatory Quality, Rule of Low, Control of Corruption, Voice and Accountability, Political Stability and Absence of Violence/Terrorism. Values for individual parts range from –2.5 to +2.5. After adding up all the values, a value of 3 is added to the total indicator to avoid negative values of the indicator.

According to the World Bank (2019), Slovakia received the value of the Worldwide Governance Indicator 3.88 (0.59 for Government Effectiveness, 1.01 for Regulatory Quality, 0.53 for Rule of Low, 0.22 for Control of Corruption, 0.86 for Voice and Accountability, 0.67 for Political Stability and Absence of Violence/Terrorism). We give a score of 6.88 (3.88 + 3) out of the maximum possible value of 15 (Mercer, 2020: 73).

2.3.2 Protection and communication for members

Question P1

For defined benefit schemes:

P1a Are there minimum funding requirements?

P1b What is the period over which any deficit or shortfall is normally funded?

P1c Describe the major features of the funding requirements.

For defined contribution schemes, are the assets required to fully meet the members' accounts?

The pension system in Slovakia uses both defined benefit and defined contribution types of funds. The investment rules are clearly set out in Act 43/2004 Coll., Part Six, starting with Section 60, and Section 80 state that the National bank of Slovakia sets out the definition of own funds that a company is required to comply with. The law does not set the maximum possible time for managing the deficit. However, according to the law, due to the nature of the funds, liabilities must be covered by real resources. Within the defined benefit fund, the maximum score of 5 points was awarded when the funding requirements were confronted with actuarial methods and the deficit could not last longer than 4 years. In our case, we award half points, i.e., 2.5 points.

In Act 650/2004 Coll., the investment rules are stated in §53 and the adequacy of own resources in §33. Strict demands are placed on the management of companies' assets; therefore, we give the full number of points, 5 points within the defined contribution fund, thus 7.5 points together.

Question P2

Are there any limits on the level of in-house assets held by a private sector pension plan? If yes, what are they?

The question is mainly focused on the independence of employee funds, which is not common in Slovakia. However, for both types of pension savings under the mentioned acts, it is stated that investments in other funds may not be transferred. If the country's pension system does not include employee funds but meets the share of the so-called in-house assets in private pension funds, the maximum number of points is awarded. Therefore, we can award the maximum number of points for the question, i.e., 2 points.

Question P3

P3a Are the members' accrued benefits provided with any protection or reimbursement from an act of fraud or mismanagement within the fund?

P3b In the case of employer insolvency (or bankruptcy), do any unpaid employer contributions receive priority over payments to other creditors, and/or are members' accrued benefits protected against claims of creditors?

According to Act 43/2004 Coll., §113, the National Bank of Slovakia has supervision over pension funds in Slovakia, which is obliged to protect the interests of consumers when exercising supervision. Assets in the pension fund are not part of the bankruptcy estate of the pension management company, nor may they be used to settle with the creditors of the pension management company. A new management company is designated to manage the funds.

The second level of supervision is represented by the depositary, i.e., the bank that is not property-related to the given fund. According to Section 104, assets in pension funds which are entrusted to the depositary in accordance with the provisions of this Act may not be the subject of enforcement of a decision or execution against the depositary, subject to enforcement of the depositary, and are not part of the bankruptcy estate. In the event of the bankruptcy of the depositary, the owners will not lose their funds, because the depositary does not own the funds.

For each question we award the maximum number of points, i.e., 2 points for each question.

Question P4

When joining the pension plan, are new members required to receive information about the pension plan?

According to Act 43/2004 Coll., §64, which deals with the old-age pension savings contract, the pension fund management company is obliged to inform the person interested in concluding an old-age pension savings contract before concluding the old-age pension savings contract with key information, sufficiently in advance, the statute of the pension fund, with a report on the management of the assets of each pension fund created and managed by the pension fund management company and with a report on the management of the pension management company.

According to Act 650/2004 Coll., §57 before concluding the participation contract, the supplementary pension company is obliged to inform the potential participant about the status of the supplementary pension fund, with key information, on how to take into account environmental factors including climatic, social, organizational and management factors and on the possibilities of obtaining further information. For the question, we award the maximum number of points, i.e., 2 points.

Question P5

P5a Are plan members required to receive or have access to the annual report from the pension plan?

P5b Is the annual report required to show:

– the allocation of the plan's assets to major asset classes?

– the major investments of the plan?

Act 43/2004 Coll., §105 lists all the requirements that the pension fund management company is obliged to make available to savers in the annual report on the management of its own assets and the annual reports on the management of assets in pension funds. In Act 650/2004 Coll., §65, analogous information is given regarding the obligation to inform savers about the annual reports of the funds, which also contain information on the allocation of assets and investments of the company. For both questions, we give 2 points for each question.

Questions P6

P6a Are plan members required to receive an annual statement of their current personal benefits from the plan?

P6b Is this annual statement to individual members required to show any projection of the member's possible retirement benefits?

According to Act 43/2004 Coll., §108, the Pension fund management company is obliged to send a statement from the saver's personal pension account on the last day of the calendar year.

The statement shall include, inter alia, the saver's retirement age or the estimated retirement age of the saver, the amount corresponding to the current value of the saver's personal pension account at the end of the calendar year, information on the saver's contributions to his personal pension account during the last 12 months, on the amount of appreciation of the saver's personal pension account at the end of the calendar year, information on remuneration, costs and fees due to the saver, information on the retirement age pension forecast, which includes the baseline scenario, the optimistic scenario and the pessimistic scenario based on possible economic scenarios if the saver is not a beneficiary old-age pension or early retirement pension paid by program selection. The same obligations are set out in Act 650/2004 Coll., §66 item a.

The models of personal account statements and statements in supplementary pension savings are set out in Measure of the MLSAaF SR (2021). For both questions we give a maximum 2 points for each question.

Question P7

Do plan members have access to a complaints tribunal which is independent from the pension plan?

The mentioned acts contain specific information on complaints, a citizen can file a complaint within the framework of financial consumer protection with the National Bank of Slovakia, which is the financial market supervisory authority.

According to Act 43/2004 Coll., §114, the subject of supervision is not, however, the resolution of disputes arising from the contractual relations of a pension management company, the hearing and decision-making of which is regulated by the competent court or other body according to a special regulation. If the National Bank of Slovakia finds a violation of consumer rights, it may impose sanctions on a financial institution, see §115. Analogously, Act 650/2004 Coll. deals with supervision in §69 and sanctions in §71.

We give the question a maximum rating of 2 points.

2.3.3 Costs

C1a What percentage of total pension assets is held in various types of pension funds?

C1b What percentage of total pension assets are held by the largest ten pension funds/providers?

The long-run efficiency of a pension system depends on the costs that will affect the pension income of the fund members themselves. Different types of pension funds have different cost structures. As pension funds increase in size, their costs as a proportion of assets will decrease and benefits will be passed on to fund members. The aim is to minimize costs and improve profit.

Pension funds in Slovakia are managed by pension fund management companies under the second pillar and by supplementary pension companies under the third pillar. The net asset value of the guaranteed and non-guaranteed pension funds of all 5 pension fund management companies operating in Slovakia as of 31 December 2020 was 10 336.64 mil. € (Association of Pension Fund Management Companies, 2021). The net asset value of the four supplementary pension companies operating in Slovakia as of 31 December 2020 was 2 672.61 mil. € (Annual reports of individual companies, 2021). In the first question, a value was assigned from 1 (for individual private funds) to 10 (for centralized funds). The weight of these values depended on their share of assets in each type of fund. In Slovakia, we only have individual funds, so we assign a rating of 1 point to the first question.

In 2020, pension assets in Slovakia were held in 17 pension funds of pension fund management companies and in 19 pension funds of supplementary pension companies. 73.87% of all pension assets were held in the 10 largest pension funds (Association of Pension Fund Management Companies, 2021).

Score 1 was awarded when these assets in the 10 largest funds accounted for less than 10% of all assets, rising to a maximum score of 5 when these assets accounted for more than 75% of all assets. In the second question, we award 4.93 points.

3 GLOBAL PENSION INDEX OF SLOVAKIA

This chapter presents all our score values found for individual questions and their associated indicators. Our final results are recorded in Tables 5–8. The sign x* indicates the real weighted score values of individual indicators on a 10-point scale.

Table 5 Indicators of the adequacy sub-index, Slovakia 2020

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
a	0.00	9.38	3.91	10(2)	10(2+1)	0.00	10(2)	10(2)	10	7.02	5(1)
b	7.5 (1.5)		6.19	0(0)							10(2)
Normalized weights											
a	0.15	0.25	0.05	0.02	0.1	0.1	0.075	0.03	0.05	0.05	0.01
b	0.025		0.05	0.03							0.01
x^*	0.188	2.345	0.505	0.20	1.00	0.00	0.75	0.3	0.5	0.351	0.15
Sub-index value = $10 \times \sum x^* = 62.9$											

Source: Author's work

Table 6 Indicators of the sustainability sub-index, Slovakia 2020

	S1	S2	S3	S4	S5	S6	S7	S8	S9
a	7.54	0.82	10	4.17	5.33	5.98	10 (2)	5.37	10 (2)
b			6.45		1.5	5.32	10 (2)		
c			4.65						
d			3.34						
Normalized weights									
a	0.2	0.15	0.05	0.1	0.08	0.05	0.04	0.09	0.01
b			0.05		0.02	0.05	0.01		
c			0.05						
d			0.05						
x^*	1.508	0.123	1.222	0.417	0.456	0.565	0.50	0.483	0.1
Sub-index value = $10 \times \sum x^* = 53.7$									

Source: Author's work

Table 7 Indicators of the integrity sub-index, Slovakia 2020

	R1	R2	R3	R4	R5	P1	P2	P3	P4	P5	P6	P7	Costs
a	10(2)	10(2)	10(2)	10(2)	4.59(6.88)	7.5	10(2)	10(2)	10(2)	10(2)	10(2)	10(2)	1
b	10(2)	10(2)	10(2)	10(2)				10(2)		10(2)	10(2)		9.86(4.93)
c		8(4)	10(2)										
d			10(2)										
Normalized weights													
a	0.025	0.04	0.04	0.025	0.15	0.1	0.05	0.025	0.05	0.025	0.05	0.025	0.05
b	0.05	0.02	0.04	0.025				0.025		0.025	0.025		0.05
c		0.04	0.025										
d			0.02										
x^*	0.75	0.92	1.25	0.5	0.689	0.75	0.5	0.5	0.5	0.5	0.75	0.25	0.543

Sub-index value = $10 \times \sum x^* = 84.0$

Source: Author's work

Table 8 The overview of sub-indices score for the monitored countries of Europe, 2020

	Country	Overal score	Adequacy	Sustainability	Integrity
1.	Netherlands	82.6	81.5	79.3	88.9
2.	Denmark	81.4	79.8	82.6	82.4
3.	Finland	72.9	71.0	60.5	93.5
4.	Sweden	71.2	65.2	72.0	79.8
5.	Norway	71.2	73.4	55.1	90.3
6.	Germany	67.3	78.8	44.1	81.4
7.	Switzerland	67.0	59.5	64.2	83.1
8.	Ireland	65.0	74.7	45.6	76.5
9.	Slovakia	65.0	62.9	53.7	84.0
10.	United Kingdom	64.9	59.2	58.0	83.7
11.	Belgium	63.4	74.6	32.4	88.9
12.	France	60.0	78.7	40.9	57.0
13.	Spain	57.7	71.0	27-May	78.5
14.	Poland	54.7	59.9	40.7	65.9
15.	Austria	52.1	64.4	22-Jan	74.6
16.	Italy	51.9	66.7	18-Aug	74.4
	Average	65.5	70.1	49.8	80.2

Source: Author's work

The total Global Pension Index with weights of 40% for the adequacy sub-index, 35% for the sustainability sub-index and 25% for the integrity sub-index is 65.0 points from the 100-point scale. Based on the grade from A to E, which is listed in Mercer (2020: 6), Slovakia is rated by grade C+ as a country with a pension system "that has some good features, but also has major risks and/or shortcomings that should be addressed. Without these improvements, its efficacy and/or long-term sustainability can be questioned." Slovakia is one of the countries whose overall index value is 65.0 points, and it shares the eighth and ninth place with Ireland in the monitored European countries. Slovakia has the lowest value of the sustainability sub-index, namely 53.7 points.

CONCLUSION

In our paper, we have determined the Global Pension Index 2020 for Slovakia according to the original document. As the overall grade C+ shows, based on the grade from A to E, Slovakia is not among the countries with the best pension system. For example, the second pillar of retirement savings is not as profitable as it should be. This unfavourable situation regarding the appreciation of funds was caused in 2009 and subsequently in 2013 by the political intervention of the transfer of a huge amount of money from equity funds to bond funds, and savers are significantly impoverished by returns. In this way, the savings of more than 800,000 savers were transferred. This unfavourable situation endures. The net replacement rate is currently at 69%, but there is a serious concern that it will rapidly fall well below 50%. Therefore, it is essential that future retirees save in the second pillar and in the third pillar. Of course, also the third pillar currently shows serious shortcomings. These are mainly high administrative fees, which are paid even at a time when the appreciation of savings is very weak, even negative. Another serious problem is that in 2020, only 32% of the economically active population was saving in the third pillar, and about 30% of them did not pay contributions in 2020. As part of improving the pension system in Slovakia, it is necessary for the state to assume greater responsibility in the investment strategy of the funds offered. Our ambition is to offer our results to representatives of Mercer, CFA Institute and Monash Center for Financial Studies and ask them to include Slovakia in the group of monitored countries.

ACKNOWLEDGMENT

The work was supported by the Slovak Scientific Grant Agency VEGA No. 1/0150/21 under the project Profit testing of pension insurance products.

References

Act 43/2004 Coll. on the Old-Age Pension Scheme [online]. (2021). Bratislava, Slovakia: National Council of the Slovak Republic. [cit. 4.7.2021]. <https://www.nbs.sk/_img/Documents/_Legislativa/_UplneZneniaZakonov/A43-2004.pdf>.

Act 650/2004 Coll. on the Supplementary Pension Scheme [online]. (2021). Bratislava, Slovakia: National Council of the Slovak Republic. [cit. 3.7.2021]. <https://www.nbs.sk/_img/Documents/_Legislativa/_BasicActs/A650-2004.pdf>.

Act 595/2003 Coll. on Income Tax [online]. (2021). Bratislava, Slovakia: National Council of the Slovak Republic. [cit. 4.7.2021]. <https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2003/595/20210101.html>.

ASSOCIATION OF PENSION FUND MANAGEMENT COMPANIES. (2020). *Data on pension funds as of 31.12.2020* (in Slovak: Údaje o dôchodkových fondoch k 31.12.2020) [online]. Bratislava, Slovakia: Asociácia dôchodkových správcovských spoločností. [cit. 10.8.2021]. <https://www.adss.sk/preview-file/adss_20_12_31-503.pdf>.

BERSTEIN, S., MORALES, M. (2021). The role of a longevity insurance for defined contribution pension systems. *Insurance: Mathematics and Economics*, 99: 233–240. <https://doi.org/10.1016/j.insmatheco.2021.03.020>.

CFA INSTITUTE. (2021). *CFA Program* [online]. Charlottesville, Virginia: CFA Institute, 22902. [cit. 10.8.2021]. <https://www.cfainstitute.org/programs/cfa>.

Civil Code Law 40/1964 Coll. (in Slovak: Občiansky zákonník 40/1964 Zb.) [online]. (2021). Bratislava, Slovakia: National Council of the Slovak Republic. [cit. 8.7.2021]. <https://www.mindbank.info/item/2196>.

Directive (EU) 2016/2341 of the European Parliament and of the Council of 14 December 2016 on the activities and supervision of institutions for occupational retirement provision (IORPs) [online]. (2021). Brussels, Belgium: Official Journal of the European Union L 354/37. [cit. 14.12.2021]. <https://eurlex.europa.eu/legalcontent/EN/TXT/PDF/uri=CELEX:32016L2341&from=EN>.

EUROPEAN COMMISION. (2021). *Occupational pension funds* [online]. Brussels, Belgium: European Commission. [cit. 11.7.2021]. <https://ec.europa.eu/info/business-economy-euro/banking-and-finance/insurance-and-pensions/occupational-pension-funds_en>.

EUROSTAT. (2020). *Housing in Europe* [online]. Brussels, Belgium: European Commission. [cit. 11.7.2021]. <https://ec.europa.eu/eurostat/cache/digpub/housing/images/pdf/Housing-DigitalPublication-2020_en.pdf?lang=en>.

EUROSTAT. (2021a). *Projected old-age dependency ratio* [online]. Brussels, Belgium: European Commission. [cit. 12.8.2021]. <https://ec.europa.eu/eurostat/databrowser/view/tps00200/default/table?lang=en>.

EUROSTAT. (2021b). *Projected life expectancy by age (in completed years), sex, and type of projection* [online]. Brussels, Belgium: European Commission. [cit. 12.8.2021]. <https://ec.europa.eu/eurostat/databrowser/view/proj_19nalexp/default/table?lang=en>.

Global Pension Statistics [online]. (2021). Paris, France: OECD. [cit. 5.8.2021]. <https://www.oecd.org/finance/private-pensions/globalpensionstatistics.htm>.

Gross Domestic Product [online]. (2021). Bratislava, Slovakia: Štatistický úrad Slovenskej republiky. [cit. 10.7.2021].<http://datacube.statistics.sk/#!/view/sk/VBD_SK_WIN/eu0001re/v_eu0001re_00_00_00_en>.

HINRICHS, K. (2021). Recent pension reforms in Europe: More challenges, new directions. An overview. *Social Policy and Administration*, 55(3): 409–422. <https://doi.org/10.1111/spol.12712>.

HINZ R. et al. (2010). *Evaluating the Financial Performance of Pension Funds*. Washington, D.C.: The World Bank.

INTERNATIONAL LABOUR ORGANIZATION [online]. (2020). Genève, Switzerland: International Labour Organization. [cit. 15.7.2021]. <https://www.ilo.org/shinyapps/bulkexplorer43/?lang=en&segment=indicator&id=EAP_DWAP_SEX_AGE_RT_A>.

INTERNATIONAL MONETARY FUND. (2021a). *Country data* [online]. Washington, D.C.: International Monetary Fund. [cit. 12.8.2021]. < https://www.imf.org/en/Countries/SVK>.

INTERNATIONAL MONETARY FUND. (2021b). *Country data* [online]. Washington, D.C.: International Monetary Fund. [cit. 12.8.2021]. <https://www.imf.org/en/Countries/SVK#countrydata>.

JAKUBÍK, T., MELICHERČÍK, I., ŠEVČOVIČ, D. (2009). Sensitivity analysis for a dynamic stochastic accumulation model for optimal pension savings management. *Ekonomický časopis*, 57(8): 756–771.

KAŠČÁKOVÁ, A., KUBIŠOVÁ, Ľ., NEDELOVÁ, G. (2015). Social and economic situation of silver generation in Slovakia [online]. *18th AMSE Conference – Applications of Mathematics and Statistics in Economics*. [cit. 5.9.2021]. <http://amse-conference.eu/history/amse2015/doc/Kascakova_Nedelova_Kubisova.pdf>.

Measure of the Ministry of Labour, Social Affairs and Family of the Slovak Republic (MLSAaF SR) 411/2019 Coll. [online]. Bratislava, Slovakia: National Council of the Slovak Republic. [cit. 12.8.2021]. <https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2019/411/20200101>.

MERCER. (2020). *Mercer CFA Institute Global Pension Index* [online]. Melbourne, Australia: Mercer and CFA Institute. [cit. 5.4.2021]. <https://www.mercer.com.au/our-thinking/global-pension-index.html>.

MINISTRY OF LABOUR, SOCIAL AFFAIRS AND FAMILY OF THE SLOVAK REPUBLIC (MLSAaF SR). (2021a). *Minimum pension* [online]. Bratislava, Slovakia: National Council of the Slovak Republic. <https://www.employment.gov.sk/sk/socialne-poistenie-dochodkovy-system/minimalny-dochodok.html>.

MINISTRY OF LABOUR, SOCIAL AFFAIRS AND FAMILY OF THE SLOVAK REPUBLIC (MLSAaF SR). (2021b). *The second pillar – old-age pension savings* [online]. Bratislava, Slovakia: National Council of the Slovak Republic. <https://www.employment.gov.sk/sk/socialne-poistenie-dochodkovy-system/dochodkovy-system/ii-pilier-starobne-dochodkove-sporenie/>.

MINISTRY OF LABOUR, SOCIAL AFFAIRS AND FAMILY OF THE SLOVAK REPUBLIC (MLSAaF SR). (2021c). *The second pillar in numbers* [online]. Bratislava, Slovakia: National Council of the Slovak Republic. <https://www.employment.gov.sk/sk/socialne-poistenie-dochodkovy-system/dochodkovy-system/ii-pilier-starobne-dochodkove-sporenie/zhodnotenie-majetku/>.

MINISTRY OF LABOUR, SOCIAL AFFAIRS AND FAMILY OF THE SLOVAK REPUBLIC (MLSAaF SR). (2021d). The third pillar – supplementary pension savings [online]. Bratislava, Slovakia: National Council of the Slovak Republic. <https://www.employment.gov.sk/sk/socialne-poistenie-dochodkovy-system/dochodkovy-system/iii-pilier-doplnkove-dochodkove-sporenie/>.

MINISTRY OF LABOUR, SOCIAL AFFAIRS AND FAMILY OF THE SLOVAK REPUBLIC (MLSAaF SR). (2021e). *Old-age pension savings contributions* [online]. Bratislava, Slovakia: National Council of the Slovak Republic. <https://www.employment.gov.sk/sk/socialne-poistenie-dochodkovy-system/dochodkovy-system/ii-pilier-starobne-dochodkove-sporenie/prispevky-starobne-dochodkove-sporenie-2.html>.

MINISTRY OF LABOUR, SOCIAL AFFAIRS AND FAMILY OF THE SLOVAK REPUBLIC (MLSAaF SR). (2021f). *The third pillar in numbers* [online]. Bratislava, Slovakia: National Council of the Slovak Republic. <https://www.employment.gov.sk/sk/socialne-poistenie-dochodkovy-system/dochodkovy-system/iii-pilier-doplnkove-dochodkove-sporenie/zhodnotenie-majetku/>.

NN TATRY – SYMPATIA, D.D.S., A.S. (2020). *Annual report 2020* (in Slovak: Výročná správa 2020) [online]. Bratislava, Slovakia: NN dôchodková správcovská spoločnosť, a.s. [cit. 10.8.2021]. <https://www.nn.sk/archiv/sk-nn/tlacove_centrum/VS_NN_DDS_2020.pdf>.

OECD. (2020a). *Hausehold debt* [online]. Paris, France: OECD. [cit. 20.8.2021]. <https://data.oecd.org/hha/household-debt.htm#indicator-chart>.

OECD. (2020b). *Hausehold savings* [online]. Paris, France: OECD. [cit. 20.8.2021]. <https://data.oecd.org/hha/household-savings.html>.

OECD. (2020c). *Net pension replacement rates* [online]. Paris, France: OECD. [cit. 22.10.2021]. <https://data.oecd.org/pension/net-pension-replacement-rates.htm>.

Pensions at a Glance. OECD and G20 Indicators [online]. (2017). Paris, France: OECD Publishing. [cit. 14.8.2021]. <https://doi.org/10.1787/pension_glance-2017-en>.

OECD. (2012). *OECD Pensions Outlook 2012* [online]. Paris, France: OECD Publishing. [cit. 14.8.2021]. <https://dx.doi.org/10.1787/9789264169401-en>.

Pensions at a Glance. OECD and G20 Indicators [online]. (2019). Paris, France: OECD Publishing. [cit. 14.8.2021]. <https://doi.org/10.1787/b6d3dcfc-en>.

PRIVATE PENSION ASSETS SLOVAKIA. (2020). *Insurance corporations and pension funds statistics* [online]. Bratislava, Slovakia: National Bank of Slovakia. [cit. 13.7.2021]. <https://www.nbs.sk/en/statistics/financial-institutions/insurance-companies-and-pension-funds/insurance-corporations-and-pension-funds-statistics>.

SOCIAL INSURANCE AGENCY. (2021a). *Pension from the second pillar* [online]. Bratislava, Slovakia: Social insurance agency. [cit .14.8.2021]. <https://www.socpoist.sk/dochodok-z-ii-piliera-iwj/59350s>.

SOCIAL INSURANCE AGENCY. (2021b). *Woman on maternity leave and the second pillar* [online]. Bratislava, Slovakia: Social insurance agency. [cit. 14.8.2021]. <https://www.socpoist.sk/poradna/326s?prm2=13830>.

SOCIAL INSURANCE AGENCY. (2021c). *Voluntarily insured* [online]. Bratislava, Slovakia: Social insurance agency. [cit. 14.8.2021]. <https://www.socpoist.sk/kto-moze-vyuzit-dobrovolne-poistenie/55401s#2>.

SOCIAL INSURANCE AGENCY. (2021d). *The reference amount valid for 2020 is 464.60€* (in Slovak: Referenčná suma platná na rok 2020 je 464,60 eur) [online]. Bratislava, Slovakia: Social insurance agency. [cit. 12.2.2022]. <https://www.socpoist.sk/aktuality-referencna-suma-platna-na-rok-2020-je-464-60-eur/48411s68093c>.

Social Insurance Act [online]. (2021). Bratislava, Slovakia: National Council of the Slovak Republic. [cit. 10.8.2021].<https://www.zakonypreludi.sk/zz/2003-461#cast1>.

ŠPIRKOVÁ, J., SZŰCS, G., KOLLÁR, I. (2019). Detailed view of a payout product of the old-age pension saving scheme in Slovakia. *Ekonomický časopis*, 67(1): 761–777.

STABILITA, D.D.S., A.S. (2021). *Annual report on the management of own property the supplementary pension company* (in Slovak: Ročná správa o hospodárení s vlastným majetkom doplnkovej dôchodkovej spoločnosti) [online]. Bratislava, Slovakia: Stabilita, d.d.s., a.s. [cit. 10.8.2021]. <https://www.stabilita.sk/media/object/5292/rocna_sprava_o_hospodareni_s_vlastnym_majetkom_doplnkovej_dochodkovej_spolocnosti_stabilita__d.d.s__a.s._k_31_12_2020.pdf>.

SUPPLEMENTARY PENSION COMPANY OF TATRA BANK. (2021). *Annual report 2020* (in Slovak: Výročná správa 2020) [online]. Bratislava, Slovakia: Supplementary Pension Company of Tatra Bank. [cit. 10.8.2021]. <https://www.ddstatrabanky.sk/sk/dokumenty/spravy.html>.

TATRA BANK. (2021). Retirement savings (in Slovak: Sporenie na dôchodok) [online]. Bratislava, Slovakia: Tatra bank. [cit. 5.9.2021]. <https://www.tatrabanka.sk/sk/zivotne-momenty/sporenie-na-dochodok>.

STATISTICAL OFFICE OF THE SLOVAK REPUBLIC. (2021a). *Life expectancy* [online]. Bratislava, Slovakia: Statistical Office of the Slovak Republic. [cit. 2.9.2021]. <https://slovak.statistics.sk>.

STATISTICAL OFFICE OF THE SLOVAK REPUBLIC. (2021b). Gross wage and median [online]. Bratislava, Slovakia: Statistical Office of the Slovak Republic. [cit. 2.9.2021]. <http://statdat.statistics.sk>.

UNIQA D.D.S., A.S. (2020). *Annual report 2020* (in Slovak: Výročná správa 2020) [online]. Bratislava, Slovakia: Uniqa d.d.s. [cit. 10.8.2021]. <https://finstat.sk/35977540/zavierka>.

WORLD BANK. (2019). Worldwide governance indicators [online]. Washington, D.C.: World Bank. [cit. 18.8.2021]. <http://info.worldbank.org/governance/wgi/index.aspx#reports>.

WORKING AGE POPULATION. (2021). *Economically active population by age until 2020* (in Slovak: Ekonomicky aktívne obyvateľstvo podľa veku do roku 2020) [online]. Bratislava, Slovakia: Statistical Office of the Slovak Republic. [cit. 2.8.2021]. <http://datacube.statistics.sk/;#!/view/sk/VBD_SK_WIN2/pr3804qr/v_pr3804qr_00_00_00_sk>.

The Importance of Determinants of Transition from Unemployment to Self-Employment: Evidence from Slovak Micro-Data

Peter Pisár[1] | *Matej Bel University, Banská Bystrica, Slovakia*
Alexandra Mertinková[2] | *Matej Bel University, Banská Bystrica, Slovakia*
Miroslav Šipikal[3] | *University of Economics, Bratislava, Slovakia*
Mária Stachová[4] | *Matej Bel University, Banská Bystrica, Slovakia*

Received 10.2.2022, **Accepted** (reviewed) 8.3.2022, **Published** 17.6.2022

Abstract

The study empirically analyses the determinants of self-employment from unemployment in Slovakia in the period of economic boom. The previous employment of individuals before support is proving to be an important factor in the transition to self-employment. We believe that the importance lies in gaining a practical basis from past jobs, market orientation or establishing contacts before starting a business. Practical courses and support in the form of a tax loan would contribute to the creation of value-added business ideas that have a better chance on the labour market (because after support there is entrepreneurship only in less capital-intensive industries). The paper examines short-term and long-term perspectives using decision trees and random forests, which are exceptionally used in the study of public support. At the same time, research is enriched with practical perspectives, which significantly increases the information base of research.

Keywords	***DOI***	***JEL code***
Active employment policy, contribution to self-employment, decision trees, random forests, the importance of factors, Slovakia	*https://doi.org/10.54694/stat.2022.8*	*J48, H53, H50*

[1] Department of Finance and Accounting, Faculty of Economics, Matej Bel University, Tajovského 10, 975 90 Banská Bystrica, Slovakia. Corresponding author: e-mail: peter.pisar@umb.sk, phone: (+421)905272165.
[2] Department of Finance and Accounting, Faculty of Economics, Matej Bel University, Tajovského 10, 975 90 Banská Bystrica, Slovakia. E-mail: alexandra.mertinkova@umb.sk, phone: (+421)918158105.
[3] Department of Public Administration and Regional Development, Faculty of National Economy, University of Economics, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia. E-mail: miroslav.sipikal@euba.sk, phone: (+421)907780346.
[4] Department of Quantitative Methods and Information Systems, Faculty of Economics, Matej Bel University, Tajovského 10, 975 90 Banská Bystrica, Slovakia. E-mail: maria.stachova@umb.sk.

INTRODUCTION

In many European countries, support for self-employment is characterized by a huge number of support instruments, mainly public assistance for the unemployed, which is combined with education. Despite the considerable amount and popularity, support for self-employment in EU countries is financed by a very low share of total expenditure (below 1% in the long run). Support in Slovakia is below 0.05% of GDP (Eurostat, 2021). Although self-employment support is financially undersized, empirical evidence suggests positive effects of support, with most studies being in Western European countries (Germany: Niefert, 2010; Baumgartner, Caliendo, Kopeinig, 2008; Caliendo and Künn 2013; United Kingdom: Meager, Bates, Cowling, 2003; Finland: Haapanen and Tervo, 2009; France: Duhautois, Redor, Desiage, 2015; Spain: Millán and Congregado, 2010; Cueto, Mayor, Suárez, 2015).

The main idea of support is to place the unemployed in the labor market. According to several studies (Pfeiffer and Reize, 2000; Reize, 2004; Andersson and Wadensjö, 2007), self-employed persons do not differ in terms of socio-demographic characteristics from self-employed persons without support, but the differences are perceived mainly by smaller company size, less capital-intensive business and their business is growing slower. According to the study Haapanen and Tervo (2009), if it is not a push effect, but the entry into self-employment is from paid employment, it will be more sustainable due to higher human capital, motivation, and better information about business opportunities. In some cases, self-employment has a double effect, but for the unemployed it is more of a rarity.

Survival of self-employment from unemployment varies from study to study. While German studies (Pfeiffer and Reize, 2000; Reize, 2004) do not show a significant difference, our previous findings point to a significant difference (as do studies Cueto and Mato, 2009; Haapanen and Tervo, 2009). Significant differences between countries may be justified by the overall duration of the aid. While there is three years of support in Slovakia, in other countries the time is much shorter. E.g. in Sweden only for a period of 6 months (copies unemployment benefits). In Germany, according to a study by Caliendo and Künn (2013), support is also paid for the first 6 months. The Niefert study (2010) reveals the fact that subjects must prove their personal and professional suitability, which can also affect the survival of self-employment and prolong the period of receiving support.

The importance of behavioural aspects, the push effect of support and the regional specifics in which support, is provided in this article. The aim of the article is to examine the importance of factors of self-employment from unemployment in Slovakia in times of economic boom. Slovakia is the country with one of the lowest funding volumes in this area within the EU countries, and, at the same time, is not an economically strong country (GDP per capita is only 71% of the EU average). At the same time, however, it achieved a very significant improvement in the unemployment rate during the period under review; in 2012, the unemployment rate was 14.4%, so in 2017 it was only 5.94%. At the same time, Slovakia is a country with very significant regional differences, while regional characteristics are relatively rarely analysed in similar studies (Caliendo and Künn, 2013). The government has identified the areas of the south and east of Slovakia (Prešov, Košice and Banská Bystrica regions) as the least developed regions, which we include in our analysis as a significant element in survival of self-employment.

The benefit and originality also lies, in addition to the inclusion of the regional dimension, in the enrichment of research with unique factors such as the length of registration at the Office of Labour, Social Affairs and Family, the impact between the place of activity of self-employment and the residence of the supported entity. At the same time, the article is based on modern research methods (decision trees, random forests), which both classify and indicate the impact, which is an information-enriched approach compared to the traditional method of logistic regressions. The methods are considered in the field of public support as new methods, as there is no evidence of their application in the evaluation of active employment policy.

The article consists of three parts. The first part is a literature review, which reflects the current state of the issue of factors of self-employment. The methodological part of the paper defines a specific instrument in Slovakia and methods for evaluating the importance of individual factors. The results and the conclusion combine the acquired knowledge about self-employment in comparison with the opinions from practice and at the same time submit proposals for further research in this area.

1 LITERATURE SURVEY

Potential factors influencing survival of self-employment can be monitored in different phases. The first of these is the phase in obtaining support, which means that the individual is still only a candidate for the contribution. Whether or not they receive a contribution is determined by law and the structure of the conditions for obtaining contributions. There are also factors such as administrative complexity, or the time or complexity of meeting the conditions necessary to obtain support. These factors are very individual, and more and more studies rely on behavioural aspects such as motivation to start a business and the necessary preparation for it (Bořik and Caban 2013; Caliendo and Kritikos 2010).

The second phase, which can be observed during the support, is a very risky phase, namely the survival of support. Here we also observe various influences that may make it difficult to successfully implement the instrument. Based on an overview of factors in research published so far and country specifics (grouped in the previous research by Pisár, Mertinková and Šipikal, 2021), we monitor 3 categories of factors, namely:

(1) Socio-demographic factors (gender, marital status, age, education, last and previous records held at the labour office, previous employment);
(2) Regional labour market (employment activity in an underdeveloped region, place of employment activity equal to the region where the employment activity takes place);
(3) Economic factors (amount of aid, year of granting aid, economic cycle).

Studies Kuang-TaLo, Jiun-NanPan, ShuPeng (2020), Caliendo, Künn, Weißenberger (2016) have identified a positive statistically significant impact of men on keeping self-employment from unemployment. However, there are common differences between gender. Women are less tied to the labour market, earn less and have stronger family responsibilities regardless of participation status (Bořík, Ďurica, Molnárová, Švábová, 2015).

In the case of marital status, we observe that support is not more sustainable for singles and therefore it can be assumed that family support creates a better background in business. Even some studies (Niittykangas and Tervo, 2005) look for connections with previous family entrepreneurship, which can positively affect the sustainability of the current one (learning from parents, helping from a young age in entrepreneurship). Another factor related to marital status is the number of children examined by Caliendo and Kritikos (2010); Caliendo and Künn (2013); Millán and Congregado (2010) or Haapanen and Tervo (2009). Another additional variable of the study (Caliendo and Künn, 2013; Caliendo and Kritikos, 2010) reports health status or working time (Caliendo and Kritikos, 2010; Caliendo and Künn, 2013; Millán and Congregado, 2010).

Studies Holtz-Eakin, Joulfain, Rosen (1994), Parker (2004) suggest that support retention rates are higher in middle age than in younger or older self-employed people.

In the case of education, there are conflicting views. A study by Pankaj and Marcus (2019) found that self-entrepreneurs who have better financial abilities based on education and experience, achieve higher prosperity and can maintain support for longer. Studies Bořík, Ďurica, Molnárová, Švábová (2015), Parker (2004), Niefert (2010) also find that higher levels of education enter self-employment as one of the positive factors of sustainability. However, there are also conflicting views and the impact of the educational level of subjects is uncertain, according to a study by Baumgartner and Caliendo (2008). It is expected that the higher the level of education attained, the lower the likelihood of choosing to become self-employed, as other opportunities in the labour market open up for the subjects.

The length of registration at the employment office and previous job also plays an important role in maintaining employment support. Entities with business experience have higher human capital, motivation and better information about business opportunities (Haapanen and Tervo, 2009). The study further explains that starting a business in a new environment also brings unexpected risks such as search for suppliers or customers.

In summary, the theory does not clearly determine the order and importance of factors in self-employment from unemployment but has defined an appropriate selection of factors from the domestic and international environment.

2 METHODS

The method of implementing support for self-employment in Slovakia is to be found in *the contribution to self-employment* in accordance with §49 of the Employment Services Act, which is an intensively used instrument of active policy. The amount of support is granted within 30 days in the amount of 60% and then the remaining 40% of support for the past year. At the same time, the amount of support is conditioned by the place of self-employment. In the case of the least developed regions (southern and eastern Slovakia), where districts achieve an average registered unemployment rate higher than the national average, the contribution is at most 4 times the total price of labour calculated from the average wage of an employee. In the case of a lower unemployment rate, it is 3 times, while in the Bratislava region (the region with the highest GDP per capita) it is only 2.5 times.

In our conditions, the support is only for job seekers who are kept in the records of the Office of Labour, Social Affairs and Family for at least 3 months. The contribution is provided for the partial payment of costs related to the creation of a job for self-employment and the subsequent operation of self-employment for at least three years. There are no phases in Slovakia where we would educate individuals (we only monitor the preconditions for entrepreneurship in the form of a business plan). The research methodology is based on 2 models that evaluate the survival of support in the short term (after 6 months) and in the long term (after 3 years), which points to differences in time. The following table shows the methodology of the models.

Table 1 Methodology of models

Model name	Model parameters	
Model3_3SZCO	Survival of support after 3 years	The supported entity survival on the labour market (only cases of self-employed)
Model4_6SZCO	Survival of support after 6 months	The supported entity survival on the labor market (only cases of self-employed)

Source: The authors

A dependent variable in research is the survival of jobs created through support of the Contribution to self-employment. In each model 11 independent variables are analysed, as shown in Table 2. The selection of selected variables reflects other studies enriched with new factors and is also created in terms of Slovakia-specific data, which were discussed in the study Pisár, Mertinková, Šipikal (2021). The analysis includes the phase of survival of support in the period 2012–2016 in the Slovak Republic. The data sets with the data for the Contribution to self-employment were created from data that were processed based on 2 unique databases of the Ministry of Labour and Social Affairs on supported entities.

Table 2 Description of the variables that affect the probability of employment and self-employment

Name of variable	Variable description and coding
	Dependent variables
Labour	3 years after the end of the support period, the supported entity is employed as self-employed = 1, otherwise = 0 (model3_3SZCO). 6 months after the end of the support period, the supported entity is employed as self-employed = 1, otherwise = 0 (model4_6SZCO).
	Independent variables
Gender	Gender of job seeker (male = 1, female = 0).
Age	Age of the subject on the labor market in years.
Marital_status	Marital status: if job seekers is single = 1, otherwise = 0.
History_1	Last length of days at the employment office (in days).
History_2	Previous length of days at the employment office (in days).
Education	Achieved level of education. Code: 0 – without education or unfinished primary school, 1 – primary education (ZŠ), 2 – lower vocational education (NOV), 3 – secondary vocational education (SOV), 4 – complete secondary education (USOV), 5 – higher vocational education (VOV), 6 – 1st level of university education (Bc.), 7 – 2nd level of university education (Mgr./Ing.), 8 – 3rd level of university education (PhD.).
Job_previous	Previous employment of job seeker – code according to the statistical classification of occupations. Code: 1 – Managers and legislators, 2 – specialists, 3 – Technical and professional staff, 4 – Administrative staff, 5 – Service and trade workers, 6 – Skilled workers in agriculture, forestry and fishing, 7 – Skilled workers and craftsmen, 8 – Operators and fitters of machinery and equipment, 9 – Auxiliary and unskilled workers. Variables used separately as dummy variables.
Support_year	Year in which the support was granted. The scope of the monitored period is in the range 2012–2016. The code assigned to each year is as follows: 1 – 2012, 2 – 2013, 3 – 2014, 4 – 2015 and 5 – 2016.
Region	If the job seeker is from the Prešov, Košice and Banská Bystrica self-governing regions (7; 8; 6) = 1, otherwise = 0.
Same_place	If the job seeker has the same NUTS code of residence and place of work, then the variable = 1, otherwise = 0.
Aid_amount	The total amount of financial support from the instrument for self-employment allocated to the job seeker (in Euros).

Source: The authors

Given the defined goal, the research questions are as follows:

Research question 1: *Which factor is key in self-employment and what is the target value in the root node?*

We use the decision tree method to determine the predictor and the target value in the root node. The *rpart* package (Therneau, Atkinson, Ripley, 2019) and the *rpart.plot* package (Milborrow, 2021), which contains the CART algorithm (*Classification and Regression Trees*) in the R program, were used to produce the outputs. The value of the explanatory variable was assigned using the labour variable.

The decision tree (or *tree diagram*) is a decision support tool that uses a tree as a decision model. In data mining and machine learning, the decision tree is a predictive model. This means that it informs from observations about the item to conclusions about its target value. In these tree structures, leaves represent classifications and branches represent combinations of characters that lead to these classifications (Stachová and Kráľ, 2010).

The most common decision algorithm is the CART algorithm. It is a form of binary division. In our case, self-employment in the decision-making node can be divided into only two groups. Thus, each parent node can lead to two child nodes, and each of these child nodes can split itself and create additional children.

CART analysis has a number of advantages over other classification methods, including classical multidimensional logistic regression. First, it is essentially nonparametric, so no assumptions are made regarding the basic distribution of the values of the predictor variables. Thus, CART can process numerical data that is highly distorted or multimodal, as well as categorical predictors with ordinal or non-ordinal structure. This is an important feature because it eliminates the time an analyst would otherwise spend finding out if the variables are normally distributed and performing the transformation if they are not. In addition, this algorithm uses the ‚white box' model, in other words, the situation is observable in contrast to neural network models (Stachová and Kráľ, 2010).

Rese-arch question 2: *How important are the individual factors of survival of self-employment in Slovakia?*

To evaluate the importance, we use the random forest method (one tree does not vote, but a group of trees determines the importance of individual factors). *The randomForest* package (Liaw and Wiener, 2018) in the R program was used for the chosen method. A data set of predictors was inserted in item x and a dependent variable was entered in item y, which in our case is the same as in decision trees. In this method, the results for the independent variables are distinguished based on the MDA (*Mean Decrease Accuracy*) indicator. The indicator was used as measurement of factors importance. It means, that the more the accuracy of our model suffers without this factor the more important the factor is.

The random forest is a classifier of a machine learning file that consists of many decision trees and issues a class that is a mode of class output according to individual trees. Many classification trees grow in random forests. Each tree gives a classification, and we say that the tree ‚votes' for this class. The forest chooses the classification with the largest number of votes (Stachová and Kráľ, 2010).

The list of the above methods is chosen mainly for their ability to select the most important information from a large number of options. The methods are usually used as a tool to support decision-making in the areas of company productivity, risk minimization or revenue maximization and thus to reduce the company's bankruptcy. The principle of examining self-employment is very similar. While in business practice models follow the survival of productivity, we will monitor the survival of support. What factors emerge as the most important will tell us how to minimize the risk of wasting public funds or maximize the survival of such support.

Research question 3: *Why are these factors of great importance in the support of self-employment and what is the opinion of practitioners?*

In addition to generating rare and heterogeneous evidence on the importance of self-employment factors, the paper, on the other hand, seeks to contribute to the understanding of the factors that lead

to these peculiarities in Slovakia. Therefore, in the discussion, we draw certain connections regarding the behaviour of support derived from the opinions of people from practice. Opinions will thus contribute to a more realistic view of the importance of factors.

We contacted the staff of the Labour, Social Affairs and Family Office, who are the first contact in obtaining such support. At the same time, representatives of institutions directly related to self-employment also took part in the professional discussion. These are the National Bank of Slovakia, where the labour market expert participates, the Institute of Employment Policy, where the President of the Institute spoke, and the Slovak Chamber of Commerce, which represents the interests of self-employed persons. Opinions are collected by the Delphi survey method, where they express an attitude towards each factor addressed. Confirmation resp. refute the knowledge gained and express the gradual importance of the factors.

The survey was conducted in two rounds, which reflected the results already found regarding the effects of factors. More precisely, in September 2021, employees at the labour office from the districts that record the most supported subjects were contacted. A month later (October 2021), further telephone conversations were held with representatives of employment policy in Slovakia. A total of 8 employees and 3 of the institutions took part in the survey.

3 RESULTS

Our strategy for selecting variables is based on similar studies that examine the determinants of survival on the labour market. The selection strategy reflects national specificities. Supported entities are divided according to 2 models – see Table 3.

Table 3 Division of subjects in models

Name of model	Share of persons placed as self-employed entity	Share of persons not placed as self-employed entity
Model3_3SZCO	39.05%	60.95%
Model4_6SZCO	56.22%	43.78%

Source: The authors

The largest proportion of supported entities are men. The average age is 37 years. 21% of entities did not have previous employment, 44% are single and 46% of them have completed higher vocational education (in 23.4% they were job seekers from the 2nd level of higher education). Most often they come from the regions: Žilina, Prešov and Banská Bystrica, two of which are among the least developed regions in Slovakia. Their average amount of contribution reached € 3 570, while they spent an average of 1 368 days in the register (approximately 3 year and 7 months, which can be considered as long-term unemployed). We observe the largest share of previous employment in the supported entities in the field of technicians and professionals; service and trade workers and craftsmen, processors, and repairers (according to the ISCO classification).

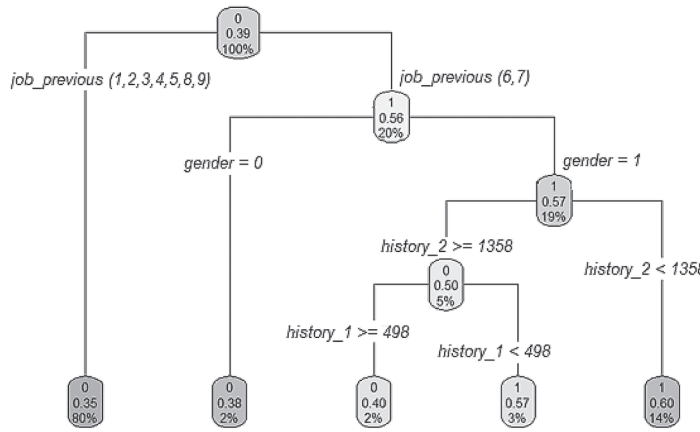
3.1 Determination of the most important factor and the target value of survival of self-employment

We follow the answers to the first research question using the decision tree method. The bushyness of the tree in the models was set to minimize the cross-validation error (xerror) and at the same time the tree was not too bushy and therefore opaque. The resulting models in the case of self-entrepreneurs are shown in Figures 1 and 2.

The results show that the *Model3_3SZCO* model distributed to entrepreneurs after 3 years in the root node and identified the previous job as a predictor. Self-employment, which met the following condition

(they were from the category 1, 2, 3, 4, 5, 8, 9), were not placed on the labour market for up to 3 years. Other subjects are redistributed in other nodes according to conditions related to predictors (e.g. education of the subject, the last and previous length of days at the employment office). In other words, with a closer understanding of the survival of support (*Model3_3SZCO*), we observe that the greatest amount of variability is explained by the variable of previous job.

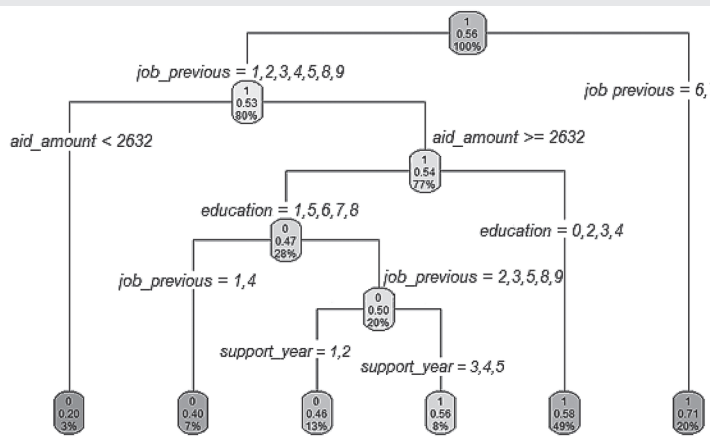
Figure 1 Decision tree for model *Model3_3SZCO*



Source: Own processing in program R

In the case of self-employment from unemployment, survival after 6 months (*Model4_6SZCO*) identified in the root node as the main predictor just the previous job, which is identical to the long-term survey. Other subjects are redistributed in other nodes according to conditions related to predictors such as aid amount, education of the subject or year of provision. In other words, with a closer understanding of survival, we observe that the greatest amount of variability is explained by the variable of previous job. Those whose previous job was from the field of skilled workers and craftsmen or skilled workers from the field of agriculture, forestry and craftsmanship were able to be placed on the labour market (the same as in the *Model3_3SZCO* model, which concerns placement after 3 years).

Figure 2 Decision tree for model *Model4_6SZCO*



Source: Own processing in program R

The predictive abilities of the created decision trees are expressed in Table 4 together with its errors of the first and second kind[5]. The results show a relatively good predictive ability of the models in the range of 60.56% to 64.01%.

Table 4 Classification table – decision trees

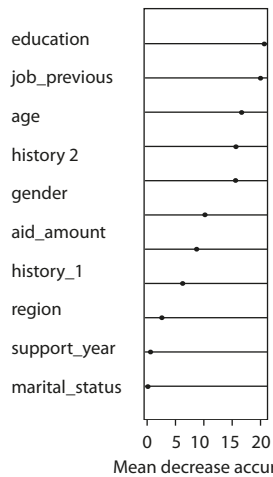
Model	***Model3_3SZCO***		**% correctly classified subjects**	**Model**	***Model4_6SZCO***		**% correctly classified subjects**
The actual classification	0	1		**The actual classification**	0	1	
0	3 769	467	54.23%	0	1 367	3 010	13.67%
1	2 034	680	9.78%	1	933	4 687	46.88%
Total predictive power	64.01%			Total predictive power	60.56%		

Source: Own processing in program R

3.2 Determination the importance of factors in survival of self-employment

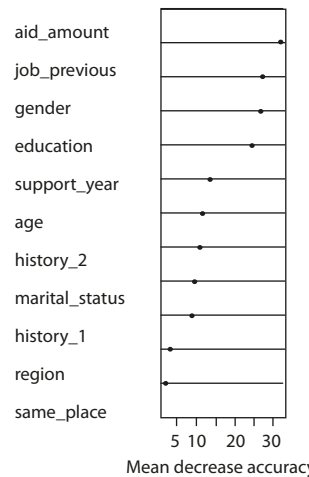
We follow the answers to the second research question by the method of random forests. The self-employment factors shown in this way (Figures 3 and 4) form the predictive power of the random forest model. If we consider deleting the upper variable, the predictive power of the model will be significantly reduced. On the other hand, if we consider reducing one of the lower variables, it may not have much effect on the predictive power of the model. Therefore, we consider the first variables to be the most important variables in the models.

Figure 3 Variable importance for model *Model3_3SZCO*



Source: Own processing in program R

Figure 4 Variable importance for model *Model4_6SZCO*



Source: Own processing in program R

[5] The error of the first kind expresses the % of truly sustainable self-employment, which the model mistakenly classified as unsustainable, and we get it if we subtract the percentage of correctly classified from 100%. The error of the second kind expresses the percentage of unsustained self-employment that the model incorrectly classified as sustainable and can be found by analogy

According to the chosen MDA indicator, based on ‚tree voting', education was identified as the most important variable influencing the long-term maintenance of self-employment support, while second place belongs to previous job (according to the decision tree is the dividing criterion in model *Model3_3SZCO*).

The results of models up to 6 months (short-term observation) show that in the model *Model4_6SZCO*, we can mark variables according to their importance as the aid amount, previous job and gender.

The model differs in the importance of variables in terms of time, while in self-employment from unemployment, the amount of funding provided plays an important role, in the long run it is the level of education. Just behind these factors is the previous job, the importance of which is significant in both respects.

The predictive ability of the created random forests is expressed in Table 5. The results show a relatively good predictive ability of the models in the range of 59.52% to 61.61%.

Table 5 Classification table – random forests

Model	***Model3_3SZCO***		**% correctly classified subjects**	**Model**	***Model4_6SZCO***		**% correctly classified subjects**
The actual classification	0	1		**The actual classification**	0	1	
0	3 359	840	20.00%	0	1 911	2 434	56.02%
1	1 808	891	66.99%	1	1 593	4 010	28.43%
Total predictive power	61.61%			Total predictive power	59.52%		

Source: Own processing in program R

In addition to the above results, we try to determine the links with views from practice on the survival of self-employment from unemployment and draw possible conclusions for Slovakia.

4 DISCUSSION

The use of decision trees and random forests was effective. These algorithms have been shown to be classification techniques that are easy to understand, interpret and can be used in the public sector. These techniques, despite leaving the reality of the data, achieved satisfactory predictive accuracy. The models were able to learn to recognize placed rather than unplaced subjects due to the lower number in the examined models.

The knowledge gained through decision trees and random forests is largely identified with practical views. The most important factors such as previous job before support and the subject's past at the employment office are always in the first place. Opinions from practice also confirmed the importance of the age factor, where the staff of the Office ranked it as the most important. A little higher, the practice would include the factor of operation in an underdeveloped region compared to our results. Both research and practice have decided that the factor marked *same_place* is the least important factor. The most controversial factor was the gender of the subjects.

What specific statements provided by practitioners (to confirm/refute the importance of the factor) will show a detailed examination of the 5 most important factors: length of last unemployment, previous job, age, education and the amount of financial support.

The most important factor from the category of socio-demographic factors is the previous job. We assume that its significance lies in obtaining a practical basis from past employment, market orientation in the given area or obtaining contacts before starting a business. In our sample, there are 21% of entities with zero pre-support experience, which may emphasize the importance of the factor.

Despite the high significance of the factor, there are no phases in Slovakia where we would monitor the practical experience of individuals before starting a business (we only monitor the preconditions for doing business in the form of a business plan). We often meet that the business plan is developed by a third party and so we do not know the real business knowledge. In other countries, it is often possible to combine support for self-employment with education (Oberschachtsiek and Scioch, 2015; Wolff, Nivorozhkin, Bernhrard, 2016).

To capture the importance of this factor, we recommended to check the practical experience of future self-employed person, but the views of practice are conflicting. An expert from the National Bank of Slovakia states as a counterargument that *´overdiversification of the experience would lead to complications, more bureaucracy and probably would also lead to discouragement of some people. Therefore, I would not even try to introduce. However, I agree with the statement that previous job has an impact on the success and survival of the self-employment´*. The director of the Slovak Chamber of Commerce very similarly expresses the opinion that *´verification of the level of business skills would be disproportionately financially demanding in practice and unnecessary, as knowledge alone is not the only necessary skill for successful entrepreneurship´*.

An interesting observation was the statement of the President of the Institute of Employment Policy, who states that *´the review of experience should take place, but only in areas of business that require deeper practical experience´*.

It can therefore be argued that previous employment is understood by both practice and empirical evidence as the most important factor in self-employment. The practice largely states that we do not need to know the business experience of self-entrepreneurs (financially and administratively demanding). The form of preparation and gaining experience can be practical courses before support, where the entity itself recognizes what it requires deeper experience (how to gain contacts, product presentation, market research and competition ...). We note that not in current areas such as taxes, accounting and levies, as this service will still be provided externally. Thus, the contribution would ensure access for all without distinction, but at the same time it would offer the possibility of education before the actual application. This would naturally select applications for self-employment and decisions would be more certain.

Another important factor is education. Education is the starting point before being placed on the labour market, which we consider to be the main reason for the importance of the factor. Education has a shorter and less intensive duration compared to previous jobs, which confirms that previous experience is more important than education.

We note that the significance of the factor is monitored in the period 2012–2016, which is a period of economic boom. The entity is essentially "pushed" to choose to become self-employed in a less developed region, as the region does not offer opportunities equivalent to its education. At the same time, he is also affected by social influences such as strong ties in the region, contacts, background, ... (more pronounced if the subject decides to be a self-employed person due to unemployment). Representatives of the institutions add that the area of business is important. e.g. those doing business in the IT sector have literally "unlimited possibilities" at the global level.

It can therefore be argued that education affects the survival of self-employment support from unemployment, especially in the long term (it is a kind of starting point for placement in the labour market). In the case of other models and from the point of view of practice, education is less important than previous practical experience and evaluates it as a moderately strong factor. The level of education mainly affects entities that are pushed to support in a less developed region.

An important factor in the field of economic factors is the amount of financial support. Sufficient financial resources are logically more important for self-employed people from unemployment than from employment[6]. The fact that this effect is mainly short-term can be explained by its importance when starting a business. In the long run, sufficient financial resources replace other factors (e.g. education, age, ...).

The assumption of the importance of funds in starting a business is largely confirmed by practice. The employees of the labour office confirm the importance, while the contribution will provide the greatest help with the input capital, such as rent or material and technical equipment. Furthermore, but they draw attention to the fact that *´the financial value of the support is sufficient only for starting a business, in the later period the business idea must generate complementary funds´*. Based on the above findings, our question was whether a higher amount of aid would not provide a better labour place, a better product/service and thus a higher added value of the investment. Some authors (Niefert, 2010) also dispute whether subsidies should not be exchanged for loans that entities would have to repay. This would increase the motivation to keep the business going and prosper in the future. The President of the Institute for Employment Policy recommends a ´golden mean´ called tax loan. Instead of a grant that covers only basic needs, the business would be subsidized by a more significant repayable amount. It would probably attract more educated people who would otherwise choose paid employment.

It can therefore be argued that the amount of aid is an important factor mainly from the point of view of self-employed persons and is mainly short-term. Financial resources are the first step to the successful survival of self-employment, but the main focus of survival shifts to the human capital of the entity, which is hidden in previous practical experience, education or length of registration at the employment office.

An important factor in self-employment (especially in terms of practice and a broader understanding of the purpose of the contribution) is age.

Age plays an important role in our analysis, with the average being 37 years. Studies Holtz-Eakin, Joulfain, Rosen (1994), Parker (2004) suggest that support retention rates are higher in middle age than in younger or older self-employed people. E.g. in a study by Bořík, Ďurica, Molnárová, Švábová (2015) an average of 34 years, Ellen et al. (2021) approximately 46 years or Parker (2004) with an age limit of 40.

The assumption of higher survival of support at middle age has been partially confirmed. There are several reasons. From the point of view of an NBS expert, this is mainly a *´combination of experience and greater knowledge of the market. At a younger age, it may be about testing what will work on the market, respectively finding your place´*.

The president of the institute justified higher sustainability by choosing a business. *´while the elderly have a choice of less risky businesses than the liberal professions (consulting, expertise, expertise), the younger ones are dominated by start-ups´*. According to the employee of the labour office, burnout, health or pension are associated with older age.

Another significant factor from the category of socio-demographic factors is the length of registration at the employment office. We assumed that the high significance of the factor is caused by the fact that subjects (whose registration at the labour office is on average over 3 years) have lost work habits. The entity is likely to enter the labour market with limited access to information on business opportunities, which may lead to an insufficient business idea (most often goods and services without added value).

[6] Entities are likely to be affected by factors such as the network of contacts, good work habits and business experience that have brought them new work experience.

Despite the questionable profitability of a business, their push effect makes it work, which ultimately has a negative impact on the survival of business.

The presumption of loss of work habits due to long-term placement in the employment office or limited access to information that will lead to an insufficient business idea has been largely confirmed from the point of view of practice.

The director of the Slovak Chamber of Commerce adds that *´work habits and real contact with the labour market usually weaken after only 6 months of unemployment´*.

It can therefore be argued that the length of registration at the employment office is understood by both practice and empirical evidence as one of the most important factors. Experience agrees that the loss of work habits and real contact make the support unsustainable in the short and long term.

CONCLUSION

The aim of the paper was to examine the importance of the factors of self-employment in Slovakia in times of economic boom. The choice of the contribution was chosen due to its significant effectiveness according to the research carried out so far and the relatively lowest support among all active instruments of the EU countries.

The contribution and originality of this state compared to other studies lies in the enrichment of research with unique factors such as the length of registration at the Office of Labour, Social Affairs and Family or specifically capture the least developed regions. At the same time, the article is based on modern research methods (decision trees, random forests), which both classify and indicate the impact, which is an information-enriched approach compared to the traditional method of logistic regressions. The methods are considered in the field of public support as new methods, as there is no evidence of their application in the evaluation of active employment policy.

The data are not from freely available databases and therefore self-employment research is a significant milestone in capturing the importance of factors and behaviour of subjects. Factors influencing sustainability were examined from both the short-term (after 6 months) and long-term (after 3 years), with the main idea of supporting the unemployed to be placed on the labour market.

We consider the use of decision trees and random forests to be effective. These algorithms have been shown to be classification techniques that are easy to understand, interpret and can be used in the public sector. These techniques, despite leaving the reality of the data, achieved high predictive accuracy. Even the error rate is relatively low due to keeping the data realistic.

The knowledge gained through decision trees and random forests is largely identified with practical views. The most important factors such as previous job before support and length of registration at the employment office are always in the first place. Other factors are age, amount of financial support and education.

We perceive the importance of these factors on two levels. The first is a consequence of past human capital building events (education, previous employment, length of support and age) and the second is the level of financial assistance, which is less important. These findings suggest possible improvements in several areas. The first is the introduction of practical courses, which could have a different information base than previously provided. Until now, future self-employed person have been able to choose consulting courses in the field of taxes, accounting or levies (for which they will still use an external service), topics such as gaining contacts, product presentation, market research or competition are more important when starting a business. The second improvement concerns the overall setting of the aid, which is rather related to the importance of the amount of aid. Just as some authors (Niefert, 2010) argue over whether subsidies should not be exchanged for loans, we argue about the so-called tax loan. Instead of a grant that covers only basic needs, it could be more efficient to subsidize with a more significant repayable amount, which can also cover value-added businesses (at the same time

more acceptable than a normal loan on the market). It would probably attract more educated people who would otherwise choose paid employment. Thus, a tax loan is not the absolute extreme of any alternative.

We note that the study is formed in terms of macroeconomic factors in times of economic boom, which can significantly affect the behaviour of individuals and policy settings as such. If we look at the current situation (caused by the COVID-19 contagion), which rather points to the downturn in economic development and the labour market, we see that active employment policies respond differently. What impact this will have and how the importance of factors will change will be the subject of further research that will build on the results of this state.

ACKNOWLEDGEMENTS

This research was supported by the Slovak Research and Development Agency (APVV), APVV-21-22678 "Innovations of employment support tools and activation of human capital in Slovakia".

References

Act No. 5/2004 Coll. Act on Employment Services and on Amendments to Certain Acts. (2004). Bratislava: National Council of the Slovak Republic.

ANDERSSON, P., WADENSJÖ, E. (2007). Do the Unemployed Become Successful Entrepreneurs? *International Journal of Manpower,* 28(7): 604–626.

BAUMGARTNER, H. J., CALIENDO, M. (2008). Turning unemployment into self-employment: effectiveness of two start-up programmes [online]. *Oxford Bulletin of Economics and Statistics,* 70(3): 347–373. <https://doi.org/10.1111/j.1468-0084.2008.00505.x>.

BOŘÍK, V., CABAN, M. (2013). *Pilotné hodnotenie dopadov vybraných opatrení aktívnej politiky trhu práce* [online]. MPSVR SR. [cit. 10.2.2021]. <http://www.evaluacia.sk/wp-content/uploads/2012/12/vborik.pdf>.

BOŘÍK, V., ĎURICA, M., MOLNÁROVÁ, M., ŠVÁBOVÁ, L. (2015). *The Net Effects of Graduate Work Experience and the Promotion of Self-employment: Technical Report* [online]. Bratislava: Ministry of Labour, Social Affairs and Family of the Slovak Republic. [cit. 10.2.2021]. <https://www.employment.gov.sk/files/slovensky/esf/op-zasi/technical-evaluation-report_final5edit.pdf>.

CALIENDO, M., KOPEINIG, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching [online]. *Journal of Economic Surveys*, 22: 31–72. <https://doi.org/10.1111/j.1467-6419.2007.00527.x>.

CALIENDO, M., KRITIKOS, A. S. (2010). Start-ups by the unemployed: Characteristics, survival and direct employment effects [online]. *Small Business Economics,* 35(1): 71–92. <https://doi.org/10.1007/s11187-009-9208-4>.

CALIENDO, M., KÜNN, S. (2013). Regional Effect Heterogeneity of Start-Up Subsidies for the Unemployed [online]. *IZA Discussion Papers,* 7460: 1108–1134. [cit. 15.9.2021]. <http://ftp.iza.org/dp7460.pdf>.

CALIENDO, M., KÜNN, S., WEIßENBERGER, M. (2016). Personality Traits and the Evaluation of Start-Up Subsidies [online]. *European Economic Review,* 86: 87–108. <https://doi.org/10.1016/j.euroecorev.2015.11.008>.

CUETO, B., MATO, J. (2009). A nonexperimental evaluation of training programmes: regional evidence for Spain [online]. *The Annals of Regional Science,* 43: 415–433. <https://doi.org/10.1007/s00168-008-0214-2>.

CUETO B., MAYOR M., SUÁREZ P. (2015). Entrepreneurship and unemployment in Spain: a regional analysis [online]. *Applied Economics Letters,* 22: 1230–1235. <http://dx.doi.org/10.1080/13504851.2015.1021450>.

DUHAUTOIS, R., REDOR D., DESIAGE L. (2015). Long Term Effect of Public Subsidies on Start-up Survival and Economic Performance: An Empirical Study with French Data [online]. *Revue d'économie industrielle,* 149: 11–41. <https://doi.org/10.4000/rei.6063>.

EK, E., ALA-MURSULA, L., VELÁZQUEZ, R. G., TOLVANEN, A., KATARIINA SALMELA-ARO, K. (2021). Employment trajectories until midlife associate with early social role investments and current work-related well-being [online]. *Advances in Life Course Research,* 47: 100391. <https://doi.org/10.1016/j.alcr.2020.100391>.

EUROSTAT. (2021). *LMP expenditure by type of action – summary tables* [online]. [cit 17.8.2021]. <https://webgate.ec.europa.eu/empl/redisstat/databrowser/view/LMP_EXPSUMM/default/table>.

HAAPANEN, M., TERVO, H. (2009). Self-employment duration in urban and rural locations. [online]. *Applied Economics,* 41(19): 2449–2461. <https://doi.org/10.1080/00036840802360278>.

HOLTZ-EAKIN, D., JOULFAIN, D., ROSEN, H. S. (1994). Sticking it Out: Entrepreneurial Survival and Liquidity Constraints [online]. *Journal of Political Economy,* 102: 53–75. [cit 15.10.2021]. <http://www.jstor.org/stable/2138793>.

KUANG-TA LO, JIUN-NAN PAN, SHI-SHU PENG. (2020). The role of gender and its potential channels to affect self-employment in Taiwan [online]. *Economic Modelling*, 89: 601–610. <https://doi.org/10.1016/j.econmod.2020.02.030>.

LIAW, A., WIENER, M. (2018). *Package 'randomForest'* [online]. [cit. 10.8.2021]. <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>.

MEAGER, M., BATES, P., COWLING, M. (2003). An evaluation of business start-up support for young people [online]. *National Institute Economic Review,* 186: 59–72. <https://doi.org/10.1177/002795010300100111>.

MILBORROW, S. (2021). *Package 'rpart.plot'* [online]. [cit. 10.8.2021]. <https://cran.r-project.org/web/packages/rpart.plot/rpart.plot.pdf>.

MILLÁN, M. J., CONGREGADO E. (2010). Determinants of self-employment survival in Europe [online]. *Small Business Economics,* 38: 231–258. <https://doi.org/10.1007/s11187-010-9260-0>.

NIEFERT, M. (2010). Characteristics and Determinantls of Start-ups from Unemployment: Evidence from German Micro Data [online]. *Journal of Small Business and Entrepreneurship,* 23(3): 409–429. <https://doi.org/10.1080/08276331.2010.10593493>.

NIITTYKANGAS, H., TERVO, H. (2005). Spatial Variations in Intergenerational Transmission of Self-Employment [online]. *Regional Studies,* 39: 319–32. <https://doi.org/10.1080/00343400500087166>.

OBERSCHACHTSIEK, D., SCIOCH P. (2015). The outcome of coaching and training for self-employment: a statistical evaluation of outside assistance support programs for unemployed business founders in Germany [online]. *Journal for Labour Market Research,* 48(1): 1–25. <https://doi.org/10.1007/s12651-014-0161-6>.

PARKER, S. C. (2004). *The economics of self-employment and entrepreneurship* [online]. Cambridge University Press. <https://doi.org/10.1017/CBO9780511493430>.

PFEIFFER, F., REIZE, F. (2000). From Unemployment to Self-Employment – Public Promotion and Selectivity [online]. *International Journal of Sociology,* 30(3): 71–99. [cit. 17.11.2021]. <https://www.jstor.org/stable/20628598>.

PISÁR, P., MERTINKOVÁ, A., ŠIPIKAL, M. (2021). What Factors Influence the Survival of Subsidised Start-ups for the Unemployed in Slovakia? [online]. *Central European Public Administration Review*, 19(2): 109–130. <https://doi.org/10.17573/cepar.2021.2.06>.

R CORE TEAM. (2020). *R: A language and environment for statistical computing* [online]. R Foundation for Statistical Computing. [cit. 8.4.2021]. <https://www.R-project.org/>.

REIZE, F. (2004). *Leaving Unemployment for Self-Employment. An Empirical Study*. ZEW Economic Studies, Heidelberg: Physica.

STACHOVÁ, M., KRÁĽ, P. (2010). Predicting Financial Distress of Slovak Companies Using Data Mining Techniques. *13th International Scientific Conference AMSE 2010.*

THERNEAU, T., ATKINSON, B., RIPLEY B. (2019). *Package 'rpart'* [online]. [cit. 10.8.2021]. <https://cran.r-project.org/web/packages/rpart/rpart.pdf>.

WOLFF, J., NIVOROZHKIN, A., BERNHRARD, S. (2016). You can go your own way! The long-term effectiveness of a self-employment programme for welfare recipients in Germany [online]. *International Journal of Social Welfare,* 25(2): 136–148. <https://doi.org/10.1111/ijsw.12176>.

The Relationship between Financial Development, Trade Openness and Economic Growth in Turkey: Evidence from Fourier Tests

Havanur Ergün Tatar[1] | *Bartın University, Bartın, Turkey*
Gökhan Konat[2] | *Abant Izzet Baysal University, Bolu, Turkey*
Mehmet Temiz[3] | *Fırat University, Elazığ, Turkey*

Received 23.12.2021 (revision received 10.3.2022), **Accepted** (reviewed) 21.3.2022, **Published** 17.6.2022

Abstract

In this study, the effects of financial development and trade openness on economic growth were investigated using annual data for Turkey over the period 1960–2017. The financial development variable is represented as the ratio of financial system deposits to GDP. The trade openness variable is represented as the ratio of the sum of exports and imports of goods and services to GDP. To examine the long-run relationship between financial development, trade openness and economic growth; Fourier-based stationarity test and its complementary Fourier-based cointegration test are used. Finally, Fourier-based causality tests are also used to examine the causality relationship between the variables. As a result of cointegration tests, a long-term cointegration relationship was found between variables. According to the Fourier Toda-Yamamoto causality analysis results, it is seen that there is a one-way causality relationship from financial development to economic growth and from financial development to trade openness.

Keywords	*DOI*	*JEL code*
Financial development, trade openness, economic growth, unit root, cointegration, causality, Fourier	*https://doi.org/10.54694/stat.2021.46*	*O1, G0, C4*

INTRODUCTION

As the concept of growth is an important macroeconomic factor, there is a large literature on which factors are affected or which factors affect it. Herein, especially after the economic transformation

[1] Department of Economy, Bartın University, 74110 Bartın, Turkey. Corresponding author: e-mail: havanurergun@gmail.com.
[2] Department of Econometrics, Abant Izzet Baysal University, 14280 Bolu, Turkey. E-mail: gokhan.konat@inonu.edu.tr.
[3] Department of Economy, Fırat University, 23119 Elazığ, Turkey. E-mail: mtemiz@firat.edu.tr

process in Turkey after 1980, the concepts of financial development and trade openness came to the fore. In the economic literature, it is accepted that financial development is one of the most important internal variables that significantly affect the economic growth of countries. Goldsmith (1969), McKinnon (1973) and Shaw (1973) pioneered the relationship between financial development and economic growth. While early economic growth theories did not explicitly include financial development as a variable, a growing theoretical and empirical literature shows that financial intermediation makes a significant contribution to economic growth by mobilizing savings, reorganizing the allocation of resources, and diversifying risks. Endogenous growth models claim that financial institutions and markets contribute to long-term economic growth by reducing information and transaction costs, influencing decisions in favor of more efficient activities, and efficiently utilizing the most promising investments (Salahuddin and Gow, 2016).

There are various views in the literature that tries to explain the channels of financial development affecting growth. Some economists focus on the view that financial development directly affects economic growth. Some economists, on the other hand, emphasize that financial development indirectly affects economic growth by fulfilling various functions in providing financial intermediation and reducing transaction costs. (Tadesse and Abafia, 2019). In contrast, some empirical studies suggest that financial development does not affect poverty (Chaouachi and Chaouachi, 2021). At this point, the main functions of financial institutions are considered as efficient allocation of economic resources, improved capital accumulation and improvement in sufficiency (Tadesse and Abafia, 2019).

The concept of financial development is defined as the increase in the services of financial intermediaries, especially banks. The transformation and development in financial markets led to sophisticated financial development. This situation has brought the concepts of financial development and growth to be discussed further (Hussain and Chakraborty, 2012).

The development of the financial system encourages "optimal capital allocation" as well as providing information on investments, which are considered as an important dynamic of growth (Guru and Yadav, 2019). Thus reduces the cost of information in the economy (Greenwood and Jovanovic, 1990). Levine (1997), and Guptha and Rao (2018) pointed out "production growth and capital accumulation" and "productivity increase" while drawing the theoretical framework between financial development and growth. Especially, Guptha and Rao (2018) stated in their studies that financial development leads to economic growth by mobilizing excess funds for the financing of investment projects. Secondly, innovation in financial technologies leads to efficient allocation of resources by reducing the asymmetric information.

Financial development makes a positive contribution to growth by affecting capital accumulation. This implies that the intersectoral specialization and thus structure of trade flows is determined by the relative level of financial intermediation. A well-developed financial sector affects growth through technological development channel. Thus increases the capacity of an economy to benefit from international trade to stimulate economic growth. However, international trade enables efficient allocation of internal and external resources. The shift of technological development to developing countries, thus less developed countries benefiting from the innovations of developed countries, contribute to economic growth through "learning by doing" (Shahbaz, 2012).

Most of the studies in the literature have analyzed the relationship between trade openness and growth. The relationship in question, which has an important place in the international economic literature, has been discussed with the hypotheses of "export-led growth", "import-led growth" or "trade-led growth". The validity of the hypothesis was investigated in various country groups.

Bencivega and Bruce (1991), Greenwood and Jovanic (1990) suggested that financial development wasis one of the major factors affecting economic growth in the long run because financial development

leads to capital accumulation, efficient allocation of resources and technological innovation. Along with these developments, economic growth is positively affected in the long run. Supply-pull and demand-pull hypotheses come to the fore in the analysis of the relationship between the financial system and growth. In the study of King and Levine (1993), the financial system is considered as the primary condition for growth. On the other hand, Aydın et al. (2013) state that an effectively functioning financial system can meet the need for funds, which play an important role in economic growth. In the light of all these evaluations, the theoretical expectation was also confirmed empirically in this study. A causal relationship from financial development to economic growth has been determined. In other words, financial development positively affects economic growth.

Studies on the relationship between financial development, trade openness and economic growth have been carried out by a wide audience over the years. The literature on the relationship between financial development and economic growth mostly supports a positive relationship between the two variables. However, there are differing views on the direction of the causal link between them. While some authors argue that the causality relationship runs from financial development to economic growth, others argue that it runs from economic growth to financial development. There are also few studies suggesting the existence of a bidirectional relationship between the variables.

Svaleryd and Vlachos (2002), Rajan and Zingales (2003), and Baltagi et al. (2009) argued in their study that commercial development was an important determinant for financial sector development. At this point, the direction of the relationship between financial development and trade openness is from trade openness to financial development. However, according to Beck (2003), economies increase their international trade volumes as they benefit from developments in the financial sector, technology and economies of scale. That is, the direction of the relationship is from financial development to trade openness. In our study, as emphasized in Beck's (2003) study, a causality from financial development to trade openness was determined. In addition, it has been revealed that both financial development and trade openness have a positive effect on economic growth.

Most panel and cross-country studies have found a positive relationship between financial development and economic growth when controlling for other growth determinants and also taking into account variable neglect bias, concurrency, and country-specific effects. These studies also support a causality running from financial development to economic growth. On the other hand, most of the time series studies have found both unidirectional and bidirectional causality between financial development and economic growth. Different results have also emerged when different proxy measures are used for financial development. However, the general literature supports the positive impact of financial development on long-term economic growth.

This study aims to analyze the relationship between trade deficit, financial development and growth by considering the subject from a different perspective and with up to date methods. Our study contributes to the existing literature by using the recently introduced Fourier-based cointegration (FSHIN) test developed by Tsong et al. (2016) and Fourier Toda-Yamamoto Causality Analysis proposed by Nazlioglu et al. (2016) which takes into account the structural changes in the model. The remainder of this article is organized as follows. The first section briefly explains the relevant literature. In the second section, the data set and the econometric methodology used are presented. The third section presents the empirical findings and the study is completed with the conclusion and recommendations section.

1 RELATED LITERATURE

Table 1 provides a brief summary of the studies on relationship between financial development, trade openness and economic growth.

Table 1 Literature review

Authors	Countries and time period	Methodology	Results
Omoke (2009)	Nigeria 1970–2005	Cointegration and Granger causality test	There is no cointegration relationship between financial development, trade openness and economic growth. Results shows that trade openness and financial development have a causal effect on economic growth.
Kenani and Fujio (2012)	Malawi 1970–2009	VECM and causality analysis	Trade openness affects economic growth and financial development indirectly affects economic growth in the short run.
Tash and Sheidaei (2012)	Iran 1966–2010	Johansen cointegration test	The joint impact of trade liberalization and financial development on economic growth is positive.
Arouri et al. (2013)	Bangladesh 1975Q1–2011Q4	ARDL bounds test, cointegration and causality	Series move together in the long run. Financial development causes economic growth. There is a feedback mechanism between trade openness and economic growth.
Lacheheb et al. (2013)	Algeria 1980–2010	ARDL bounds test, cointegration	There is a long-run relationship between trade openness, financial development and economic growth.
Menyah et al. (2014)	21 African countries 1965–2008	Panel causality test	The results show that recent attempts at financial development and trade liberalization have no significant impact on growth.
Zombe and Seshamani (2014)	Zambia 1965–2011	Cointegration VECM and causality	In the short run, it is concluded that economic growth and trade openness are the causes of financial development.
Kar et al. (2014)	Turkey 1989M1–2007M11	Linear and nonlinear causality	There is unidirectional causality between economic growth and trade openness. Economic growth leads to financial development. It has been found that financial development leads to trade openness.
Rehman et al. (2015)	Saudi Arabia 1971–2012	Cointegration and causality	Financial development, trade openness and economic growth move together in the long run. There is a one-way causality relationship from trade openness to economic growth and from economic growth to financial development.
Saaed and Hussain (2015)	Kuwait 1977–2012	VAR, cointegration and Granger causality test	According to Granger causality results based on VAR models, it was concluded that there is a causal relationship between economic growth and financial development, and between trade openness and economic growth.
Lawal et al. (2016)	Nigeria 1981–2013	ARDL bounds test	Economic growth financial development and trade openness level move together in the long run.
Ayad and Belmokaddem (2017)	16 MENA Countries 1980–2014	Panel cointegration, panel VAR model, Toda, Yamamoto, Dolado and Lutkepohl Granger causality tests	The results show that financial development and trade liberalization do not have a significant effect on economic growth.

Table 1 (continuation)

Authors	Countries and time period	Methodology	Results
Sönmez and Sağlam (2018)	Transition economies 2001–2014	Principal component analysis, panel cointegration and causality tests	An economic growth based on financial development and trade openness is realized.
Xie et al. (2018)	China 1978–2015	Bootstrap ARDL and causality analysis	There is a unidirectional causality between trade openness and economic growth, and between trade openness and financial development.
Atgür (2019)	Turkey 2004–2017	Cointegration and causality	It was concluded that there is no long-term relationship between financial development and trade openness levels and economic growth. In addition, a unidirectional causality relationship from trade openness to economic growth has been determined.

Note: ARDL, VECM and VAR, respectively, refer to autoregressive distributed lag model, vector error correction model and vector autoregressive model.
Source: Own construction

2 DATA SET AND ECONOMETRIC METHOD

In this study, the effects of financial development (FD) and trade openness (TO) on economic growth (GDP) are investigated using annual time series data for Turkey in the period 1960–2017. To examine the long-run relationship between financial development, trade openness and economic growth; Fourier-based stationarity test and its complementary Fourier-based cointegration test are used. Finally, Fourier-based causality tests are also used to examine the causality relationship between the variables. The investigated model is as follows:

$$GDP_t = \beta_0 + \beta_1 FD + \beta_2 TO + \varepsilon_t. \tag{1}$$

The variables used in Formula (1) were obtained from the official database of the World Bank. The financial development variable is represented as the ratio of financial system deposits to gross domestic product (as %). The trade openness variable is represented as the ratio of the sum of exports and imports of goods and services to gross domestic product. And per capita gross domestic product (GDP, constant 2010 US$) is used to represent the economic growth variable.

Table 2 Definition of variables and acronyms

Data			Source
Per capita gross domestic product	Gross domestic product	GDP	World Bank
The ratio of financial system deposits to gross domestic product	Financial development	FD	World Bank
The ratio of the sum of exports and imports of goods and services to gross domestic product	Trade openness	TO	World Bank

Source: Own construction

Descriptive statistics of the variables used in the study are presented in Table 3.

Table 3 Descriptive statistics of variables

	FD	TO	GDP
Mean	24.093	31.344	7 119.606
Median	21.119	33.178	6 389.336
Maximum	46.335	55.762	14 975.090
Minimum	8.679	5.727	3 134.577
Standard deviation	11.460	16.5241	3 113.956
Skewness	0.649	–0.090	0.842
Kurtosis	2.351	1.472	2.826
Jargue – Bera	5.096 (0.078)	5.723 (0.057)	6.931 (0.031)

Note: Values in parentheses indicate probability values.
Source: Own construction

According to the Jarque-Bera normality test results, it can be seen that variables considered in the model do not exhibit normal distribution. In addition, it is found that the variables have a kurtosis below the normal.

2.1 Data set and model analysis

The subject of structural break was first introduced to the literature by Perron (1989) and it was stated that ignoring these sudden changes in the series could lead to false and misleading results. However, with the developing literature, it has been emphasized that the change in the series may not be sudden but soft, and many Fourier-based tests have been proposed in order to catch these soft changes (Enders and Lee, 2004; Becker et al., 2004; Becker et al., 2006; Christopoulos and Leon-Ledesma, 2010, 2011; Enders and Lee, 2012; Omay, 2015; Bozoklu et al., 2020). Sometimes some tests lose their validity in cases where the structure of the breaks is not sharp and smooth transitions are experienced. For these cases, nonlinear and smooth transition unit root tests have been developed. In unit root tests where the breaks are sharp or the break structures are determined by nonlinear models, the number of breaks and the structure of the nonlinearity are determined beforehand. However, in cases where the structure and number of breaks cannot be determined beforehand, both test groups cannot provide sufficient power for stability tests. Incorrect determination and modeling of the number and location of the breaks present a problem just like the neglect of the fractures. In this context, unit root tests based on frequency component selection have been developed by using the Fourier function approach, where there is no requirement to predetermine the refraction numbers and structures (Chi-Wei, 2012: 22). Becker et al. (2006), using Fourier functions, extended the KPSS stationarity test developed by Kwiatkowski et al. (1994). In this way, the situation where the number of structural breaks in the functional form is not known is allowed. These Fourier functions are intended to capture a large number of smooth changes whose number, position and shape have no effect on the strength of the test. Data creation process for Becker et al. (2006) Fourier KPSS (FKPSS) test is as follows:

$$y_t = X_t^{'}\beta + Z_t^{'}\gamma + r_t + \varepsilon_t \; r_t = r_{t-1} + u_t \; . \tag{2}$$

Here ε_t shows stationary errors and u_t shows the error process for the independent identical distribution (iid) with variance σ_u^2 . $Z_t = \left[\sin\left(2\pi kt/T\right), \cos\left(2\pi kt/T\right)\right]'$ is defined like this. T represents sample

size considered and k represents the number of frequency. To investigate whether the y_t series is level or trend stationary, respectively $X_t = [1]$ ve $X_t = [1, t]'$ determined. In this test based on KPSS null hypothesis expresses the stationarity. In order to calculate the test statistic for the constant or with trend model under this null hypothesis assumption, the following models are estimated at first and the residuals are obtained:

$$y_t = \alpha_0 + \gamma_{1k} \sin\left(\frac{2\pi kt}{T}\right) + \gamma_{2k} \cos\left(\frac{2\pi kt}{T}\right) + e_t, \tag{3}$$

$$y_t = \alpha_0 + \beta t + \gamma_{1k} \sin\left(\frac{2\pi kt}{T}\right) + \gamma_{2k} \cos\left(\frac{2\pi kt}{T}\right) + e_t. \tag{4}$$

Respectively for constant, constant and trend models are determined as $\tau_\mu(k)$ and $\tau_\tau(k)$. In order to determine the optimal frequency number, frequency values from 1 to 5 are tried for Formulas (3) and (4). The value at which the sum squares residual is minimum is selected as the appropriate frequency value.

The test statistic for the two models is calculated in the same way and is expressed as:

$$\tau(k) = \frac{1}{T^2} \frac{\sum_{t=1}^{T} \tilde{S}_t(k)^2}{\tilde{\sigma}^2}. \tag{5}$$

Here $\tilde{e}_j$ represents residuals from constant and constant and trend models and determined as $\tilde{S}_t(k) = \Sigma_{j=1}^{t} \tilde{e}_j$. Before proceeding to the stationarity testing phase, the significance of the Fourier functions included in the model is tested. To use here, Becker et al. (2006) obtained the F statistic for the significance test of the terms as follows:

$$F_i(k) = \frac{\left(KKT_0 - KKT_1(k)\right)/2}{KKT_1(k)/(T-q)}, i = \mu, \tau \tag{6}$$

Here k is the number of frequencies and q is the number of independent variables. In Formulas (3) and (4), KKT_0 without trigonometric terms and $KKT_1(k)$ are calculated by considering trigonometric terms. The F test can only be used when the stationarity basic hypothesis cannot be rejected. It is important that the coefficients included in the model are statistically significant. Otherwise, it is recommended to use the standard KPSS test (Yılancı, 2017: 57).

2.2 Fourier Shin cointegration analysis[4]

The concept of cointegration, which was first proposed by Engle-Granger (1987), has shown a rapid development like unit root tests and many tests have been added to the literature. One of them is the cointegration test proposed by Shin (1994), which is an improved version of the KPSS stationarity test to the cointegration form. In the following process, similar to the unit root test literature, the importance of considering structural changes in long-term relationships for cointegration tests has been mentioned. Therefore, tests taking into account the structural changes are proposed for the long-term relationship between the series (Gregory and Hansen, 1999; Johansen et al., 2000; Arai and Kurozumi, 2007; Hatemi-J, 2008). One of these tests is the Fourier-based cointegration (FSHIN) test developed by Tsong et al. (2016). The feature that distinguishes this test from other tests is that, as in Fourier structures, the number

[4] Tsong et al. (2016).

and form of structural changes are not determined a priori. Therefore, strong results are obtained (Yılancı, 2017: 58). The null hypothesis of this extended version of Becker et al. (2006) FKPSS test to cointegration form is the existence of cointegration. The FSHIN test procedure is defined as follows:

$$y_t = d_t + x_t^{'}\beta + \eta_t,\ \eta_t = \gamma_t + \upsilon_{1t},\ \gamma_t = \gamma_{t-1} + u_t,\ x_t = x_{t-1} + \upsilon_{2t}. \tag{7}$$

Here u_t~iid$(0, \sigma_u^2)$ and y_t has a random walk process. Therefore as v_{1t} and v_{2t} exhibit a stationary process y_t and x_t are stationary at the first difference. d_t in Formula (7) is for both constant and constant and trend model respectively defined as $d_t = \delta_0 + f_t$ and $d_t = \delta_0 + \delta_1 t + f_t$, and $f_t = \alpha_k \sin\left(\frac{2\pi kt}{T}\right) + \beta_k \cos\left(\frac{2\pi kt}{T}\right)$ represents the Fourier function. Here k, t and T represent number of frequency, trend, and sample size, respectively. From here, the following equation is obtained:

$$y_t = \delta_0 + \alpha_k \sin\left(\frac{2\pi kt}{T}\right) + \beta_k \cos\left(\frac{2\pi kt}{T}\right) + x_t^{'}\beta + \upsilon_{1t}. \tag{8}$$

The following test statistic is used to test the null hypothesis:

$$CI_f^m = \frac{1}{T^2}\hat{\omega}_1^{-2}\sum_{t=1}^{T} S_t^2. \tag{9}$$

Here $\hat{\omega}_1^2$ represents a consistent estimator of long-run variance of v_{1t} in Formula (7) and S_t represents the partial sum of the least squares residues obtained from Formula (8).

2.3 Fourier Toda-Yamamoto causality test

After the determination of the cointegration relationship between the variables, a causality test based on the Vector Autoregressive (VAR) model introduced by Toda and Yamamoto (1995) has been proposed. The VAR structure proposed by Sims (1980) has been proposed as an alternative to large-scale structural models. VAR removes the constraints arising from economic theory in structural models and provides convenience for multivariate analysis. Thanks to this advantage, the VAR model is thought to be more useful than univariate models. The VAR model is defined as a system of equations in which the lagged values of each internal variable and other variables are take place on the right side of the equation and shown as follows:

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_{p+d} y_{t-(p+d)} + \varepsilon_t. \tag{10}$$

In the causality test of Toda and Yamamoto (1995), after obtaining the highest order of stationarity and optimal lag length, which are indicated by d_{max} and p, respectively, a VAR model is obtained at the level of $(d_{max} + p)$. Toda and Yamamoto (1995) performed the causality test with the help of Wald test statistics. If the obtained test statistics value is greater than the critical value, the null hypothesis stating that there is no causality is rejected.

Nazlioglu et al. (2016), on the other hand, proposed a new test by incorporating Fourier terms into the model, taking into account the structural breaks. They included structural changes in the familiar VAR model and extended the constant term assumption. In other words, instead of the constant term in the VAR model, Fourier terms are added to capture the changes that may occur in the dependent variable. Instead of the constant term in Formula (10), Fourier terms are added as in Formula (8) and it is represented as follows:

$$y_t = \alpha_0 + \gamma_{1k} \sin\left(\frac{2\pi kt}{T}\right) + \gamma_{2k} \cos\left(\frac{2\pi kt}{T}\right) + \beta_1 y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_{p+d} Y_{t-(p+d)} + \varepsilon_t. \quad (11)$$

Here k represents the frequency number. Thanks to these added terms, possible structural breaks are captured with sinus and cosinus waves, without knowing the breaking time, the number of breaks, and the way of breaking. Nazlioglu et al. (2016) suggested that the F test statistic should be used instead of the Wald test, since the χ^2 distribution is weak in causality tests in terms of small sample features. After determining the optimal lag and frequency of the Fourier terms, the test is performed and the null hypothesis that there is no causality is tested.

3 EMPRICAL FINDINGS

In this study investigating the effects of financial development and trade openness on economic growth for Turkey, Becker et al. (2006) Fourier KPSS stationarity test results and Tsong et al. (2016) Fourier Shin Cointegration test results are presented in the tables below. In the continuation, KPSS stationarity test results and Shin (1994) cointegration test results, which form the basis of these tests, respectively, are also reported. Lastly, to these, the results obtained as a result of the Fourier Toda-Yamamoto causality analysis proposed by Nazlioglu et al. (2016) and taking into account the structural changes are also presented.

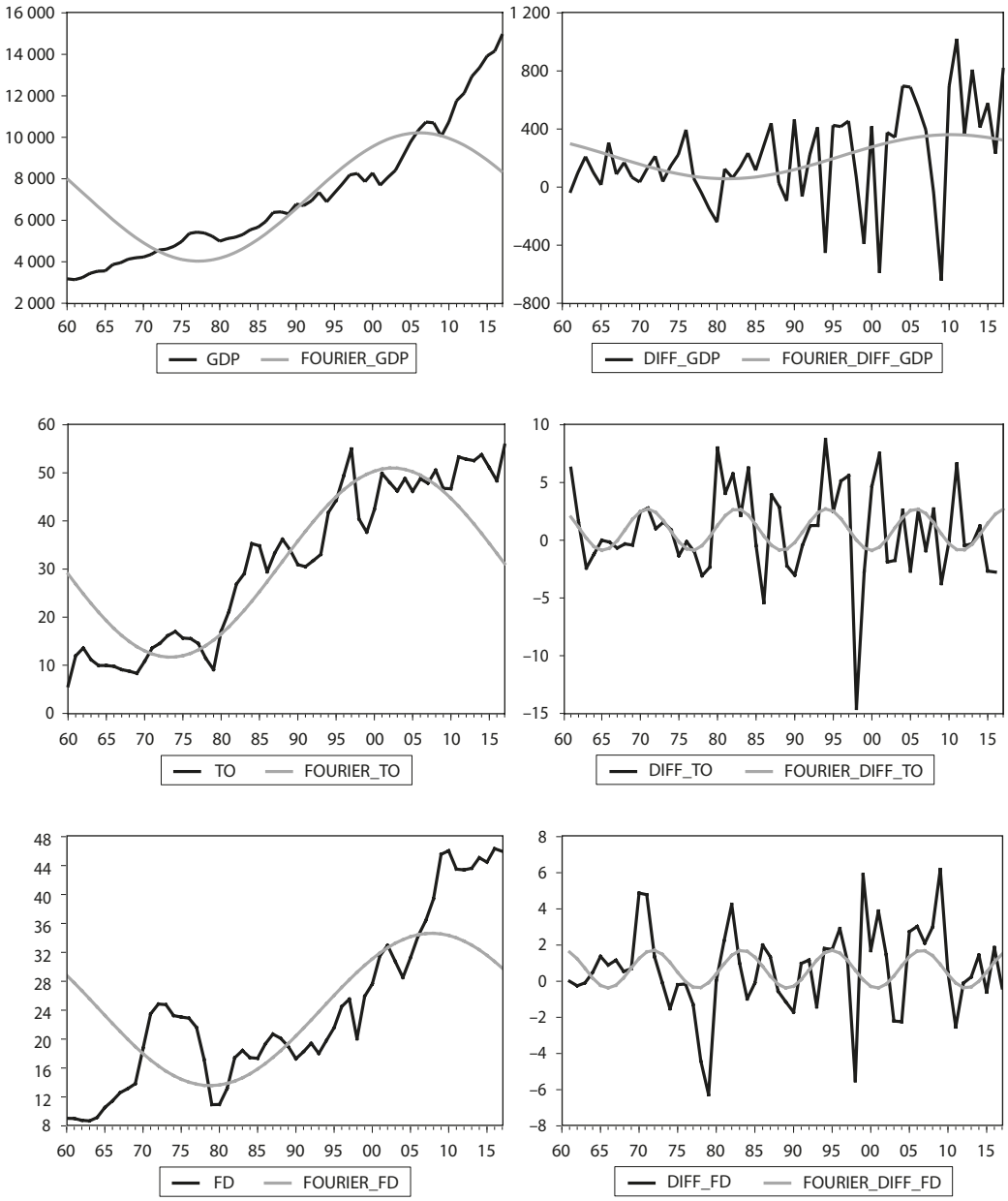
Table 4 KPSS and Fourier KPSS unit root test results

	KPSS test stat.	Frequency	Min. SSR	Fourier KPSS test stat.	Critical values			F test stat.
					%1	%5	%10	
ΔGDP	0.895	1	275 000	0.457	0.269	0.172	0.131	27.783***
GDP	0.625*	1	5 383 768	0.184*	0.269	0.172	0.131	3.309
ΔTO	0.900	1	4 336.839	0.433	0.269	0.172	0.131	71.151***
TO	0.103***	5	782.237	0.264***	0.738	0.462	0.351	3.128
ΔFD	0.778	1	4 266.452	0.405	0.269	0.172	0.131	20.752***
FD	0.112***	5	293.963	0.139***	0.738	0.462	0.351	2.802

Note: *, ** and *** shows respectively %10, %5 and %1 significance level. KPSS test and the critical values required for the F test, which is used to test the significance of trigonometric terms in levels %1, %5 and %10 respectively 0.739, 0.463, 0.347 and 6.730, 4.929, 4.133.
Source: Own construction

According to the Fourier KPSS stationarity test results, it is seen that the GDP, TO and FD variables are not stationary at the level, but become stationary after taking their first difference. Therefore, it is concluded that all three series are I(1). Since the significance test of trigonometric terms was used only when the null hypothesis was not rejected, the F test was performed again for three variables whose difference was taken, and it was found that trigonometric terms were not significant in these three variables. For this purpose, KPSS test was applied for difference series and it was concluded that both series were I(1) according to both Fourier KPSS and traditional KPSS test results. In addition, the time path graph of the Fourier estimates of the variables is presented in Figure 1.

Figure 1 Time paths of series with fourier approximations



Source: Own construction

According to the time path graphs of the Fourier predictions obtained from Figure 1, it is seen that the appropriate Fourier predictions are realized and long-term oscillations can be captured.

The findings of the tests carried out to test the long-term cointegration relationship are reported in Table 5.

Table 5 Shin and Fourier Shin cointegration test results

	F test atat.	Frequency	Min. SSR	Test statistic	Critical values		
					%1	%5	%10
FSHIN cointegration test	11.572***	2	41 959	0.108***	0.132	0.182	0.328
Shin cointegration test	–	–	–	0.115***	0.163	0.221	0.380

Note: *** shows significance at %1 level. The critical values required for the F test, which is used to test the significance of trigonometric terms in levels %1, %5 and %10 respectively 5.774, 4.066, and 3.352.
Source: Own construction

As a result of both FSHIN and Shin cointegration tests, it is found that there is a long-term cointegration relationship between economic growth, trade openness and financial development. This result shows that trade openness and financial development move together with economic growth in the long run. It is also seen that the F statistic is significant for the trigonometry terms for the FSHIN test. The coefficient estimates of the long-term relationship determined between the variables were investigated with the Dynamic Least Squares (DOLS) method. It is stated that this technique proposed by Stock and Watson (1993) produces strong and consistent predictions even in the presence of endogenity and autocorrelation problems in explanatory variables. In order to overcome the internality problem, in addition to the level values of the explanatory variables, the lag of the first differences (lag) and the antecedents (lead) should be included in the model. In addition, to overcome autocorrelation problem Generalized OLS method should be used. The findings of the DOLS method are presented i n Table 6.

Table 6 DOLS Long-Run coefficient estimator results

	Coefficient	Standard error	Statistic value
TO	71.327	13.536	5.269 (0.000)***
FD	167.547	20.946	7.998 (0.000)***
SIN	–156.702	198.083	2.671 (0.432)
COS	358.672	182.992	1.960 (0.055)*
C	847.167	317.054	–0.791 (0.010)**

Note: *, ** and *** respectively shows significance at the level %10, %5 and %1.
Source: Own construction

It is seen that both trade openness and financial development series are statistically significant and have a positive effect on growth. In addition, it was found that the cosinus term among the Fourier functions included in the model was statistically significant, and the sinus term was not statistically significant.

According to the Fourier Toda-Yamamoto causality analysis results, it is seen that there is a one-way causality relationship from financial development to economic growth and from financial development to trade openness. For the other variables, no causality is determined according to the test result.

Table 7 Fourier Toda-Yamamoto causality test results

H_0 null hypothesis	Optimal lags	Optimal frequency	Wald stat.	Asymptotic p-value	Bootstrap-value
FD does not cause *GDP*	1	3	4.311	0.038**	0.041**
GDP does not cause *FD*	1	3	0.175	0.675	0.684
TO does not cause *GDP*	1	1	1.918	0.166	0.172
GDP does not cause *TO*	1	1	2.286	0.131	0.131
FD does not cause *TO*	1	1	2.929	0.087*	0.094*
TO does not cause *FD*	1	1	0.114	0.736	0.740

Note: ** and * indicate 5% and 1% significance level, respectively. Analyzes were performed with 1 000 bootstrap simulations.
Source: Own construction

CONCLUSION

In the economic literature, financial development and trade liberalization are identified as key factors supporting economic growth in general. In theory, the financial system mediates the allocation of financial resources, and financial development increases both the size and efficiency of the allocation of resources. Especially for developing countries, economic growth occurs when a country has an efficient financial system. An advanced financial system encourages investments, funds business opportunities, mobilizes savings, and manages risks. All these functions stimulate the economy and thus support its growth.

In this study, the effects of financial development and trade openness on economic growth were tested with annual data covering the period 1960–2017 for Turkey. For this purpose to measure the degree of integration of the variables; KPSS stationarity test, which is one of the traditional tests, and Becker et al. (2006), the Fourier KPSS stationarity test, which is an extended version of the KPSS test with Fourier functions, are applied. As a result of these two stationarity analysis, it is concluded that three variables became stationary after taking their first difference, that is, I(1). In order to measure the long-term cointegration relationship between the variables considered, Shin and Fourier Shin cointegration tests, which are accepted as the continuation of these stationarity tests in the literature, were carried out. According to the cointegration test results, it is seen that there is a long-term relationship between the variables. In addition, as a result of the long-term coefficient estimation, it is concluded that the coefficients of the trade openness and financial development series, which are taken as independent variables, are significant and positive. In addition, it is seen that the cosinus term, which is one of the Fourier functions included in the model, is also significiant. Finally, according to the results of the Fourier-based causality analysis, a one-way causal relationship from financial development to economic growth and from financial development to trade openness.

According to the results of the analysis, financial development and trade openness had a positive effect on growth. The empirical results of the study are consistent with the studies of Shahbaz (2012) and Alsamara et al. (2019). The positive functioning of the financial markets in Turkey and the increase in the level of trade openness have great importance in obtaining these results. It is important to continue the positive economic transformation, especially with the progress in the post-1980 liberalization process. At this point, some transformations should be implemented at both national and international level.

Structural reforms should be accelerated in order to keep the competition dynamism alive in the country and to be ready for competing with foreign countries. International integration should be achieved with broader participation and multilateral trade agreements.

With the acceleration of the globalization phenomenon, multinational companies (MNCs) have started to take part actively in international retail chains. While this process makes easier accessing to products at more affordable prices for individuals, it has made companies more open to competition. At this point, it has become essential for countries to allocate more resources to R&D and innovation. However, in this way, the domestic market becomes ready for foreign competition. Considering these aspects in terms of trade policy, policymakers in Turkey should focus more on export policies.

In developing countries such as Turkey the existence of a strong financial structure that can quickly adapt to international financial conditions is essential in order to avoid the risk of increased capital flows arising from trade openness. Considering the challenging structure in global competitive conditions, priority should be given to policies regarding the efficiency of the financial system. Because, for the continuity of the positive effect of financial development on growth, it is important to provide financial deepening. At this point, it is necessary to reduce financial fragility and diversify financial instruments. Similarly, priority should be given to long-term policies for the effective and efficient allocation of resources.

In recent years, when the policies of foreign expansion became obvious and capital movements accelerated around the World, financially successful openness policies should be maintained in developing countries such as Turkey. In particular, the speculative effects of capital movements should be minimized and the amount of foreign borrowing should be reduced.

References

ALSAMARA, M., MRABET, Z., BARKAT, K., ELAFIF, M. (2019). The impacts of trade and financial developments on economic growth in Turkey: ARDL approach with structural break [online]. *Emerging Markets Finance and Trade*, 55(8): 1671–1680. <https://doi.org/10.1080/1540496X.2018.1521800>.

ARAI, Y., KUROZUMI, E. (2007). Testing for the Null Hypothesis of Cointegration with a Structural Break [online]. *Econometric Reviews*, 26(6): 705–739. <https://doi.org/10.1080/07474930701653776>.

AROURI, M., UDDIN, G. S., NAWAZ, K., SHAHBAZ, M., TEULON, F. (2013). Causal linkages between financial development, trade openness and economic growth: Fresh evidence from innovative accounting approach in case of Bangladesh. *Ipag Business School Working Paper*, 2013-037.

ATGÜR, M. (2019). Finansal Gelişme, Ticari Açıklık ve Ekonomik Büyüme İlişkisi: Türkiye Örneği. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 33(2): 553–572.

AYAD, H., BELMOKADDEM, M. (2017). Financial development, trade openness and economic growth in MENA countries: TYDL panel causality approach. *Theoretical and Applied Economics*, 24(1): 610.

AYDIN, M. K., AK, M. Z., ALTUNTAŞ, N. (2013). Çevre ülkelerinde finansal gelişme ile büyüme arasındaki ilişki: Panel Veri Analizi [online]. *H. Ü. İktisadi ve İdari Bilimler Fakültesi Dergisi*, 31(2): 1–14. <https://doi.org/10.17065/huniibf.103641>.

BALTAGİ, B., DEMETRİADES, P., LAW, S. H. (2009). Financial development, openness, and ınstitutions: evidence from Panel Data. *Journal of Development Economics*, 89(2): 285–296.

BECK, T. (2003). Financial dependence and international trade [online]. *Review of International Economics*, 11(2): 296–316. <https://doi.org/10.1111/1467-9396.00384>.

BECKER, R., ENDERS, W., HURN, S. (2004). A general test for time dependence in parameters [online]. *Journal of Applied Econometrics*, 19: 899–906. <https://doi.org/10.1002/jae.751>.

BECKER, R., ENDERS, W., LEE, J. (2006). A Stationary Test in the Presence of an Unknown Number of Smooth Breaks [online]. *Journal of Time Series Analysis*, 27(3): 381–409. <https://doi.org/10.1111/j.1467-9892.2006.00478.x>.

BENCİVENGA, V. R., BRUCE, D. S. (1991). Financial İntermediation and endogenous 24 growth [online]. *The Review of Economic Studies*, 58(2): 195–209. <https://doi.org/10.2307/2297964>.

BOZOKLU, S., YILANCI, V., GORUS, M. S. (2020). Persistence in per capita energy consumption: A fractional integration approach with a Fourier function [online]. *Energy Economics*, 91: 104926. <https://doi.org/10.1016/j.eneco.2020.104926>.

CHAOUACHI, M., CHAOUACHI, S. (2021). Financial Development and Poverty Reduction in Crisis Periods: Panel Data Evidence from Six Countries of ECOWAS [online]. *Statistika: Statistics and Economy Journal*, 101(2): 187–202. <https://www.czso.cz/documents/10180/143550797/32019721q2_chaouachi.pdf/3791b08a-a288-4a12-9d5d-7b27c8739ef5?version=1.1>.

CHI-WEI, S. (2012). Flexible fourier stationary test in purchasing power parity for central and eastern European countries. *Ekonomický časopis*, 60(1): 19–31.
CHRISTOPOULOS, D. K., LEÓN-LEDESMA, M. A. (2010). Smooth breaks and non-linear mean reversion: Post-Bretton Woods real exchange rates [online]. *Journal of International Money and Finance*, 29(6): 1076–1093. <https://doi.org/10.1016/j.jimonfin.2010.02.003>.
CHRISTOPOULOS, D. K., LEON-LEDESMA, M. A. (2011). International output convergence, breaks, and asymmetric adjustment [online]. *Studies in Nonlinear Dynamics & Econometrics*, 15(3). <https://doi.org/10.2202/1558-3708.1823>.
ENDERS, W., LEE, J. (2004). Testing for a unit root with a nonlinear Fourier function. *Econometric Society 2004 Far Eastern Meetings*, Vol. 457.
ENDERS, W., LEE, J. (2012). A unit root test using a Fourier series to approximate smooth breaks [online]. *Oxford Bulletin of Economics and Statistics*, 74(4): 574–599. <https://doi.org/10.1111/j.1468-0084.2011.00662.x>.
ENGLE, R. F., GRANGER, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing [online]. *Econometrica*, 55(2): 251–276. <https://doi.org/10.2307/1913236>.
GOLDSMITH, R. W. (1969). *Financial Structure and Development*. New Haven, CT: Yale University Press.
GREENWOOD, J., JOVANOVIC, B. (1990). Financial development, growth, and the distribution of income. *The Journal of Political Economy*, 98(5): 1076–1107.
GREGORY, A. W., HANSEN, B. H. (1996). Residual-based tests for cointegration in models with regime shifts. [online]. *Journal of Econometrics,* 70(1): 99–126. <https://doi.org/10.1016/0304-4076(69)41685-7>.
GUPTHA, K. S. K., RAO, R. P. (2018). The causal relationship between financial development and economic growth: An experience with BRICS economies [online]. *Journal of Social and Economic Development*, 20(2): 308–326. <https://doi.org/10.1007/s40847-018-0071-5>.
GURU, B. K., YADAV, I. S. (2019). Financial development and economic growth: Panel evidence from BRICS [online]. *Journal of Economics, Finance and Administrative Science*, 24(47): 113–126. <https://doi.org/10.1108/JEFAS-12-2017-0125>.
HATEMI-J, A. (2008). Tests for cointegration with two unknown regime shifts with an application to financial market integration [online]. *Empirical Economics*, 35(3): 497–505. <https://doi.org/10.1007/s00181-007-0175-9>.
HUSSAIN, F., CHAKRABORTY, D. K. (2012). Causality between financial development and economic growth: Evidence from an Indian state. *Romanian Economic Journal*, 15(35): 27–48.
JOHANSEN, S., MOSCONI, R., NIELSEN, B. (2000). Cointegration analysis in the presence of structural breaks in the deterministic trend [online]. *Econometrics Journal*, 3: 216–249. <https://doi.org/10.1111/1368-423X.00047>.
KAR, M., NAZLIOĞLU, Ş., AĞIR, H. (2014). Trade openness, financial development and economic growth in Turkey: Linear and nonlinear causality analysis. *BDDK Bankacılık ve Finansal Piyasalar Dergisi*, 8(1): 63–86.
KENANI, J. M. VE FUJIO, M. (2012). A Dynamic Causal Linkage Between Financial Development, Trade Openness And Economic Growth: Evidence From Malawi. *Interdisciplinary Journal of Contemporary Research in Business,* 4(5): 569–583.
KİNG, R. G., LEVİNE, R. (1993). Finance, entrepreneurship and growth [online]. *Journal of Monetary Economics*, 32(3): 513–542. <https://doi.org/10.1016/0304-3932(93)90028-E>.
KWIATKOWSKI, D., PHILLIPS, P. C. B., SCHMIDT, P., SHIN, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root [online]. *Journal of Econometrics*, 54: 159–178. <https://doi.org/10.1016/0304-4076(92)90104-Y>.
LACHEHEB, M., ADAMU, P., AKUTSON, S. (2013). Openness, financial development and economic growth in Algeria: an ARDL bound testing approach [online]. *International Journal of Economics, Finance and Management Sciences*, 1(6): 400–405. <https://doi.org/10.11648/j.ijefm.20130106.28>.
LAWAL, A. I., NWANJI, T. I., ASALEYE, A., AHMED, V. (2016). Economic growth, financial development and trade openness in Nigeria: An application of the ARDL bound testing approach [online]. *Cogent Economics & Finance*, 4(1): 1258810. <https://doi.org/10.1080/23322039.2016.1258810>.
LEVINE, R. (1997). Financial development and economic growth: views and agenda. *Journal of Economic Literature*, 35: 688–726.
MCKINNON, R. I. (1973). *Money and capital in economic development*. Washington, DC: Brookings Institution, 1973.
MENYAH, K., NAZLIOĞLU, S., WOLDE-RUFAEL, Y. (2014). Financial development, trade openness and economic growth in African countries: New insights from a panel causality approach [online]. *Economic Modelling*, 37: 386–394. <https://doi.org/10.1016/j.econmod.2013.11.044>.
NAZLIOĞLU, S., GORMUS A., SOYTAS, U. (2016). Oil Prices and Real Estate Investment Trusts (Reits): Gradual-Shift Causality and Volatility Transmission Analysis [online]. *Energy Economics*, 60: 168–175. <https://doi.org/10.1016/j.eneco.2016.09.009>.
OMAY, T. (2015). Fractional frequency flexible Fourier form to approximate smooth breaks in unit root testing [online]. *Econ. Lett*., 134: 123–126. <https://doi.org/10.1016/j.econlet.2015.07.010>.
OMOKE, P. C. (2010). The Causal Relationship among Financial Development, Trade Openness and Economic Growth in Nigeria. *Trade Openness and Economic Growth in Nigeria,* 2(2): 137–147.
PERRON, P. (1989). The great crash, the oil price shock, and the unit root hypothesis [online]. *Econometrica: Journal of the Econometric Society*, 57(6): 1361–1401. <https://doi.org/10.2307/1913712>.

RAJAN, R. G., ZİNGALES, L. (2003). The great reversals: the politics of financial development in the twentieth century [online]. *Journal of Financial Economics*, 69: 5–50. <https://doi.org/10.1016/S0304-405X(03)00125-9>.

REHMAN, M. I., ALI, N., NASIR, N. M. (2015). Linkage between financial development, trade openness and economic growth: Evidence from Saudi Arabia. *Journal of Applied Finance and Banking*, 5(6): 127–141.

SAAED, A. J., HUSSAIN, M. A. (2015). The causal relationship among trade openness, financial development and economic growth: Evidence from Kuwait. *Journal of Emerging issues in Economics, Finance and Banking*, 4(1): 1385–1409.

SALAHUDDIN, M., GOW, J. (2016). The effects of Internet usage, financial development and trade openness on economic growth in South Africa: a time series analysis [online]. *Telematics and Informatics*, 33(4): 1141–1154. <https://doi.org/10.1016/j.tele.2015.11.006>.

SHAHBAZ, M. (2012). Does trade openness affect long run growth? Cointegration, causality and forecast error variance decomposition tests for Pakistan [online]. *Economic Modelling*, 29: 2325–39. <https://doi.org/10.1016/j.econmod.2012.07.015>.

SHAW, E. S. (1973). *Financial deepening in economic development*. New York: Oxford Universtty Press.

SHIN, Y. (1994). A residual-based test of the null of cointegration against the alternative of no cointegration [online]. *Econometrics Theory*, 10(1): 91–115. <https://doi.org/10.1017/S0266466600008240>.

SIMS, C. A. (1980). Macroeconomics and reality [online]. *Econometrica*, 48(1): 1–48. <https://doi.org/10.2307/1912017>.

SÖNMEZ, F. E., SAĞLAM, Y. (2018). Finansal Gelişme ve Ticari Açıklık ile Ekonomik Büyüme Arasındaki İlişki: Avrupa Dönüşüm Ekonomileri Örneği [online]. *Anadolu Üniversitesi Sosyal Bilimler Dergisi*, 18(4): 59–72. <https://doi.org/10.18037/ausbd.552681>.

STOCK, J., WATSON, M. W. (1993). A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems [online]. *Econometrica*, 61(4): 783–820. <https://doi.org/10.2307/2951763>.

SVALERYD, H., VLACHOS, J. (2002). Markets for risk and openness to trade: How are they related? [online]. *Journal of International Economics*, 57: 369–395. <https://doi.org/10.1016/S0022-1996(01)00153-2>.

TADESSE, T., ABAFIA, J. (2019). The causality between financial development and economic growth in Ethiopia: Supply leading vs demand following hypothesis [online]. *Journal of Economics and Financial Analysis*, 3(1): 87–115. <https://doi.org/10.1991/JEFA.V3I1.A25>.

TASH, M. N. S., SHEIDAEI, Z. (2012). Trade liberalization, financial development and economic growth in the long term: The case of Iran [online]. *Business and Economic Horizons*, 8(2): 33–45. <https://doi.org/10.15208/beh.2012.9>.

TODA, H. Y., YAMAMOTO, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes [online]. *Journal of Econometrics*, 66: 225–250. <https://doi.org/10.1016/0304-4076(94)01616-8>.

TSONG, C. C., LEE, C. F., TSAI, L. J., HU, T. C. (2016). The Fourier approximation and testing for the null of cointegration [online]. *Empirical Economics*, 51(3): 1085–1113. <https://doi.org/10.1007/s00181-015-1028-6>.

XIE, H., CAI, Y., SAM, C. Y., CHANG, T. (2018). Revisit Financial Development, Trade Openness And Economic Growth Nexus In China Using a New Developed Bootstrap ARDL Test [online]. *Economic Computation & Economic Cybernetics Studies & Research*, 52(4). <https://doi.org/10.24818/18423264/52.4.18.09>.

YILANCI, V. (2017). Analysing the relationship between oil prices and economic growth: A fourier approach. *Ekonometri ve İstatistik e-Dergisi*, 27: 51–67.

ZOMBE, C., SESHAMANI, V. (2014). Financial Development, Trade Openness, and Economic Growth in Zambia. *Journal of Modern Accounting and Auditing*, 10(7): 803–815.

Is Gender Earnings Gap a Reality? Signals from Indian Labour Market

Sonu Madan[1] | *Indira Gandhi University, Meerpur, India*
Surender Mor[2] | *BPS Women University, Khanpur Kalan, India*

Received 22.7.2021 (revision received 22.3.2022), **Accepted** (reviewed) 27.4.2022, **Published** 17.6.2022

Abstract

We examined the persistence of the gender earnings gap across diverged occupational groups and the workers owning diverged work status in India using the relevant information on 94 446 workers from the Periodic Labour Force Survey (2017–18). The marginal mean earning of workers is estimated using GLM: ANCOVA. The findings report the persistence of significant gender earnings gap across the occupational structure and work status of workers. The elimination of demotivating factors leading to the gender earnings gap, removal of gender discrimination, enhancing the self-esteem of females, raising productivity potential by augmenting the professional/vocational education and policies for increased female work participation is the need of the hour.

Keywords	***DOI***	***JEL code***
Educational attainments, gender earnings gap, GLM: ANCOVA, occupational groups, work status	*https://doi.org/10.54694/stat.2021.19*	*J16, J24, J31*

INTRODUCTION

Earning from work is a key indicator of a nation's economic well-being and has remained a challenge towards attaining decent working conditions and inclusive growth in India (Madan and Goel, 2019). The persistence of the gender earnings gap from work has been a common feature of the Indian labour market (Das, 2012). The gender earnings gap occurs when workers with the same educational attainment, expertise and work experience earn differently because of their gender, irrespective of their socio-economic characteristics (Poddar and Mukhopadhyay, 2019). Workers of different offspring are paid differently even within the same occupational groups despite possessing similar work profiles and skill levels. Female participation in the labour market is a gauge of productivity potential and growth of a nation and an informative indicator of the progress and status of females in society (Nazier, 2017). The current estimates drew attention towards

[1] Department of Economics, Indira Gandhi University, Meerpur, Rewari, Haryana 122 502, India. E-mail: sonu.economics@igu.ac.in, phone: (+91)8685911117.
[2] Department of Economics, BPS Women University, Khanpur Kalan, Sonipat, Haryana 131 305, India. E-mail: surendermor71@gmail.com, phone: (+91)9729185100.

54 percent of India's population in the working-age group (15–59 years), wherein females account for a significant proportion, i.e., 25 percent, signifying their relevance in the labour market (Agarwal, 2017).

Moreover, earning gap from work has been one of the key reasons for labour mobility across economic sectors and regions (Weeden, 1998; Stephen, 1998; Weichselbaumer and Winter, 2005; Livanos and Pouliakas, 2012). The gender earnings gap, as indicated from the female/male earnings ratio, is commonly witnessed to be less than one and is documented in several studies (Hampton and Heywood, 1993; Anker, 1997; Hoffner and Greene, 1997; Ashraf and Ashraf, 1998; Nor, 1998). The overall gender pay earning has widened between 1983 and 2004 by 0.03 log points in India, especially in specific services and industries, characterised by a high female employment rate (Dutta and Reilly, 2008).

Numerous research outcomes have been documented towards exploring the causes of the gender earnings gap. Labour productivity mainly depends on the educational attainments of workers, labour market conditions, occupation safety, business environment, public investment in infrastructure, advancement, and adoption of technology, etc. Therefore, education is one of the factors for productivity growth, and it may, in turn, lead to better wages, safe working conditions, wage security, increased profits, increase in revenue to Governments etc. Females tend to spend less years acquiring formal education, which affects their productivity and adversely affects their earnings. Kingdon (1998) empirically tested labour market discrimination against women using household-level data and revealed that women lack incentives to invest in schooling than boys and reap less return than boys in the labour market.

Similarly, Azam (2012) examined the evolution of wages based on individual-level earning data from the urban area from 1983–2004 and shows that the return of secondary and tertiary education has increased since 1990, resulting in wage inequality. Mohanty (2021) examined the gender earnings gap among workers with similar technical qualifications using employment data from the National sample survey 2011–12 in India and revealed that women lagged in attaining technical education and unemployment. The findings further attribute marriage, having children and low linkage with the labour market as significant factors for low monetary rewards for females compared to male workers. Kijima (2006) examined India's age gap and inequality since economic reforms (1991) and showed that earning inequality in the urban area had begun well before 1991. The study revealed that increased return to skills and increasing demand for skilled labour yield skill premium. The experience of developed nations also reveals that education and skills are immensely helpful in high growth and raising the wage level and living standard (Billard, 2017) besides promoting entrepreneurship by reducing the fear of failure (Mor, Madan, Chhikara, 2020). However, Gangel and Ziefle (2009) attributed motherhood and family responsibilities impact women earnings negatively instead of differences in human capital endowments.

1 GENDER EARNINGS GAP EXPLANATIONS

The worldwide average labour force participation rate stands at about 62 percent of the working-age population (approximately 3.3 billion individuals). Among all employed, 54 percent (1.8 billion) are wage/salaried workers[3] (ILO, 2018a). For most workers, earning wages/salaries constitute a significant proportion of their total household income ranging from about 40 percent in low and middle-income countries to 60–80 percent in high-income economies (ILO, 2017) and have prominently been witnessed in European countries (de Pleijt and van Zanden, 2021). At the same time, all working people are not

[3] The persons who worked in other farm or non-farm enterprises and received receiving piece wage or salary and paid apprentices, both full time and part-time in return regularly (i.e., not based on a daily or periodic renewal of work contract; NSO, 2019).

[4] Self-employed workers operate their enterprises on their own account or with a few partners without hiring any labour during the reference period. They could have had unpaid helpers to assist them in the enterprise's activity (NSO, 2019).

paid employees; rather many are either self-employed/own-account workers[4] or contributing to family businesses, especially in low and middle-income countries. More than 70 percent of workers, whose primary income derives from self-employment, are engaged in small-scale, unincorporated entrepreneurial activities. This indicates the need of exploring the gender earnings gap from all self-employment activities rather than from small businesses, as only a handful of studies have attempted to investigate the economic consequences of self-employment for male and female workers separately.

Occupational segregation and the gender earnings gap are found to be inversely correlated. Though there is a gender earnings gap in all occupational categories, a representation of females in the higher end of earning spectrum (legislators, senior officials, and managers) indicates that they are aware of their rights and face the lowest gender earnings gap. But at the same time, these constitute only one percent of the total female workers (ILO, 2018b), and majority of females are employed in low skilled occupations, are paid low wages, and have a lower probability of getting social security benefits compared to men. Male workers earn a premium for providing long hours in accordance with requirements. Earlier work in this line has considered differences in human capital accumulation of workforce by gender while making a preference for any occupation for livelihood.

There are two major theories of choosing self-employment as a carrier option over regular wage employment. One is the disadvantaged worker argument, and another is the class mobility hypothesis (Budig, 2006). The former claims the absence of an attractive mix of human capital and inability to obtain employment, whereas the latter argues for escapism for undesirable employment opportunities to choose self-employed to improve their economic situation. Moreover, compensating differentials argue that females with greater family responsibilities trade earnings from work in lieu of work time flexibility to meet family commitments and childcare. This also explains the reason for the return of female workers to non-professional self-employment. But it is less influential for interpreting females' return to professional self-employment (Budig, 2006).

1.2 Significance and scope of the study

In the developing era, females have increased their productivity-enhancing capabilities and have increased their employability across diverged occupations. At the same time, they also have emerged as self-employed workers in every occupation. Considering the view, the present study evolves around the exploring the persistence gender earning gap in general and tends to examine the same across broad occupational groups and also for self-employed workers and regular wage employees to get the concrete picture of the scenario. Though, the difference in educational attainments of workers has been an important force of earning from work, but certain studies do not support any skill-based reason for earning gap and have found the gender of workers taking the lead in this concern (Goldin, 2014; Miller, 2016). With this, differences in workers' education need to be neutralised to examine the gender earnings gap across broad occupational groups and work status of workers to capture the real effect of occupation work status on the gender earnings gap. The real contribution of the paper lies in examining the gender earning gap after neutralizing the effect of differences in educational attainments/skill level of workers. The paper deals with specific research questions such as: Is the impact of occupational segregation on earnings the same for female and male workers? Is the impact of work status on earnings of male and female workers in segregated occupations the same? The answers to these questions are critical for understanding the impact of making occupations and work status choices on gender economic equity. This helps to underline the importance and urgency of framing state policies and their strict implementation to ensure females' active participation in the workforce. In this backdrop, the present endeavour is a fresh attempt to provide crucial insights to policymakers for mainstreaming females into the workforce for efficient and effective utilisation of human resources for the socio-economic progress of India.

The paper unfolds as follows. Section 2 deals with the literature review and develop hypotheses of the study, while Section 3 pertains to the methodology employed during the study. Section 4 dedicates the main findings and discussion, whereas final section concludes the paper with suggestions.

2 LITERATURE REVIEW AND HYPOTHESIS FORMATION

2.1 Persistence of gender earnings gap

The gender earnings gap is a common feature of the labour market as there is unequal allocation of high paying jobs reflecting labour market segmentation by gender, particularly in civil services and unionised workplaces (Pendakur and Pendakur, 2007; Anderson, Hegewisch, Hayes, 2015). In response to compensating earning variation, female workers earn lesser, leading to wider gender earnings gap (Bonin et al., 2007; Azmat and Barbera, 2014). Female work participation has declined in urban areas despite having a wide spectrum of job opportunities, and the decline is more pronounced in the case of illiterate, lower caste, and economically poor females (Ara, 2016). One of the reasons for the gender earnings gap is the existence of wide gender-employment associations across societies, which causes a tipping point for males to work with occupations with too many females to safeguard their masculine identities (Akerlof and Kranton, 2000; George and Rachel, 2000). Earning of female workers would increase by about 10 percent if they were rewarded in the labour market on the same basis as for males (Lissenburgh, 2000). Moreover, the reluctance of male workers to associate with females at workplace (Goldin, 2013), holding bigot attitude towards appropriate roles of females at workplace results in lower female work participation (Pan, 2015), and male workers tend to earn more than their female counterparts (Madan and Mor, 2021). Further, because of classic compensating differential equilibrium (Rosen, 1986), females tend to place a higher value on temporal flexibility, whereas male workers earn premiums for providing long hours of work in workplaces that face higher costs of providing the amenity.

H_{01}: *Gender does not form any basis for earnings gap in any society.*

2.2 Gender earnings gap and occupation

The persistence of occupational segregation is a strong feature of the labour market. Occupation is found to explain larger variation in the wage-earning of the workforce from work (Cortes and Pan, 2017; Madan and Mor, 2020; Madan, 2019). Generally, high paid work opportunities are associated with managerial, professional, and technical related work, requiring higher cognitive, managerial and technical skills with high promotional prospects. Working as clerical support workers, skilled workers in agri-business, service workers can provide moderate earning for work and require skill-oriented education to perform routine official tasks. Lower-level occupation is associated with the secondary labour market, and workers face relatively flat earnings from work. The occupational choices of females depend upon family structure to accommodate family requirements and work (Yee, 2007; ILO, 2015).

Several studies witnessed more gender earnings gaps in higher-level managerial and professional occupational categories (Turner, Christern, Murphy, 2017). Female dominated occupations pay less than male-dominated occupations with similar attributes (Levanon, England, Allison, 2009; Blau and Kahn, 2017). The under-representation of females in male-dominated professions could account for the gender earnings gap as occupation and type of industry explain more than half of the variation in the gender earnings gap (Blau and Kahn, 2017). The separation of occupations based on gender is one of the most lasting socio-structural characteristics of the labour market and the German labour market. After witnessing increasing labour force participation still has a relatively worse labour market for female workers than male workers (Wiepcke, 2011). Different sectors of different occupations differ on a variety of attributes such as earnings stability, earning variance, injury, casualty risk, degree of competition, working hours etc., and gender differences in attitudes toward risk and competition could directly affect

the choice of occupation and, consequently, gender earnings gaps. Female workers are more risk-averse than their male counterparts, which is the reason for female over-representation in low-risk professions/ occupations with lower earning variation.

H_{02}: *Occupational diversity is not a reason of gender earnings gap.*

2.3 Gender earnings gap and work status of workers

Work status of workers as self-employed or regular wage employees has been viewed as an important policy measure walk to move the unemployed labour force out of poverty. The earning of self-employed workers is seen as lesser than salaried employees with the same traits. In this line, Evans and Leighton (1989) hold that many self-employed workers are in small retail businesses and not growth-creating innovators for which they did not earn at par with salaried workers. Despite lower initial earnings compared to salaried workers, self-employed workers sustain their work (Hamilton, 2000). Expanding literature examines the causes of women's increased participation in self-employment (Budig, 2006). Young women do not prefer to work in Egypt's private sector due to the fear of sexual harassment at the workplace, the lack of signed work contracts besides the lower-earning and have long hours and hence do not contribute to pension plans owing to lack of job contracts. In contrast, the jobs in the public sector in Egypt are relatively women-friendly in terms of working hours, workplace gender propriety and the less hierarchical relations and hence preferred by the young women (Ghada, 2010).

H_{03}: *Gender earnings gap does not differ for self-employed and regular wage/salaried workers.*

3 RESEARCH METHODOLOGY

3.1 Database of the study

The study employs a database provided by the Periodic Labour Force Survey (PLFS) conducted by National Statistical Office (NSO) from July 2017 to June 2018. The information on selected indicators related to earning of the Indian workforce engaged in numerous economic activities in diverse occupations has been obtained. Purposefully, information on the monthly earning of 94 446 workers working in broad nine occupational as self-employed or and regular wage/salaried has been considered. Herein, information on the monthly earnings of 78 916 male and 15 530 female workers has been deemed to arrive at the gender earnings gap following the work status of workers in diverged occupations.

3.2 Specification of variables

The study attempts to explore the gender earnings gap of workers while considering their occupations and work status. Herein, the natural log of earning, measured in ₹ (INR), is considered the response variable and treated as a randomised continuous variable. The earning of workers may differ in accordance with the nature of work prescribed by diverged occupations. As a result, nine occupational groups have been considered under the International Standard Classification of Occupations-08 (ILO, 2012) to broadly explore earning variations across occupations. These nine broad occupational groups have been categorised as managers (A), professionals (B), technicians and associate professionals (C), clerical support workers (D), service and sales workers (E), skilled agricultural, forestry and fishery workers (F), craft and related trade workers (G), plant and machine operators and assemblers (H) and elementary workers (I) and are treated as a categorical variable. Similarly, two categories of workers have been considered to define the work status of workers, i.e., self-employed and regular wage/salaried workers. Hereby, work status also is a categorical variable. Further, educational attainments of workers may affect their earning potential, as workers with higher education generally get higher wages, regardless of gender and occupation. At this moment, controlling for years of education would help to improve the likelihood of finding a statistically significant interaction effect between wage, occupation and gender, if it exists. In this way, years of education is treated as a covariate to neutralise its effect while measuring the gender earnings

gap for self-employed and regular wage workers across diverged occupations. Thus, the mean difference in the earning of workers is measured in the presence of educational attainments of workers, considering it as a covariate. This also helps in reducing the error term, against which effects of variables/factors are considered under study.

3.3 Model specification and estimation techniques

The study employs GLM: ANCOVA, a special case of dummy variable regression, to estimate overall mean differences among groups in the presence of covariate(s) in the model (Culpepper, and Aguinis, 2011; Fields, 2016; Rasch, Verdooren, Pilz, 2019). While estimating the mean difference in the dependent variable among defined groups, a continuous variable may be an important explanatory variable contributing to the heterogeneity among defined groups. In this study, while estimating the gender gap in mean log earnings of workers across nine groups of occupations and two groups of work status, years of education have been considered an important variable for its effect on earning of workers. Herein, statistical control is required to explain variation in dependent variables across defined groups as independent variables. The analysis procedure employed for this statistical control is the analysis of covariance (ANCOVA).

3.4 Covariate

Educational attainments of workers, measured in years, are considered a covariate. Including education, a continuous variable, as a covariate reduces the error variance while capturing the effect of factors (occupation, work status and gender) on variation in mean earning. While estimating the gender earnings gap, the mean difference in the earning of the workforce from work is estimated for separate groups of workers as per their occupation and work status, years of education is considered a covariate. Now, estimated marginal means are adjusted for mean years of formal education of workers, i.e., 9 years of formal education. The rationale behind this adjustment process is to neutralise the effect of variations in the educational attainment of workers. If the mean years of education of any comparison groups are above average than that of another group (s) in comparison, then the mean score of that group on the dependent variable will be lowered and vice-versa. The degree to such adjustments on the mean score for any group depends on how far above or below average that group stands on the control variable, i.e., comparison group. Adjustment of mean scores on the dependent variable in this fashion provides the best estimates of various comparison groups as they had identical means on the control variable(s). Herein, workers' education is treated as a covariate to neutralise the effect of the mean earning gap of workers across diverged occupational groups and for different work statuses of workers.

4 RESULTS AND DISCUSSION

4.1 Persistence of gender earnings gap

The study found a significant variation in the estimated marginal mean earning by gender ($F_{1, 94409} = 1\,660.583$, $p < 0.01$). The estimated marginal mean of log earning of male workers, i.e., 9.356 (₹ 11 568), is witnessed to be higher than their female counterparts, i.e., 8.800 (₹ 6 634.24), indicating a difference of Ln 0.556 in their mean earning (Table 1). It indicates that the earnings of male workers are 1.744 times more than that of female workers in general. The study found a significant earning gap of male and female workers regardless of their occupation and work status, which signals the prevalence of gender discrimination. As education/skill effect of all workers is neutralized, hereby gender can be considered as a basis of earning gap among workers. It's worth highlighting the research findings of Mor et al. (2020), which underlined those male managed ventures survive for a longer period than their female counterparts. An ample of studies have brought out the reasons for gender earnings gaps. Among many, gender differences in human capital endowments (Gangel and Ziefle, 2009), glass ceiling as well as sticky floors for female workers (Nazier, 2017), motherhood and family responsibilities (Presser, 1995; Casper and O'Connell,

1998; Bianchi, 2000), gender prejudices related to an occupational preference (Leuze and Strauß, 2016) have been some of the important reasons for the persistence of gender earnings gap. So far Indian labour market is concerned. Females require flexible working hours to handle household responsibilities such as childcare concerns and management of household tasks. High paying work opportunities with specific skill requirements and working hours are more rigid are considered less attractive for female workers. Despite lack of financial resources, females choose not to work with organizations with rigid working hours in India. *With this, 1st maintained hypothesis of the absence of the gender earnings gap can be rejected.*

Table 1 Mean earning gap by gender of workers

Sr. No	Gender of worker	$Log_e'X^b$	Mean earning difference	$Antilog_e'X^c$
(i)	Male workers	9.356[a]	.556*	1.744
(ii)	Female workers	8.800[a]	–.556*	

The effect of linearly independent pairwise comparisons among the estimated marginal means: F test

	Sum of Squares	DOF	Mean Square	F
Contrast	694.385	1	694.385	1 660.583*
Error	39 477.818	94 409	0.418	

Notes: Response variable: Ln (earning of workers in ₹); [a] indicates that covariates appearing in the model are evaluated at 9 years of formal education. * indicates significant at 0.01 level of significance; [b] natural Log of mean monthly earning of workers; [c] antilog of mean monthly earning gap by gender.
Source: Author's calculations

4.2 Prevalence of gender earnings gap across occupations

Table 2 provides the mean log earnings of workers by occupation. Segregated factorial analysis about occupation indicates that the grand mean of log earnings for all workers are found to be 9.078 (₹ 8 760.42), ranging from 9.408 (₹ 12 185) for managers to 8.779 (₹ 6 496.37) for craft and related trade workers. There is a significant variation in the estimated marginal mean earning of workers among various occupational groups as indicated by $F_{8, 94409} = 313.471$, $p < 0.01$.

Table 2 clearly indicates a significant earning gap among workers in diverged occupations. So far as occupational group A is concerned, the mean earning of workers is significantly higher than the workers in other occupations except for workers in occupational group D (clerical support workers). Similarly, the mean earning gap of workers in occupational group B is less than those working with occupational group A but greater than those in other occupational categories. This difference is found significant for all workers except for those working in group D. Similarly, the mean earnings of workers in occupational group C is less than those working with occupational group A, group B and group D but greater than those in other occupational categories.

So far as the mean earning gap of workers in occupational group D is concerned, the mean earning of workers for this occupational group is significantly less than those working with occupational group A, group B and group C but greater than those in other occupational categories. Similarly, the mean earning gap of occupational group E is significantly less than those working with occupational group A, group B, group C and group D but greater than those in other occupational categories. The mean earning gap of occupational group F is significantly less than those in other occupational categories, except for workers working with occupational group G. At the same time, the mean earning gap of occupational group G is significantly less than those in all other occupational groups.

Table 2 Mean earning of workers across occupations

Number of occupational groups	Name of occupational groups	Mean earning of workers
A	Managers	9.408[a]
B	Professionals	9.294[a]
C	Technicians and associate professionals	9.133[a]
D	Clerical support workers	9.387[a]
E	Service and sales workers	9.014[a]
F	Skilled agricultural, forestry and fishery workers	8.901[a]
G	Craft and related trade workers	8.779[a]
H	Plant & machine operators and assemblers	8.952[a]
I	Elementary workers	8.834[a]
	Grand mean	**9.078[a]**

The effect of linearly independent pairwise comparisons among the estimated marginal means: F test

	Sum of squares	DOF	Mean square	F
Contrast	1 048.643	8	131.080	313.471*
Error	39 477.818	94 409	0.418	

Notes: Response variable: Ln (earning of workers in ₹); [a] indicates that covariates appearing in the model are evaluated at 9 years of formal education; * indicates significant at 0.01 level of significance.
Source: Author's calculations

The mean earning of workers working with occupational group H and group I is significantly less than those in other occupational groups except the mean earning of workers in occupational group H (Table 3). It clarifies that the mean earning of workers in diverged occupational groups differ significantly.

Table 3 Mean difference in the log monthly earning of workers of specified occupational group with other occupational groups

Occupational groups	Occupational groups								
	1	2	3	4	5	6	7	8	9
A	–	.113*	.275*	.021	.394*	.506*	.628*	.456*	.574*
B	–.113*	–	.162*	–.092	.280*	.393*	.515*	.342*	.461*
C	–.275*	–.162*	–	–.254*	.119*	231*	.354*	.181*	.299*
D	–.021	.092	.254*	–	.373*	.485*	.608*	.435*	.553*
E	–.394*	–.280*	–.119*	–.373*	–	.112*	.235*	.062	.180*
F	–.506*	–.393*	–.231*	–.485*	–.112*	–	.122*	–.050	.068
G	–.628*	–.515*	–.354*	–.608*	–.235*	–.122*	–	–.173*	-.055*
H	–.456*	–.342*	–.181*	–.435*	–.062	–.050	.173*	–	.118*
I	–.574*	–.461*	–.299*	–.553*	–.180*	–.068	.055*	–.118*	–

Note: * significant at 1 percent levels of significance.
Source: Author's calculations

Herein, the occupational earnings gap, as indicated in the present study, is following the investigation by Cortes and Pan (2017), which explored that upper-tier work opportunities are typically associated with managerial, professional, technical professionals, whereas clerical support workers, skilled workers in agri-business, service workers signify middle-level occupational categories. The defined work opportunities differ in cognitive, managerial and technical skills leading to earnings gap among workers, as indicated in the research findings of Turner et al. (2017).

Table 5 signifies the gender earnings gap across diverged occupations, work status and gender of workers. It makes clear that significant gender earnings gap exists across all occupational groups. A perusal of statistics in Table 5 clarifies that the gender earnings gap is witnessed to be maximum in occupational group G as the earnings of male workers are estimated to be 2.35 times more than that of female workers. In this same line, the gender earnings gap is high for workers in occupational group C (2.17 times) followed by group H (1.85 times), in favour of male workers. It is observed to be least for workers in occupational group D (1.13 times), preceded by occupational group E (1.64 times) and occupational group A (1.66 times).

Resultantly, it can be concluded that there exists significant gender earning gap of workers working with diverged occupational groups and education/skill of workers are not responsible as its effect is neutralized to drive out the effect of factor under consideration. The prevalence of the gender earnings gap across occupations, as brought up by this study, is consistent with the research findings of several studies. Certain studies have underlined the choice of occupational groups for work (Turner et al., 2017) on various parameters. Studies on labour market segmentation by gender (Georgellis and Wall, 2005; Pendakur and Pendakur, 2007; Levanon et al., 2009; Anderson et al., 2015; Madan, 2019) have underlined earning variations under gender dominating occupations. This makes clear that gender earning gaps across occupations is India is in line with other countries for which segregated skill requirement and work experience are the main reasons. *Hence, our 2nd maintained hypothesis of type of occupation or occupational diversity is not the reason of wage-earning of workers across can be rejected.*

4.3 Earnings gap and work status of workers

So far as the work status of workers is concerned, there exists significant variation in the estimated marginal means in the monthly earning of self-employed and wage/salaried workers as indicated by the value of F statistic ($F_{1, 94409} = 546.217$, $p < 0.01$), ranging from Ln (8.918) (₹ 7 465.14) for self-employed workers to Ln (9.238) (₹ 10 280.46) for regular salaried employees. The mean log earning of regular wage workers is significantly greater than own account workers indicating that the earning of regular wage earners is 1.377 times more than own-account workers (Table 4).

Table 4 Monthly mean earning gap from work by work-status of workers

Sr. No	Work status	$Log_e'X^b$	Mean earning difference	$Antilog_e'X^c$
A	Self-employed workers	8.918[a]	–.320*	1.377
B	Regular wage/salaries employees	9.238[a]	.320*	

The effect of linearly independent pairwise comparisons among the estimated marginal means: F test

	Sum of squares	DOF	Mean square	F
Contrast	228.405	1	228.405	546.217*
Error	39 477.818	94 409	0.418	

Notes: Response variable: Ln (earning of workers in ₹); [a] indicates that covariates appearing in the model are evaluated at 9 years of formal education; * indicates significant at 0.01 level of significance; [b] natural Log of mean monthly earning of workers; [c] antilog of mean monthly earning gap by gender.
Source: Author's calculations

Several studies, herein, supported earnings gap in accordance with the work status of workers (Evans and Leighton, 1989; Hamilton, 2000). Moreover, divergence in work status has different requirements related to skill, finance, and scale of operation, leading to earning gap from work.

Though the gender earnings gap is a common feature for all workers, it is more prominent among self-employed workers than regular wage/salaried workers. Self-employed male workers earn 1.95 times more than female workers, whereas regular wage male workers earn 1.55 times than female workers, on average. This clarifies that the gender gap persists in the earnings of workers regardless of their work status. The perusal of statistics, in this concern, shows that the gender earnings gap for self-employed workers and regular salaried workers differ in accordance with occupational categories (Table 5). Numerous studies provide support for the gender earnings gap in this concern. Different occupations require different skill requisites, financial requirements, operation scale, and labour market endowments, leading to an earning gap among workers.

Most females choose to become self-employed due to childcare concerns, flexible working timings (Presser, 1995; Casper and O'Connell, 1998; Bianchi, 2000), and do not spend sufficient time on their work. Moreover, female self-employed workers tend to start with work wherein financial requirements are comparatively less (Georgellis and Wall, 2000b), leading to a gender earnings gap. Further, the dominance of male workers in gainful employment options is also one of the reasons for the gender earnings gap (Georgellis and Wall, 2000a).

Table 5 Gender earnings gap by occupation and work status

Broad occupational groups and description	Work status	Gender	*Log_e'X	Gender earnings gap	**$Antilog_e$'X	N
Managers (Category A) Chief executives, senior officials, legislators, administrative and commercial managers, production and specialised services managers, hospitality, retail and other services managers	Self-employed workers	Male	9.393	0.941	2.389	6 546
		Female	8.522			882
		Total	8.958			7 428
	Regular salaried/ wage workers	Male	9.929	0.129	1.153	1 898
		Female	9.787			257
		Total	9.858			2 155
	Total	Male	9.661	0.757	1.660	8 444
		Female	9.154			1 139
		Total	9.408			9 583
Professionals (Category B) Science and engineering professionals, health professionals, teaching professionals, business and administration professionals, information and communications technology professionals, legal, social and cultural professionals	Self-employed workers	Male	9.353	0.706	2.036	1 722
		Female	8.642			279
		Total	8.998			2 001
	Regular salaried/ wage workers	Male	9.773	0.365	1.439	3 710
		Female	9.409			1 917
		Total	9.591			5 627
	Total	Male	9.563	0.308	1.713	5 432
		Female	9.025			2 196
		Total	9.294			7 628

Table 5 (continuation)

Broad occupational groups and description	Work status	Gender	*Log_e'X	Gender earnings gap	**$Antilog_e$'X	N
Technicians and associate professionals (Category C) Science and engineering associate professionals; health associate professionals; business and administration associate professionals; legal, social, cultural, and related associate professionals; information and communications technicians)	Self-employed workers	Male	9.387	0.866	2.512	797
		Female	8.466			118
		Total	8.927			915
	Regular salaried/ wage workers	Male	9.656	0.647	1.887	4 005
		Female	9.021			2 420
		Total	9.339			6 425
	Total	Male	9.522	0.617	2.177	4 802
		Female	8.744			2 538
		Total	9.133			7 340
Clerical support workers (Category D) Occupation as general and keyboard clerks; customer services clerks; numerical and material recording clerks and other clerical support workers	Self-employed workers	Male	9.322	0.179	1.074	81
		Female	9.251			20
		Total	9.286			101
	Regular salaried/ wage workers	Male	9.577	0.163	1.197	3 206
		Female	9.397			875
		Total	9.487			4 081
	Total	Male	9.449	0.163	1.133	3 287
		Female	9.324			895
		Total	9.387			4 182
Service and sales workers (Category E) Personal service workers; sales workers; personal care workers and protective services workers	Self-employed workers	Male	9.275	0.469	1.486	6 435
		Female	8.879			785
		Total	9.077			7 220
	Regular salaried/ wage workers	Male	9.248	0.679	1.813	6 497
		Female	8.653			1 539
		Total	8.951			8 036
	Total	Male	9.262	0.607	1.642	12 932
		Female	8.766			2 324
		Total	9.014			15 256
Skilled agricultural, forestry and fishery workers (Category F) Market-oriented skilled agricultural workers; market-oriented skilled forestry, fishery, and hunting workers; subsistence farmers, fishers, hunters and gatherers	Self-employed workers	Male	9.001	0.651	1.775	19 762
		Female	8.427			2 341
		Total	8.714			22 103
	Regular salaried/ wage workers	Male	9.337	0.564	1.644	339
		Female	8.840			42
		Total	9.089			381
	Total	Male	9.169	0.649	1.707	20 101
		Female	8.634			2 383
		Total	8.901			22 484

Table 5 (continuation)

Broad occupational groups and description	Work status	Gender	*Log$_e$'X	Gender earnings gap	**Antilog$_e$'X	N
Craft and related trade workers (Category G) Building and related trades workers, excluding electricians; metal, machinery and related trades workers; handicraft and printing workers; electrical and electronic trades workers; electronics and telecommunications installers and repairers; food processing, wood working, garment and other craft and related trades workers	Self-employed workers	Male	9.200	1.179	3.180	4 472
		Female	8.043			1 412
		Total	8.621			5 884
	Regular salaried/ wage workers	Male	9.217	0.614	1.751	4 525
		Female	8.657			460
		Total	8.937			4 985
	Total	Male	9.208	1.059	2.358	8 997
		Female	8.350			1 872
		Total	8.779			10 869
Plant & machine operators and assemblers (Category H) Stationary plant and machine operators; assemblers; drivers and mobile plant operators	Self-employed workers	Male	9.247	0.827	2.199	3 153
		Female	8.459			68
		Total	8.853			3 221
	Regular salaried/ wage workers	Male	9.111	0.477	1.742	5 093
		Female	8.556			153
		Total	9.051			5 246
	Total	Male	9.260	0.581	1.852	8 246
		Female	8.644			221
		Total	8.952			8 467
Elementary occupations (Category I) Cleaners and helpers; agricultural, forestry and fishery labourer; labourer in mining, construction, manufacturing, and transport; food preparation assistants; preparation assistants; street and related sales and service workers; refuse workers and other elementary workers	Self-employed workers	Male	9.105	0.637	1.738	2 905
		Female	8.552			285
		Total	8.828			3 190
	Regular salaried/ wage workers	Male	9.118	0.677	1.745	3 770
		Female	8.561			1 677
		Total	8.839			5 447
	Total	Male	9.111	0.649	1.742	6 675
		Female	8.556			1 962
		Total	8.834			8 637
Total	Self-employed workers	Male	9.254	0.783	1.958	45 873
		Female	8.582			6 190
		Total	8.918			52 063
	Regular salaried/ wage workers	Male	9.459	0.43	1.556	33 043
		Female	97			9 340
		Total	9.238			42 383
	Total	Male	9.356	0.504	1.744	78 916
		Female	8.800			15 530
		Total	**9.078**			94 446

Notes: Response variable: Ln (earning of workers in ₹); [a] indicates that covariates appearing in the model are evaluated at 9 years of formal education; [b] mean difference indicate earning gap of female and male workers (female earning – male earning); * natural Log of mean earning of workers; ** antilog of mean earning gap by gender.

Source: Author's calculations

So far as the gender earnings gap among self-employed workers is concerned, it is witnessed to be highest for workers in occupational group G (by 3.18 times) followed by occupational group C (by 2.51 times), occupational group A (by 2.38 times) and occupational group B (by 2. 03 times). Hereby, it is evident that the gender earnings gap is a common feature of the Indian labour market as witnessed in many countries of the world (Table 5). The gender earnings gap is least in occupational group C, preceded by occupational group D. Similarly, an examination of earnings of regular workers makes clear that the gender wage gap of regular wage/salaried workers is comparatively less than for self-employed workers. The gender earnings gap, in favour of male workers, is found to be highest for workers in broad occupational group C (1.88 times), followed by those in occupational group E (1.81 times). *This makes us refute 3rd hypothesis gender earning gap does not differ in between self-employed and regular salaried workers as gender earning gap is significantly more in salaried workers than self-employed.*

CONCLUSION AND SUGGESTIONS

The study highlights the fact that there exists a considerable earnings gap in the labour market. A significant part of gender earnings gap among workers has been explained in general and by occupational diversity and work status of workers working as a self-employed or regular wage worker. At the same time, the effect of educational attainments has been neutralised to fetch real earnings gap across occupations, work status and gender separately as education/skill provide a basis for earning gap of workers. However, the persistence of the gender earnings gap within occupational groups and within the same work status reflects the prevalence of the gender earnings gap. The study found significant gender earnings gap across occupations and the work status of workers. The occupational choices of females depend not only on future promotional and growth prospects but also on the family structure to accommodate family requirements and work. This makes females choose such occupations to work wherein they can accommodate their family requirements resulting in lesser earnings compared to their male counterparts. Working as self-employed or regular wage/salary workers is also a cause of earnings gap among workers. The earnings of self-employed is witnessed to be lesser than salaried employees. Females choose to become self-employed due to childcare concerns, movement constraints to work outside and other household responsibilities and require flexibility in working timings. At the same time, they cannot devote sufficient time towards their work and invest financial resources compared to their male counterparts, which results in a wider gender earnings gap for self-employed female workers.

Herein, the study recommends the removal of gender discrimination to raise the self-esteem of female aspirants, enabling them to contribute with more productivity. At the same time, it is important to raise the productivity potential of the female workforce. Herein, professional/vocational education is an important measure. Further, special provisions, e.g., easy finance, marketing, advertisement facilitating, need to be given to the self-employed, especially for female workers, in compensation for their unremunerative services rendered at home in *'bringing-up the civilisations'* for humanity. This helps in the promotion of entrepreneurship culture in society.

Furthermore, the gender earnings gap may reduce female workers' enthusiasm to put less effort, which constitutes half of the labour force. It might reduce the incentives to invest in female education and training, which may negatively affect productivity growth. Again, 'demotivational factors' leading to the gender earnings gap need to be eliminated to ensure equal monetary reward for workers with similar skills and attributes across occupations as these led to depression and social tensions. There are many factors such as family background, cultural differences, mode & type of schooling, managerial capabilities etc. of workers which may affect the earning potential of workers but lack of data/information on the same is the limitation of the study.

ACKNOWLEDGEMENTS

Unit level microdata has been obtained from the official website of the National Statistical Office (NSO), New Delhi, which is exempted from the individual consent of subjects.

References

AGARWAL, M. (2017). *Skill Development in India: The Supply Side Story*. In: KAPUR, D., MEHTA, P. B. (eds.) *Navigating the Labyrinth: Perspectives on India's Higher Education*, Orient Blackswan.

AKERLOF, G. A., KRANTON, R. E. (2000). Economics and Identity [online]. *The Quarterly Journal of Economics*, 115(3): 715–753. <http://www.jstor.org/stable/2586894>.

ANDERSON, J., HEGEWISCH, A., HAYES, J. (2015). The Union Advantage for Women [online]. *Briefing Paper*, IWPR #R409, Washington, DC: Institute for Women's Policy Research. <https://iwpr.org/wp-content/uploads/wpallimport/files/iwpr-export/publications/(R409)%20Union%20Advantage.pdf>.

ANKER, R. (1997). Theories of occupational segregation by sex: an overview. *International Labour Review,* 136(3): 315–39.

ARA, S. (2016). Gender and jobs: Evidence from urban labour market in India [online]. *The Indian Journal of Labour Economics,* 58: 377–403. <https://doi.org/10.1007/s41027-016-0027-2>.

ASHRAF, J., ASHRAF, B. (1998). Earnings in Karachi: Does gender make a difference [online]. *Pakistan Economic and Social Review,* 36(1): 33–46. <https://www.jstor.org/stable/25825167>.

AZAM, M. (2012). Changes in Wage Structure in Urban India, 1983–2004: a Quantile Regression Decomposition [online]. *World Development,* 40(6): 1135–1150. <https://doi.org/10.1016/j.worlddev.2012.02.002>.

AZMAT, G., BARBERA, P. (2014). Gender and the labor market: What have we learned from field and lab experiments? [online]. *Labor Economics,* 30: 32–40. <https://doi.org/10.1016/j.labeco.2014.06.005>.

BIANCHI, S. M. (2000). Maternal employment and time with children: dramatic change or surprising continuity? [online]. *Demography,* 37(4): 401–414. <https://doi.org/10.1353/dem.2000.0001>.

BILLARD, L. (2017). Study of salary differentials by gender and discipline [online]. *Statistics and Public Policy*, 4(1): 1–14. <https://doi.org/10.1080/2330443X.2017.1317223>.

BLAU, F. D., LAWRENCE, M. K. (2014). The gender wage gap: Extent, trends, and explanations. [online]. *Journal of Economic Literature*, 55(3): 789–865. <https://doi.org/10.1257/jel.20160995>.

BONIN, H. et al. (2007). Cross sectional earnings risk and occupational sorting: The role of risk attitudes [online]. *Labour Economics,* 14(6): 926–937. <https://doi.org/10.1016/j.labeco.2007.06.007>.

BUDIG, M. J. (2006). Gender, self-employment, and earnings [online]. *Gender and Society*, 20(6): 725–753. <https://doi.org/10.1177/0891243206293232>.

CASPER, L. M., O'CONNELL, M. (1998). Work, income, the economy, and married fathers as childcare providers [online]. *Demography*, 35(4): 243–250. <https://doi.org/10.2307/3004055>.

CORTES, P., PAN, J. (2017). Occupation and Gender. *IZA Discussion Paper*, No. 10672, IZA Institute of Labor Economics, Deutsche Post Foundation.

CULPEPPER, S. A., AGUINIS, H. (2011). Using Analysis of Covariance (ANCOVA) with Fallible Covariates [online]. *Psychological Methods*, 16(2): 166–178. <https://doi.org/10.1037/a0023355>.

DAS, P. (2012). Wage Inequality in India: Decomposition by Sector, Gender and Activity Status [online]. *Economic and Political Weekly,* 47(50): 58–64. <http://www.jstor.org/stable/41720467>.

DE PLEIJT, A., VAN ZANDEN, J. L. (2021). Two worlds of female labour: Gender wage inequality in western Europe, 1300–1800 [online]. *Economic History Review,* 74(3): 611–638. <https://doi.org/10.1111/ehr.13045>.

DUTTA, P., REILLY, B. (2008). The gender pay gap in an era of economic change: Evidence for India, 1983 to 2004. *Indian Journal of Labour Economics,* 51(3): 341–360.

EVANS, D. S., LEIGHTON, L. S. (1989). Some empirical aspects of entrepreneurship [online]. *American Economic Review*, 79(3): 519–535. <https://www.jstor.org/stable/1806861>.

FIELDS, A. (2016). *Discovering statistics using IBM SPSS Statistics*. Analysis of Covariance, ANCOVA (GLM 2), Sage Publications Ltd., Chapter 11.

GANGEL, M., ZIEFLE, A. (2009). Motherhood, labor force behavior, and women's careers: an empirical assessment of the wage penalty for motherhood in Britain, Germany, and the United States [online]. *Demography,* 46(2): 341–369. <https://doi.org/10.1353/dem.0.0056>.

GEORGE A. A., RACHEL, E. K. (2000). Economics and Identity [online]. *Quarterly Journal of Economics*, 115(3): 715–753. <https://doi.org/10.1162/003355300554881>.

GEORGELLIS, Y., WALL, H. J. (2000a). What makes a region entrepreneurial? Evidence from Britain [online]. *Annals of Regional Science,* 34(3): 385–403. <https://doi.org/10.1007/s001689900014>.

GEORGELLIS, Y., WALL, H. J. (2000b). Who are the self-employed? *Federal Reserve Bank of St. Louis Review*, 82(6): 15–23.

GEORGELLIS, Y., WALL, H. J. (2005). Gender differences in self-employment [online]. *International Review of Applied Economics,* 19(3): 321–42. <https://doi.org/10.1080/02692170500119854>.
GHADA, B. (2010). When there is 'no Respect' at Work: Job Quality Issues for Women in Egypt's Private Sector [online]. *OIDA International Journal of Sustainable Development*, 1(1): 67–80. <https://ssrn.com/abstract=1661196>.
GOLDIN, C. (2014). A grand gender convergence: its last chapter [online]. *American Economic Review,* 104(4): 1091–1119. <https://doi.org/10.1257/aer.104.4.1091>.
GOLDIN, C. (2013). A Pollution Theory of Discrimination: Male and Female Differences in Occupations and Earnings. In: *Human Capital and History: the American Record*, University of Chicago Press.
HAMILTON, B. H. (2000). Does entrepreneurship pay? An empirical analysis of the returns to self-employment. *Journal of Political Economy,* 108(3): 604–631.
HAMPTON, M. B., HEYWOOD, J. S. (1993). Do workers accurately perceive gender wage discrimination? [online]. *Industrial and Labor Relations Review,* 47(1): 35–49. <https://doi.org/10.1177/001979399304700103>.
HOFFNER, E., GREENE, M. (1997). Gender discrimination in the public and private sectors: a sample selectivity approach [online]. *Journal of Socio-Economics,* 25(1): 105–114. <https://doi.org/10.1016/S1053-5357(96)90056-6>.
ILO. (2012). *International Standard Classification of Occupations*. Geneva, Vol. 1.
ILO. (2015). *Global Wage Report. Wage and Income Inequality* [online]. Geneva. <https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_324678.pdf>.
ILO. (2017). *Global Wage Report 2016/17: Wage inequality in the workplace* [online]. Geneva. <https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_537846.pdf>.
ILO. (2018a). *India Wage Report-Wage policies for decent work and inclusive growth* [online]. Geneva. <https://www.ilo.org/wcmsp5/groups/public/---asia/---ro-bangkok/---sro- new_delhi/documents/publication/wcms_638305.pdf>.
ILO. (2018b). *Global Wage Report. What lies behind gender pay gaps* [online]. Geneva. <https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_650553.pdf>.
KIJIMA, Y. (2006). Why did wage inequality increase? Evidence from urban India 1983–99 [online]. *Journal of Development Economics*, 81(1): 97–117. <https://doi.org/10.1016/j.jdeveco.2005.04.008>.
KINGDON, G. G. (1998). Does the labour market explain lower female schooling in India? [online]. *Journal of Development Studies,* 35(1): 39–65. <https://doi.org/10.1080/00220389808422554>.
LEUZE, K., STRAUSS, S. (2016). Why do occupations dominated by women pay less? How 'female- typical' work tasks and working-time arrangements affect the gender wage gap among higher education graduates [online]. *Work, Employment and Society,* 30(5): 802–820. <https://doi.org/10.1177/0950017015624402>.
LEVANON, A., ENGLAND, P., ALLISON, P. D. (2009). Occupational Feminization and Pay: Assessing Causal Dynamics Using 1950–2000 U.S. Census Data [online]. *Social Forces*, 88(2): 865–891. <https://doi.org/10.1353/sof.0.0264>.
LISSENBURGH, S. (2000). Gender Discrimination in the Labour Market. *PSI Research Discussion Series*, Policy Studies Institute, UK.
LIVANOS, I., POULIAKAS, K. (2012). Educational segregation and the gender wage gap in Greece [online]. *Journal of Economic Studies,* 39(5): 554–575. <https://doi.org/10.1108/01443581211259473>.
MADAN, S. (2019). Wage differentials among workers: an empirical analysis of the manufacturing and service sectors [online]. *Indian Journal of Labour Economics*, 62(4): 731–47. <https://doi.org/10.1007/s41027-019-00195-4>.
MADAN, S., GOEL, R. (2019). Outcomes of labour market for informal workers: an analysis of influential factors in Haryana. *Man and Development,* 41(2): 51–70.
MADAN, S., MOR, S. (2020). Skill and wage-earning potential: Evidence from Indian labour market [online]. *Statistika: Statistics and Economy Journal,* 100(4): 245–260. <https://www.czso.cz/documents/10180/125507861/32019720q4_392-407_madan_analyses.pdf/a5c58983-5310-4c78-89e0-b6c10a069129?version=1.1>.
MADAN, S., MOR, S. (2021). Do occupation, work status and gender cause variations in wages: Case of Indian labour market [online]. *International Journal of Economic Policy in Emerging Economies*. <https://doi.org/10.1504/IJEPEE.2021.10040188>.
MOR, S., MADAN, S., CHIKHARA, R. (2020). The risk-seeking propensity of Indian entrepreneurs: a study using GEM data [online]. *Strategic Change,* 29(3): 311–319. <https://doi.org/10.1002/jsc.2330>.
MILLER, C. C. (2016). As Women take over a male-dominated field, the pay drops [online]. *New York Times*, March 18. <https://www.nytimes.com/2016/03/20/upshot/as-women-take-over-a-male-dominated-field-the-pay-drops.html>.
MOHANTY, S. (2021). A distributional analysis of the gender wage gap among technical degree and diploma holders in urban India [online]. *International Journal of Educational Development*, 80: 102322. <https://doi.org/10.1016/j.ijedudev.2020.102322>.
MOR, S., MADAN, S., ARCHER, G. R., ASHTA, A. (2020). Survival of the smallest: a study of microenterprises in Haryana, India [online]. *Millennial Asia,* 11(1): 57–78. <https://doi.org/10.1177/0976399619900609>.
NAZIER, H. (2017). The conditional gender wage gap in Egypt: premium or penalty? Topics in Middle Eastern and African Economies. *Proceedings of Middle East Economic Association,* 19(2): 67–95.
NOR, L. M. (1998). An overview of gender earning differentials in peninsular Malaysia. *Journal of Economics and Management,* 6(1): 23–49.

NSO. (2019). *Periodic Labour Force Survey (PLFS), 2017–18, Unit level Data*. Ministry of Statistics and Programme Implementation, National Statistical Office, Government of India.
PAN, J. (2015). Gender Segregation in Occupations: the Role of Tipping and Social Interactions. *Journal of Labor Economics*, 33(2): 365–408.
PENDAKUR, K., PENDAKUR, R. (2007). Minority earnings disparity across the distribution [online]. *Canadian Public Policy,* 33(1): 41–61. <https://doi.org/10.3138/cpp.v33.1.041>.
PODDAR, S., MUKHOPADHYAY, I. (2019). Gender Wage Gap: Some Recent Evidence from India [online]. *J. Quant. Econ.,* 17: 121–151. <https://doi.org/10.1007/s40953-018-0124-9>.
PRESSER, H. B. (1995). Job, family, and gender: Determinants of non-standard work schedules among employed Americans in 1991 [online]. *Demography*, 32(4): 577–598. <https://doi.org/10.2307/2061676>.
RASCH, D., VERDOOREN, R., PILZ, J. (2019). *Applied statistics: Theory and problem solutions with R* [online]. John Wiley and Sons, Ltd., Chapter 9. <https://doi.org/10.1002/9781119551584>.
ROSEN, S. (1986). The Theory of Equalising Differences [online]. In: ASHENFELTER, O., LAYARD, R. (eds.) *The Handbook of Labour Economics*, University of Chicago, Amsterdam: Elsevier-North Holland, 1: 641–92. <https://doi.org/10.1016/S1573-4463(86)01015-5>.
STEPHEN, M. (1998). Recent shifts in wage inequality and the wage returns to education in Britain. *National Institute Economic Review,* 166(1): 87–95.
TURNER, T., CHRISTERN, C., MURPHY C. (2017). Occupations, age and gender: Men and Women's earnings in the Irish labour market [online]. *Economic and Industrial Democracy,* 38(2): 1–20. <https://doi.org/10.1177/0143831X17704910>.
WEEDEN, K. A. (1998). Revisiting occupational sex segregation in the United States, 1910– 1990: results from a log-linear approach [online]. *Demography,* 35(4): 475–487. <https://doi.org/10.2307/3004015>.
WEICHSELBAUMER, D., WINTER, E. R. (2005). A meta-analysis of the gender wage gap [online]. *Journal of Economic Surveys,* 9(3): 479–511. <https://doi.org/10.1111/j.0950-0804.2005.00256.x>.
WIEPCKE, C. (2011). Gender-specific job choices-implications for career education as part of economic education [online]. *Int. J. of Pluralism and Economics Education,* 2: 355–368. <https://doi.org/10.1504/IJPEE.2011.046023>.
YEE, K. M. (2007). Work orientation and wives' employment careers [online]. *Work and Occupations,* 34(4): 430–462. <https://doi.org/10.1177/0730888407307200>.

Application of the Hybrid Forecasting Models to Road Traffic Accidents in Algeria

Fatih Chellai[1] | *Ferhat Abbas University, Setif, Algeria*

Received 3.11.2021 (revisions received 3.12.2021, 12.2.2022), **Accepted** (reviewed) 15.3.2022, **Published** 17.6.2022

Abstract

Road traffic accidents are a growing public health concern. In this study, we focused on analyzing and forecasting the monthly number of accidents, number of injuries, and number of deaths in Algeria over the period (2015–2020). For this purpose, hybrid forecasting models based on equal weights and in-sample errors were fitted, and we compared them with the seasonal autoregressive moving average (SARIMA) models. The three models retained for forecasting until 2022 are all hybrid models, one based on equal weight and two models based on in-sample errors (using the RMSE indicator). Furthermore, the hybrid models outperformed the SARIMA models for short (6 months), medium (12 months), and long horizon (24 months). The forecasting results showed that we expect an increase in the number of accidents, the number of deaths, and the number of injuries over the next 12 months. Policymakers must enhance strategies for prevention and road safety, especially in rural areas, where the highest rate of fatalities is recorded.

Keywords	***DOI***	***JEL code***
Road traffic accidents, hybrid forecasting models, seasonal time series analysis	*https://doi.org/10.54694/stat.2021.37*	*C22, C32, C53*

INTRODUCTION

Road traffic accidents are a real public health issue, adding that, its negative effects go beyond the health dimension to the social and economic dimensions Racioppi et al. (2004). According to the World Health Organization (WHO), road traffic accidents (RTA) cause 1.3 million fatalities with more than 5 million people injured during 2020, see WHO (2021). The other dark side is the social and economic consequences of road traffic accidents which include degradation of the quality of life and psychiatric impacts for the victim and its family (Mayou et al., 1993), loss of productivity, the cost of the legal system, and medical costs (Ansani et al., 2020; Bardal and Jørgensen, 2017; Chen et al., 2019).

Specifically, Algeria is in the top ranking of the most affected developing countries owing to road traffic accidents. Statistics show an increase of 42.6% in the number of RTAs in the first five months of 2021 compared with 2020. The same tendency was recorded for the number of injuries and the number

[1] Department of Based Education, Ferhat Abbas University, Setif 1, El Bez Campus, 19000, Algeria. E-mail: fatih.chellai@univ-setif.dz, phone : (+213)0663526184.

of fatalities, which increased by 40.8% and 21.6%, respectively, compared to the first five months in 2020. On the other side, the number of vehicles in Algeria was estimated at more than 6.4 million in 2018; in the same year 255 538 new vehicles have been registered. However, according to the data delivered by the National Office of Statistics ONS Algeria (2021) there were 6.1 million cars in 2017, representing an annual increase of 4.1%. The challenge of reducing the number of accidents has led many researchers to provide solutions to determine the effective factors in accident occurrence and to present comprehensive safety programs and strategies. The main aim of this study is to demonstrate the optimality and accuracy of hybrid models in forecasting the trajectories of the number of RTAs, mortalities, and injuries in Algeria.

To the best of the author's knowledge, few studies have applied hybrid methods to predict the patterns of road traffic accidents. This study is a new step in providing high-quality statistics in terms of reliability and regularity over a relatively long period, which can help decision-makers monitor and evaluate the efficiency of prevention strategies. The key objectives are (*i*) to explore the patterns of road accidents in Algeria by considering the spatial dimension, and (*ii*) to introduce hybrid models for forecasting the trajectories of the number of accidents, the number of injuries, and deaths.

The remainder of this article is organized as follows: the next section discusses the most frequently used models in forecasting. Second section presents the main features of the hybrid forecasting models. The third section describes the RTA data for Algeria. The fourth section presents the main results of modeling and forecasting. The last two sections discuss, conclude and summarize the findings of the study.

1 OVERVIEW OF FORECASTING METHODS

With the development of computer and simulation techniques, several statistical models (linear and non-linear) have emerged and have been applied in the field of modeling and forecasting of time series data, the most important of which is the Box-Jenkins method Box and Jenkins (1970), which is a useful linear model that has proven its efficiency and importance in the field of forecasting Ihueze and Onwurah (2018). In a general context, Hyndman and Athanasopoulos (2018) provide a good reference for the best practices and forecasting principles. We also mention the book of the time-series analysis by Hamilton (2020). On the other hand, with the development of machine learning and big data, non-linear models such as artificial neural networks (ANN) have emerge and have also been applied with great acceleration over the past years, principally by Delen et al. (2006) and Rezaie et al. (2011). However, in practice, we face several challenges in choosing the optimal model for data and prediction, and we use information criterions as well as accuracy measures for forecasting. In light of this choice, the approach of merging these candidate models has emerged to reach a single prediction result, which we call the combined forecasting method, see Granger and Ramanathan (1984), and Armstrong (2001). The application of this approach is abundant in different fields; Zhang (2003) applied a hybrid method by combining ARIMA and artificial neural networks (ANNs) and concluded that such a combination can improve the forecasting accuracy better than the single original method. In the same axis Wang et al. (2013) combined ARIMA and the ANNs and tested to forecast different datasets. Abdollahi (2020) applied a hybrid model to forecast the dynamics of oil prices. We also mention the study carried out by Rezaie et al. (2011), who applied artificial neural networks to predict the severity of road accidents, and revealed that several factors, such as human factors and weather factors could increase the crash severity in urban highways. Using the ARIMA and ARIMAX models, Ihueze and Onwurah (2018) attempted to predict the trajectories of road accidents in Anambra State (Nigeria). They concluded that transfer function models (ARIMAX) are preferred over ARIMA models.

Yusuf et al. (2015) used fuzzy logic to develop a hybrid approach for forecasting enrolment and car road accidents. The findings of the study showed that the presented method performs better the forecasting results comparing to other existing methods. Barba et al. (2014) presented a combination of ARIMA models

and autoregressive neural networks (ANNs) to improve the forecasting of traffic accidents. Following a two-step strategy of combination, they revealed that an ARIMA-HSVD (Hankel matrix) performed better in the forecasting results compared with other combinations. In recent study Sangare et al. (2021) developed a new combination framework with a Gaussian mixture model (GMM) and a support vector Classifier (SVC) to forecast urban traffic, they revealed that the new hybrid approach performed better than the road accident baseline statistical models.

2 HYBRID FORECASTING MODELS

Recently, the approach of forecast combining has been widely applied in different fields of research, the main idea of which is to combine the forecasts from different techniques such as ARIMA models, ETS (Error, Trend, Seasonal), and ANN,... the process of combination is based on the weights of each technique to the final forecast output; in practice, we can select the weights through in-sample errors, which was first introduced by Bates and Granger (1969), and the other way to select the weights is by cross-validation. The theoretical background of our study is based on the work of Bates and Granger)1969), which was the original study on this topic. In the same context Yang (2004) revealed that empirical research advocates that hybrid models usually improve forecasting accuracy over the original approaches.

For our application, we used five forecasting methods: (1) the ARIMA model; (2) theta model, which was developed by Assimakopoulos and Nikolopoulos (2000); (3) neural network models Hill et al. (1996) based on feed-forward and sing hidden layers and lagged inputs; (4) the exponential smoothing state-space model (ETS) Hyndman et al. (2002); and (5) the TBATS model (exponential smoothing state-space model with Box-Cox transformation, ARMA errors, trends, and seasonal components), De Livera et al. (2011).

In this study, we are interested in forecasting the number of road accidents, number of deaths, and number of injuries. We define ψ as a forecasting method that provides us the forecasts of $\widehat{y_1}, \widehat{y_2}, \ldots, \widehat{y_k}$ at different horizons up to k, we can simply estimate the average risk $R(\psi; k)$ to measure the accuracy of the ψ method. To achieve this objective, we have a class of forecasting Ω that contains several statistical approaches $\Omega = \{\psi_1, \psi_2, ..., \psi_m\}$, where: m may be finite or infinite. After forecasting, a probable departure e_t of the predicted values $\widehat{y_t}$ from the real valuescan occur, and the optimal approach is that gives us the minimal discrepancy e_t.

In the context of the hybrid forecasting models, any forecasting technique ψ is called a combined forecasting procedure if the forecasts outcomes $\widehat{y_1}, \widehat{y_2}, \ldots, \widehat{y_k}$ constitute a measurable function on the real values $y_1, y_2, ..., y_k$ and the values of $\hat{y}_{i,j}$ where $1 \leq i \leq k$ and $1 \leq j \leq m$.

In a general form, the model that provides the combination of these forecasting methods can be defined as follows:

$$\widehat{y_t} = \sum_{i=1}^{n} w_i \widehat{y_i}(t),$$

where w_i are the weights designed for each forecasting method, generally we have $\sum_i w_i = 1$, but this is not always the case, see for more details Lean et al. (2005).

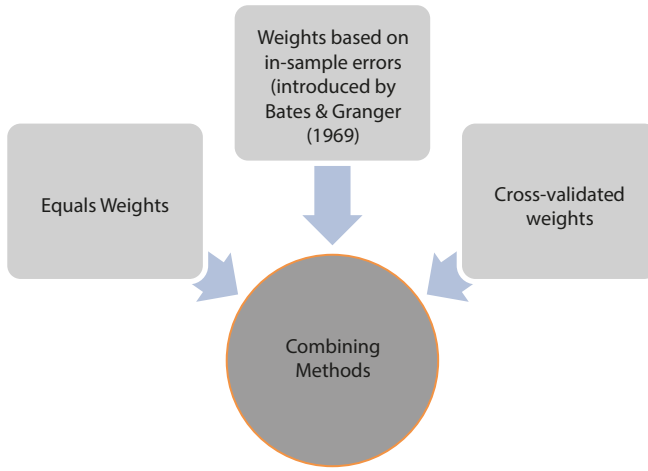
The process of combination can be applied with three approaches:

(1) Equal weights (i.e.) $w_i = \frac{1}{n}$; which is a standard and robust method as depicted by Lean et al. (2005).

(2)Weights based on in-sample errors; the main idea of this method is to resolve a system of equations (generally quadratic programming) to find at the end the optimal weights,

$$\begin{cases} Min\left(w_i\ f\left(e_t\right)\right) \\ \sum_{i=1}^{m} w_i = 1 \quad , w_i \geq 0 \ .i = 1, 2, \ldots, m \quad .t = 1, 2, \ldots, T. \end{cases}$$

Figure 1 Hybrid forecasting models



Source: Own construction

The idea of this method was firstly introduced by Bates and Granger (1969), and we have several options to choose the function $f(e_t)$ as an indication, the R package we work on provides three functions (or errors measures): MAE: Mean Absolute Error, MASE: Mean Absolute Scaled Error and RMSE: Root Mean Square Error.

(3) Cross-validated weights; cross-validation of time series data with user-supplied models and forecasting functions is also supported to evaluate the model accuracy.

The comparison (and model evaluation) between different combined forecasting classes and between the SARIMA models was conducted using the following accuracy measures: the root mean squared errors, $RMSE = \sqrt{\frac{\sum_t (y_t - \hat{y}_t)^2}{n}} = \sqrt{\frac{\sum_t (e_t)^2}{n}}$. The mean absolute errors, $MAE = \frac{\sum_t |y_t - \hat{y}_t|}{n} = \frac{\sum_t |e_t|}{n}$. The maximum of the absolute percentage error, $MAPE = Max\left(\left|\frac{y_t - \hat{y}_t}{y_t}\right|\right) \times 100$. The mean percentage errors, $MPE = \frac{100\%}{n}\sum_{t=1}^{n}\left(\frac{y_t - \hat{y}_t}{y_t}\right)$.

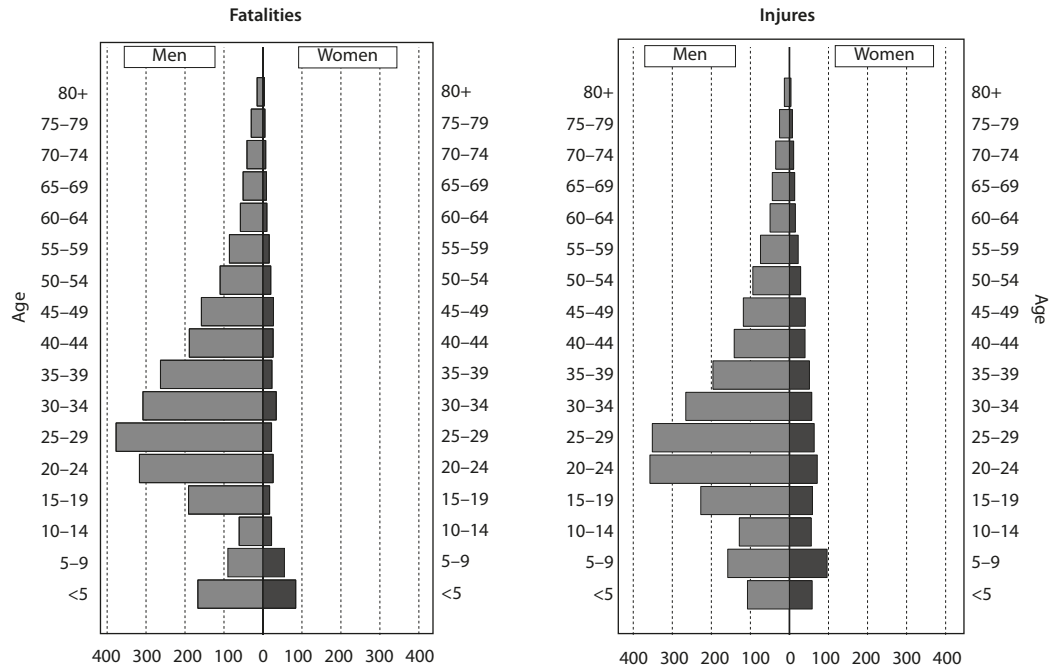
3 AN OVERVIEW OF ROAD TRAFFIC ACCIDENTIN ALGERIA

To analyze the trajectories of road traffic accidents in Algeria, we used data that are provided by the DNSR (Délégation Nationale à la Sécurité Routière) in 2020. The entire dataset was completed after obtaining a license and an official request was submitted to the administrative body of the Ministry of the Interior. In practice, however, the process of collecting information on traffic accidents is undertaken by the Civil Protection, the National Gendarmerie (in rural areas), and the police (in urban areas), after this step, the DNSR mission is to prepare, clean, and organize these data and elaborate periodical reports about the road traffic accidents in the country.

As a global fact, road traffic injuries are among the foremost factors of mortality in the general population. Specifically, children and young adults aged 5–29 years were most affected by these accidents. This is true for Algeria, as the pyramids in Figure 2 showed the distribution of deaths and injuries according to sex and age. We can see that the most affected age-category is 25–29 years for both sexes. Figures in 2020, and as a cumulative frequency, 43.6 % of the deaths and 55.9% of injuries were among persons aged under 29 years. Regardless of the age category, gender statistics show that males are more likely

to be involved in road traffic crashes than females. Approximately three-quarters (73.1%) of all road traffic deaths occur among young men under the age of 25 years, who are almost three times as likely to be killed in a road traffic crash as young women.

Figure 2 Distribution of traffic accident victims by age and sex in 2020



Source: Author's computations based on data provided by the DNSR (2020)

According to statistics provided by the Office National des Statistiques (ONS Algeria), the total number of under-five child mortality was 22 240 in 2019 (for both sexes). The data delivered by the DNSR showed that the number of fatalities among children under five years of age was 25. Consequently, 1.1% of the total deaths among children under five years of age were caused by traffic accidents. Furthermore, the mortality rates were much higher for boys than girls. As conjectural figures, and based on the recent statistics provided by the DNSR in June 2021, we found a clear increase either in the number of accidents and the number of death and injuries, this is shown in Table 1.

Table 1 Evolution of the number of crashes, number of injuries, and number of deaths between 2020 and 2021

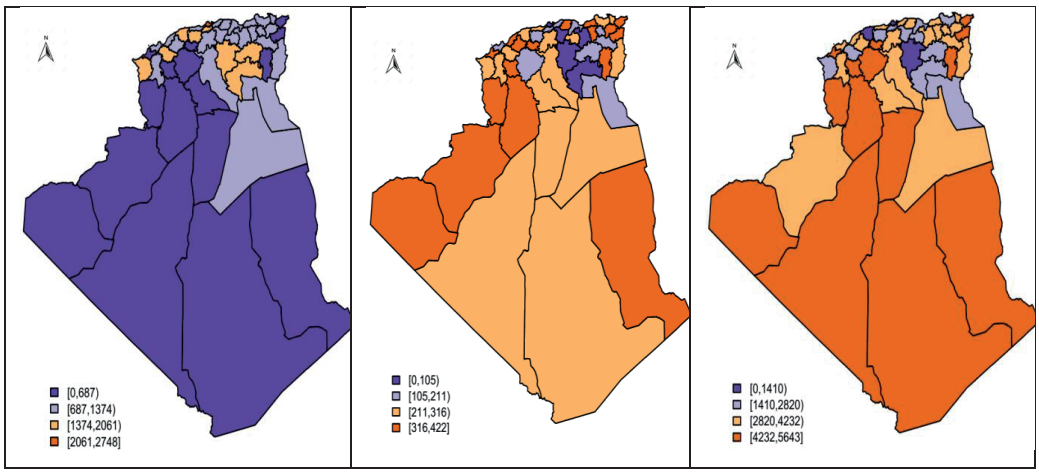
Year	Accidents	Injuries	Deaths
2020	7 216	9 708	1 065
2021	10 292	13 664	1 295

Source: Author's computations based on data provided by DNSR (2020, 2021)

In the first five months of 2020, the number of accidents was 7 216, but in the same period in 2021, the number of accidents was 10 292, with an increase of 42.6 % compared to 2020, the same tendency was recorded for the number of injuries and the number of fatalities, respectively, with an increase of 40.7 %

and 21.6% compared to the first five months in 2020. This high variation between the two periods can be explained by the strategies of the containment due to Covid-19 taking by the government at the beginning of 2020, which has been (after July 2020) removed (partially) by allowing traffic between and among states.

Figure 3 Spatial distribution of the number of crashes, number of injuries, and number of deaths in Algeria



Source: Author's computations based on data provided by the DNSR (2020)

There was a significant difference in the number of accidents, the number of fatalities, and the number of injuries and across the 48 states of Algeria, as shown in the three maps in Figure 2. Furthermore, we estimated the road fatality rate (RFT), and heterogeneity still existed among the states. This spatial analysis can be used as an indicator of comparison among regions and allows policymakers to develop suitable strategies for each region to improve road safety in the country. If detailed statistics for the 48 states of Algeria are available, we believe that spatial modeling approaches, such as geographically weighted regression (GWR), could provide more insights into the spatial differentiation of road traffic accidents in these states, and a recent study conducted by Wachnicka et al. (2021) showed the reliability of this statistical method to identify regional differences in road traffic accidents in Europe.

At another decomposition level, statistics showed that the number of accidents in urban areas was 15 211, representing 66.2% of the total number of accidents. This was twice the number of accidents in rural areas. Compared to 2019, the number increased by 5.1% in urban areas and decreased by 16.3% in rural areas. However, the number of deaths due to traffic accidents is mainly in rural areas. In 2019, the total number of deaths due to road accidents was 2 599, that was 79.4% of the total number, four times higher than in urban areas. The rural traffic network is characterized by ease of traffic, which allows high speeds and a low level of surveillance, representing the greatest challenge in terms of security.

4 RESULTS

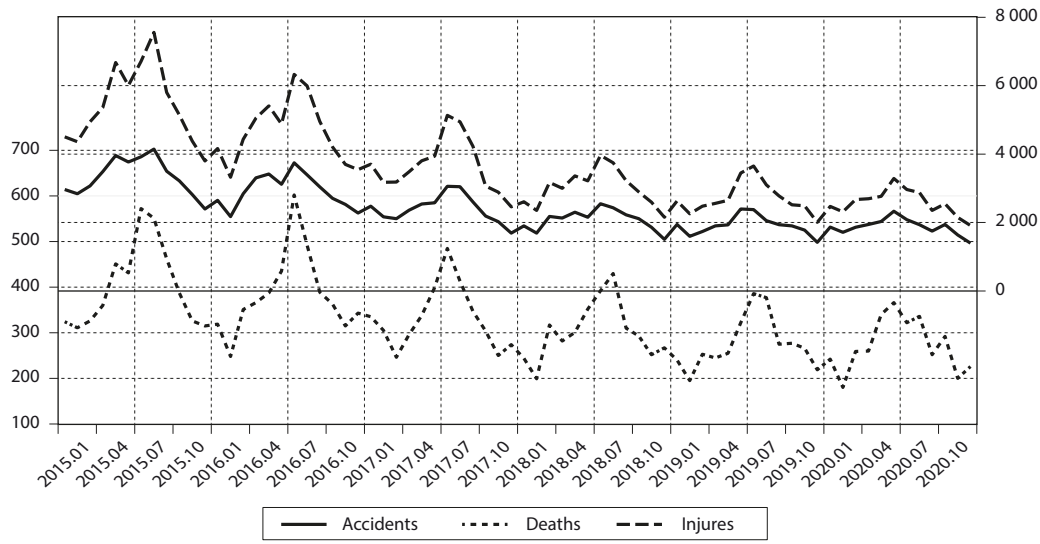
4.1 Stationary time series analysis

The plots in Figure 3 describe the trajectories of the number of accidents, deaths, and injuries in Algeria over the period (2015–2020) as the monthly frequency. Graphically, the three variables exhibited the same pattern over the study period, with a slightly decreasing trend over the period in which a seasonal component in the time series was clearly observed. The descriptive statistics showed that the average number (per month) of accidents was 2 426, of deaths, and 328 and 3 682 injuries per month, respectively. Based on the coefficient of variation (CV), we found the same level of dispersion for

the number of accidents and number of deaths (27.3% and 26.5 %, respectively). In contrast, the highest dispersion was recorded for the number of injuries, with a coefficient of variation of 35.1%. We noted a slight decrease in the number of road traffic accidents in the last year (2020), as well as in the number of injuries and deaths. This decrease may be due to lockdown strategies caused by the Covid-19 pandemic, as health authorities in Algeria have taken several safety measures to reduce transportation by 2020, which has witnessed a large spread of the virus.

An analysis of the outliers was conducted in order to test the validity of the hypothesis of the effect of the lockdown on the reduction in the number of accidents. Specifically, we aim to investigate the presence of level shifts, transient changes, and innovation outliers in the time series over the ten last months (March 2020 to December 2020). For this objective, we used the "tsoutliers" (v0.6-8; Javier, 2019) R package, which was developed on the approach of Chen and Liu (1993). Except for a transient change in the "Deaths" time series that was identified in 2020, the results revealed no presence of a significant switching in level for the three time series during the last year (i.e. in 2020). In contrast, before 2020, the test results showed the presence of a seasonal-level shift (SLS) in the number of "Accidents" time series. At the same time, transient and level changes were observed in the "Injuries" time series during 2016 and 2017, more details are in the Appendix 3.

Figure 4 Time series plots of the evolution of the number of accidents, number of deaths, and number of injuries in Algeria over the period (2015–2020)



Source: Author's computations based on data provided by the DNSR (2020)

For the normality assumption, we conducted statistical tests on the stationary time series, see the Appendix 2. The kernel densities in the left and right axis borders of the plots confirmed the non-normality distribution of the stationary deaths time series and the normality of the injuries and accidents variables; these was also tested by the test of normality of Jarque and Bera. As detailed information about the shape of the data distribution, the Fisher skewness parameters showed that the number of deaths time series followed an asymmetric distribution (skewed left), the kurtosis coefficients indicated that the dispersion of the extremes values is higher in the time series of "number of deaths" compared to the other variables (number of accidents and number of injuries), a detailed reference in measuring skewness is conducted by Doane and Seward (2011).

Since the data doesn't exhibit the trend, and this is for the three variables (Accidents, Injuries, and Deaths), we select to test the unit root hypothesis for "intercept" only, for the optimal lag selection in testing we follow the automatic option based on the Akaike Information Criterion (AIC), Akaike (1974). The critical values of the test were based on simulations, and all statistical programs provided critical values at different levels of significance, according to the sample size. In the literature, we find several seasonal unit root tests, but the most commonly used is the Hylleberg, Engle, Granger, and Yoo (HEGY) test; see Hylleberg et al. (1990). For application, Ronderos (2019) provided a simple and comprehensive procedure using the Eviews program.

Table 2 Seasonal Unit Root Test for the three time-series using the HEGY method

All seasonal frequencies	Tabulated test statistics at different significance level and different sample size			Calculated test statistics		
	1%	5%	10%	Accidents	Deaths	Injuries
n = 40	28.09	7.38	3.43	2.6123	2.7475	3.9065
n = 60	28.137	7.36	3.49			
n = 53*	28.12	**7.37**	3.47			

Note: n = 53* included observations after adjustment, which are obtained using linear interpolation. For the frequencies, typically, we worked on 0-frequency, 2π/4, 6π/4 and π.
Source: Author's calculation

The null hypothesis of this test states that a unit root exists at a specified frequency periodicity. To test this hypothesis, we select the option "all frequencies", and we worked on adjusted sample size n = 53 which is close to the lower and upper values of the simulation, in our case, the lower is n = 40 and the upper is n = 60. Thus, because the critical values of the test statistics (2.61, 2.74, and 3.90) were smaller than the critical value (7.37) at the 5% significance level, the null hypothesis is accepted. The seasonality in the statistical series is expressed by the third quarter corresponding to major holiday trips, and the fourth quarter to social re-entry and the onset of bad weather as the most dangerous periods in terms of road traffic. In general, seasonality in road traffic accidents is partly due to traffic trends and partially due to weather conditions.

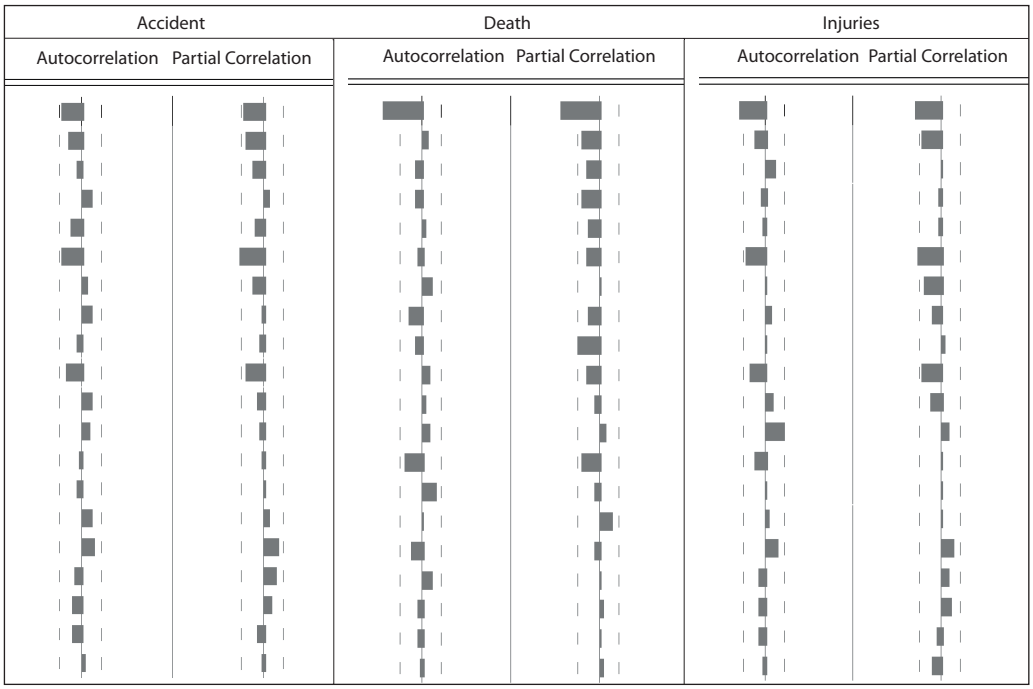
Table 3 The Augmented Dicky-Fuller Unit-Root test

Series	*p. value* at level			*p. value* at first (trend) difference		
	None	Intercept	Intercept & Trend	None	Intercept	Intercept & Trend
Accidents	0.3421	0.0684	0.2244	0.0000	0.0001	0.0004
Fatalities	1	0.1361	0.9971	0.3138	0.0154	0.0203
Injuryies	1	0.9986	0.0171	0.5541	0.0016	0.0056

Source: Author's calculation

As shown in the last two columns of Table 3, the stationarity assumption of the three time series is confirmed by the computed p-values of the Augmented Dicky-Fuller test, which are all lower than the (0.05) significance level. Furthermore, this stationarity was confirmed by the autocorrelation (ACF) and partial autocorrelation (PACF) functions shown in Figure 4, which behave as stationary processes. More precisely, the plots of these functions show the absence of autocorrelation(moving average component) and partial autocorrelation (autoregressive component) in the series of accidents (Figure 5(a)) .The two components (moving average and autoregressive) are statistically significant for the series of deaths and injuries and (b) and (c), respectively.

Figure 5 Autocorrelation and partial auto-correlations functions of the first difference of time series



Source: Author's construction

4.2 Forecasting results and model comparison

After data preparation and stationarity analysis, the main thing remaining to achieve is model identification, selecting the optimal model, and forecasting the trends of the three variables. First, for the model combination (and weight method), we worked on two methods: equal weight ($w_i = \frac{1}{n}$), and in-sample errors. As indicated in the method section, we worked on five principal methods of forecasting; this selection is justified by two reasons: these methods are the most used in forecasting of time series, and also they are all available in the "forecastHybrid" (v.5.0.19; Shaub and Ellis, 2020) R package which helps researchers to usethis combining approach in other studies.

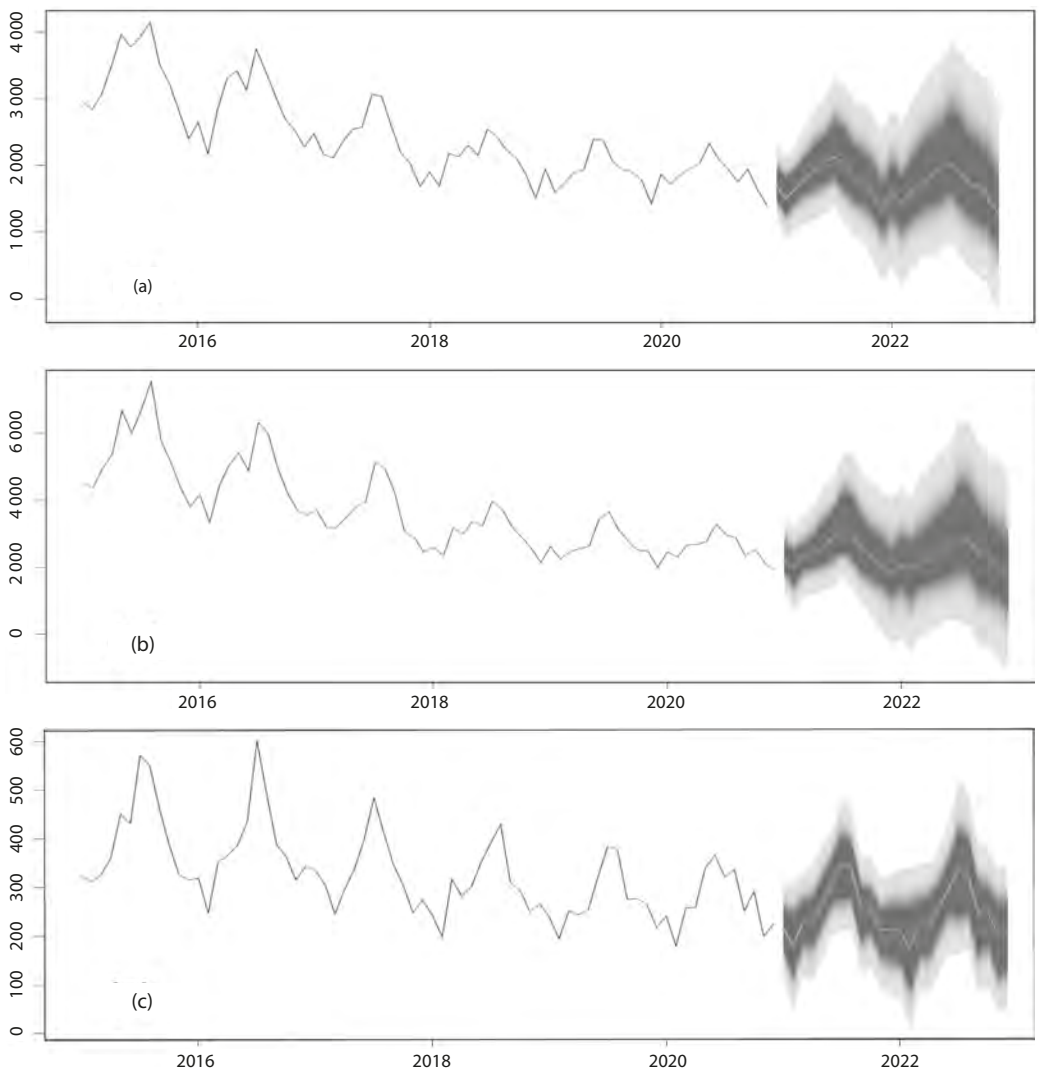
Table 4 Accuracy of the Hybrid models comparing with SARIMA models

Variables	Models	ME	RMSE	MAE	MPE	MAPE	ACF1
Accidents	Hybrid-M	–15.344	159.366	127.216	–1.2002	5.721	0.049
	SARIMA	19.963	179.506	129.932	1.161	5.814	0.407
Injuries	Hybrid-M	–44.197	268.981	209.042	–2.067	6.419	0.068
	SARIMA	15.121	329.972	229.096	0.972	6.973	0.383
Deaths	Hybrid-M	0.4302	33.4053	24.8803	0.014	7.969	0.614
	SARIMA	–5.3106	30.4623	23.2241	–3.031	7.921	0.006

Note: ME – Mean Error, **RMSE** – Root Mean Squared Error, **MAE** – Mean Absolute Error, **MPE** – Mean Percentage Error, **MAPE** – Mean Absolute Percentage Error, **ACF1** – Autocorrelation of errors at *lag 1*.
Source: Author's computation based on R program

Table 4 presents the accuracy measures of the optimal hybrid-models and the SARIMA models. The performances indicators (ME, RMSE, MAE, MPE, MAPE, and ACF1) showed that the hybrid models outperformed the SARIMA models for the three variables(accidents, injuries, and deaths). For the ARIMA models, the optimal ones were: a SARIMA$(1,1,1)(1,1,0)_{12}$ for the accident variable, and a SARIMA$(0,1,1)(2,1,0)_{12}$ for the injuries variable and a SARIMA$(1,0,0)(1,1,0)_{12}$ with drift for deaths variable, detailed characteristics of the selected SARIMA model are in the Appendix 1. The predictions estimated by the hybrid-models and SARIMA models were validated using the dataset of the last 12 months (in 2020). The models retained for forecasting are all hybrid-models; one based on equal weight for the accident variable and the two rest models based on in-sample errors (of RMSE indicator).

Figure 6 Hybrid-models forecasts for the number of accidents (a), number of Injuries (b), and number of deaths (c)



Note: The dashed blue surfaces correspond to the upper and lower bounds of confidence intervals of prediction at the $\alpha = 0.1$, and the dashed gray surfaces et the $\alpha = 0.05$ significance levels of predictions.
Source: Author's plot using R program

As can be seen in Figure 5, we expect an increase in the number of accidents and the number of injuries and deaths, and we expect that the number of accidents in August 2021 to be (on average) 2011 accidents, 2 945 injuries, and 335 deaths. The forecasts for the coming year (2022) follow nearly the same trajectories for the three variables.

5 DISCUSSION

In this study, a descriptive and predictive analysis was carried out on road traffic accidents in Algeria over the period (2015–2020), where the hybrid forecasting models were estimated and compared with the Box-Jenkins models. The findings revealed that the combined methods outperformed the SARIMA models, and we expect an increase in the number of accidents, number of deaths, and number of injuries over the next 12 months. Detailed statistics and estimation results have been presented, but the challenge is how to transform these figures into strategies.

This study presents, for the first time, the application of hybrid models to forecast the trajectories of road traffic accidents in Algeria. Our findings suggest the optimality of the hybrid models over the Box–Jenkins model. Compared with previous studies, this finding is broadly consistent with the study by Barba et al. (2014), which revealed the performance of combining Hankel matrix (HSVD)-ARIMA models with ARIMA models in forecasting traffic accidents in Chile. Similar results have been reported by Yusuf et al. (2015). Recently, Sangare et al. (2021) stated that the approach of the combination forecasting method was more accurate than baseline statistical methods in forecasting urban traffic accidents.

It was found that the number of deaths due to road traffic accidents in rural areas was four times higher than that in urban areas. This result is in good agreement with those of previous studies. For example, Cabrera-Arnau et al. (2020) explored road accident data from England and Wales, and reported that fatal crashes were more likely in rural areas than in urban areas. This pattern was demonstrated by Darma et al. (2017) in Malaysia, who revealed that the number of traffic fatalities in rural zones (66% of total deaths) was higher than that in urban zones. Accordingly, by exploring national surveillance data in China, Wang et al. (2019) showed that rural areas have higher road traffic mortality rates than urban areas do.

The results showed that men (regardless of age) were more likely to be involved in RTAs than were women; this finding is consistent with previous studies on this topic. The same finding was reported by Razi-Ardakani et al. (2018), who confirmed that men had a higher risk of road traffic accidents and attempted to analyze the factors behind sex differences in traffic accident severity. By analyzing data on road traffic accidents in Ecuador, Algora-Buenafé et al. (2017) indicated that 81.1% of fatal traffic accidents corresponded to men and 18% to women. Similarly, Wang et al. (2019) revealed that men in China had higher road accident mortality rates than women.

In terms of outlier analysis, there was a significant transient change in the total number of deaths after the lockdown due to the Covid-19 pandemic; However, this result was not conclusive in the case of Algeria. By contrast, recent studies in other countries have demonstrated the effects of pandemics on road safety. For example, Katrakazas et al. (2020) reported that the number of road accidents in Greece was reduced by 41% during the lockdown. In the same issue, Saladié et al. (2020) stated that the daily number of accidents was reduced by 74.3% during the period of lockdown in Tarragona province, Spain.

CONCLUSION

Summing up the results, it can be concluded that this study has shown the past and future dynamics of road traffic accidents in Algeria in terms of the number of accidents, injuries, and deaths. The optimality of the hybrid models over the Box-Jenkins model for forecasting RTAs is demonstrated. However, further studies are required to determine the optimal number of forecasting methods for the combined process. Several other questions remain to be addressed in order to better understand the pattern of road traffic

in Algeria. Specifically, future studies using regression models should be useful for estimating the effects of vehicle characteristics, road conditions, driver characteristics, and weather conditions on the dynamics of road traffic accidents in Algeria.

ACKNOWLEDGEMENT

This research received no specific grant from any funding agency, commercial or not-for-profit sectors. The author would like to thank the responsible of the Délégation Nationale à la Sécurité Routière (DNSR) to provide us the access to the dataset and other information on road Accidents. We would like to thank the two anonymous reviewers for helpful comments and improvements.

References

ABDOLLAHI, H. (2020). A novel hybrid model for forecasting crude oil price based on time series decomposition. *Applied Energy*, 267: 115035.

AKAIKE, H. (1974), A new look at the statistical model identification [online]. *IEEE Transactions on Automatic Control*, 19(6): 716–723. <https://doi.org/10.1109/TAC.1974.1100705>.

ALGORA-BUENAFÉ, A. F., SUASNAVAS-BERMÚDEZ, P. R., MERINO-SALAZAR, P., GÓMEZ-GARCÍA, A. R. (2017). Epidemiological study of fatal road traffic accidents in Ecuador. *Australasian Medical Journal*, 10(3): 238.

ARMSTRONG, J. S. (2001). Combining forecasts. In: *Principles of forecasting*, Boston, MA: Springer: 417–439.

ASSIMAKOPOULOS, V., NIKOLOPOULOS, K. (2000). The theta model: a decomposition approach to forecasting. *International journal of forecasting*, 16(4): 521–530.

BARBA, L., RODRÍGUEZ, N., MONTT, C. (2014). Smoothing strategies combined with ARIMA and neural networks to improve the forecasting of traffic accidents. *The Scientific World Journal*.

BARDAL, K. G., JØRGENSEN, F. (2017). Valuing the risk and social costs of road traffic accidents – Seasonal variation and the significance of delay costs. *Transport Policy*, 57: 10–19.

BATES, J. M., GRANGER, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20(4): 451–468.

BOX, G., JENKINS, G. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

CABRERA-ARNAU, C., PRIETO CURIEL, R., BISHOP, S. R. (2020). Uncovering the behaviour of road accidents in urban areas. *Royal Society open science*, 7(4): 191739.

CHEN, C., LIU, L.-M. (1993). Joint Estimation of Model Parameters and Outlier Effects in Time Series [online]. *Journal of the American Statistical Association*, 88(421): 284–297. <https://doi.org/10.2307/2290724>.

CHEN, S., KUHN, M., PRETTNER, K., BLOOM, D. E. (2019). The global macroeconomic burden of road injuries: Estimates and projections for 166 countries. *The Lancet Planetary Health*, 3(9): e390–e398.

DARMA, Y., KARIM, M. R., ABDULLAH, S. (2017). An analysis of Malaysia road traffic death distribution by road environment. *Sādhanā*, 42(9): 1605–1615.

DE LIVERA, A. M., HYNDMAN, R. J., SNYDER, R. D. (2011), Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496): 1513–1527.

DOANE, D. P., SEWARD, L. E. (2011). Measuring skewness: a forgotten statistic? *Journal of Statistics Education*, 19(2).

GRANGER, C. W., RAMANATHAN, R. (1984). Improved methods of combining forecasts. *Journal of forecasting*, 3(2): 197–204.

HAMILTON, J. D. (2020). *Time series analysis*. Princeton university press.

HILL, T., O'CONNOR, M., REMUS, W. (1996). Neural network models for time series forecasts. *Management Science*, 42(7): 1082–1092.

HYLLEBERG, S., ENGLE, R. F., GRANGER, C. W., YOO, B. S. (1990). Seasonal integration and cointegration. *Journal of Econometrics*, 44(1–2): 215–238.

HYNDMAN, R. J., KOEHLER, A. B., SNYDER, R. D., GROSE, S. (2002). A state space framework forautomatic forecasting using exponential smoothing methods. *International J. Forecasting*, 18(3): 439–454.

HYNDMAN, R. J., ATHANASOPOULOS, G. (2018). *Forecasting: principles and practice*. OTexts.

IHUEZE, C. C., ONWURAH, U. O. (2018). Road traffic accidents prediction modelling: an analysis of Anambra State, Nigeria. *Accident Analysis & Prevention*, 112: 21–29.

KATRAKAZAS, C., MICHELARAKI, E., SEKADAKIS, M., YANNIS, G. (2020). A descriptive analysis of the effect of the COVID-19 pandemic on driving behavior and road safety. *Transportation research interdisciplinary perspectives*, 7: 100186.

LEAN, Y., SHOUYANG, W. A. N. G., LAI, K. K., NAKAMORI, Y. (2005). Time series forecasting with multiple candidate models: selecting or combining? *Journal of Systems Science and Complexity*, 18(1): 1–18.
LÓPEZ DE LACALLE, J. (2019). *Tsoutliers: Detection of Outliers in Time Series* [online]. R package version 0.6-8. <https://CRAN.R-project.org/package=tsoutliers>.
MAYOU, R., BRYANT, B., DUTHIE, R. (1993). Psychiatric consequences of road traffic accidents. *British Medical Journal*, 307(6905): 647–651.
ONS. (2021). *Parc. Automobile* [online]. Office National des Statistiques, Algeria. <https://www.ons.dz/IMG/pdf/e.nat31_12_2018.pdf>.
RAZI-ARDAKANI, H., ARIANNEZHAD, A., KERMANSHAH, M. (2018). A Study of Sex Differences on Road Crash Severity. *Proceedings of the 3rd International Conference on Civil, Structural and Transportation Engineering (ICCSTE'18)*.
REZAIE MOGHADDAM, F., AFANDIZADEH, S., ZIYADI, M. (2011). Prediction of accident severity using artificial neural networks. *International Journal of Civil Engineering*, 9(1): 41–48.
RONDEROS, N. (2019). *Seasonal Unit Root Tests* [online]. <http://blog.eviews.com/2019/04/seasonal-unit-root-tests.html>.
SALADIÉ, Ò., BUSTAMANTE, E., GUTIÉRREZ, A. (2020). COVID-19 lockdown and reduction of traffic accidents in Tarragona province, Spain. *Transportation research interdisciplinary perspectives*, 8: 100218.
SANGARE, M., GUPTA, S., BOUZEFRANE, S., BANERJEE, S., MUHLETHALER, P. (2021). Exploring the forecasting approach for road accidents: Analytical measures with hybrid machine learning. *Expert Systems with Applications*, 167: 113855.
SHAUB, D., ELLIS, P. (2020). *Forecast Hybrid: Convenient Functions for Ensemble Time Series Forecasts* [online]. R package version 5.0.19. <https://CRAN.R-project.org/package=forecastHybrid>.
WACHNICKA, J., PALIKOWSKA, K., KUSTRA, W., KIEC, M. (2021). Spatial differentiation of road safety in Europe based on NUTS-2 regions. *Accident Analysis & Prevention*, 150: 105849.
WANG, L., ZOU, H., SU, J., LI, L., CHAUDHRY, S. (2013). An ARIMA-ANN hybrid model for time series forecasting. *Systems Research and Behavioral Science*, 30(3): 244–259.
WANG, L., NING, P., YIN, P., CHENG, P., SCHWEBEL, D. C., LIU, J., HU, G. et al. (2019). Road traffic mortality in China: Analysis of national surveillance data from 2006 to 2016. *The Lancet Public Health*, 4(5): e245–e255.
WHO. (2021, June). *Road traffic injuries* [online]. Geneva: World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>.
YUSUF, S. M., MU'AZU, M. B., AKINSANMI, O. (2015). A Novel Hybrid Fuzzy Time Series Approach with Applications to Enrollments and Car Road Accidents. *International Journal of Computer Applications*, 129(2): 37–44.
YANG, Y. (2004). Combining forecasting procedures: Some theoretical results. *Econometric Theory*, 20(1): 176–222.
ZHANG, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50: 159–175.

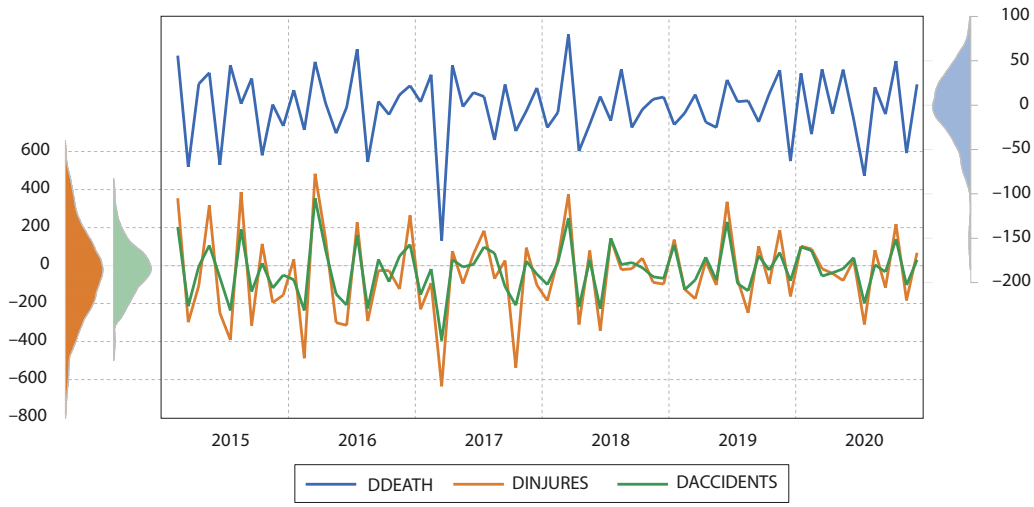
APPENDICES

Appendix 1 Characteristics of SARIMA models

Model	Variable	AIC	AICc	BIC	Variance	Log-Likelihood
ARIMA(1,1,1)(1,1,0)$_{12}$	Accidents	802.65	803.39	810.96	41429	–397.33
ARIMA(0,1,1)(2,1,0)$_{12}$	Injuries	876.03	876.77	884.34	139 991	–434.02
ARIMA(1,0,0)(1,1,0)$_{12}$ with drift	Deaths	613.98	614.71	622.36	1 410	–302.99

Note: SARIMA model is defined as: $ARIMA \underbrace{(p, d, q)}_{\text{(Non-seasonal part of the model)}} \underbrace{(P, D, Q)_m}_{\text{(Non-seasonal part of the model)}}$, where *p*: Trend Autoregressive component, *d*: Difference order, *q*: Moving Average component. The seasonal part of the model is designed as *P*: Seasonal Autoregressive component, *D*: Seasonal Difference order, *Q*: Seasonal Moving Average component, *m*: the frequency of the time series; here *m* = 12 which means monthly data and it can exhibits an annual seasonal cycle.
Source: Own construction

Appendix 2 Plot of stationary time series with kernel densities in the axis borders



Source: Own construction

Appendix 3 Summary of outliers' analysis

Variable	*id*	Type(*)	Time	Coef.	T-stat
Accident	1	AO	2016:02:00	−318.9	−3.394
	2	AO	2016:07:00	356.9	3.700
	3	TC	2017:03:00	−460.2	−3.639
	4	IO	2019:06:00	616.4	3.579
Injuries	1	SLS	2016:08:00	−972.5	−3.654
	2	TC	2017:03:00	−949.6	−3.969
	3	LS	2017:10:00	−779.1	−3.274
Deaths	1	LS	2017:03:00	−92.84	−8.440
	2	TC	2020:07:00	−106.34	−3.298

Note: (*) "AO" additive outliers, "LS" level shifts, "TC" temporary changes, "IO" innovative outliers and "SLS" seasonal level shifts.
Source: Own construction

Can Individual Human Financial Behaviour Be Mathematically Modelled? A Case Study of Elon Musk's Dogecoin Tweets

Juraj Medzihorský[1] | *Matej Bel University, Banská Bystrica, Slovakia*
Peter Krištofík | *Matej Bel University, Banská Bystrica, Slovakia*

Received 11.2.2022, **Accepted** (reviewed) 23.3.2022, **Published** 17.6.2022

Abstract

The price of Dogecoin has been influenced by Elon Musk's tweets on several occasions. Moreover, there are repeating patterns in the Dogecoin prices. However, is there also a pattern to the timing of the tweets? Applying linear regression, we have been able to make the reverse analysis – to use hard financial data (prices) to analyse the human behaviour (tweets) that preceded and influenced the financial data. Selected tweets could be paired thanks to the projections of their timing on the regression line that had been created over the prices. Our model exhibits inaccuracies only in the order of the days. That is surprising, as pump schemes do not usually require such a high level of long-term deterministic timing.

Keywords	*DOI*	*JEL code*
Behavioral economics, cryptocurrencies, pump-and-dump scheme, linear regression, time series analysis	*https://doi.org/10.54694/stat.2022.9*	*G17, G41*

INTRODUCTION

Human behaviour has become an important component of financial market research, in addition to hard financial data. The proximate determinants of stock prices are supply and demand. That is, human activity is affected by sentiment as well as by firms` results. A clear example was the rise in the value of GameStop, caused by Reddit users (Morgia et al., 2021). Cryptocurrencies, and especially memecoins are more influenced by sentiment than are stocks, because they lack an agreed valuation standard. Social media amplifies these effects.

[1] Matej Bel University, Faculty of Economics, Department of Finance and Accounting, Tajovského 10, 975 90 Banská Bystrica, Slovakia. Corresponding author: e-mail: juraj.medzihorsky@umb.sk, phone: (+421)484466111.

Many studies have focussed on modelling human behaviour. For example, Pentland (2006), Aipperspach et al. (2006), Lieder and Griffiths (2019). It turns out that some features can apply to both economic and non-economic behaviour. For example, Aipperspach et al. (2006) showed that the highly-skewed power-law distribution could model human movements in a house. The same distributions are also suitable for modelling company size (Lyócsa and Výrost, 2018), and also transactions between crypto wallets. Anomalies from these expected distributions can be caused by non-human activity – in the case of crypto wallets – by trading-bots (Zwang et al., 2018).

The use of trading-bots is not new. But they can now be much more efficient thanks to pump-and-dump schemes. These are common in cryptocurrency markets (Kamps and Kleinberg, 2018; Xu and Livshits, 2019). And so is the impact of influencers like Elon Musk. The relations between his tweets and the price of bitcoin was confirmed by Tandon et al. (2021), and for Dogecoin by Cary (2021). His tweets can have an almost immediate major impact on the price of any altcoin. Occasionally, such impact can happen regardless of the intention. A quite bizarre recent example was Elon Musk's tweet that mentioned J. R. R. Tolkien's idea about free public ducks (Twitter, 2022). This caused a significant price rise in the homonymous cryptocurrency.

The aim of our paper is to model the timing of Elon Musk's Dogecoin tweets. Our hypotheses are:

H1: Behaviour of a market-influencer can be reversely derived from asset price movements.

H2: There is a pattern to the timing of Elon Musk's tweets.

1 MATERIAL AND METHODS

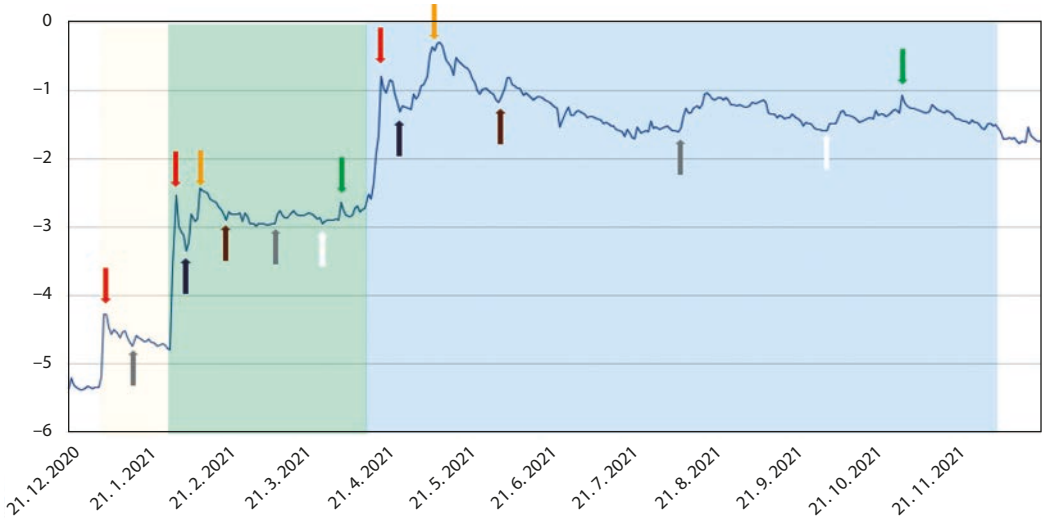
The methods used in the analysis are linear regression, and exploratory analysis. There are two datasets (Coindesk, 2022; Twitter, 2022): the time series of Dogecoin prices (daily maximums of DOGE/USD) since the large pump event of 28/01/2021, and the time series of Elon Musk's Dogecoin-related tweets during the same period. Daily maxima are used, as we intend to analyse pumps. During modelling all dates are converted to simple numbers. For example, 28/01/2021 = 1, and so on. We begin by setting out the time model of Dogecoin prices. As Figure 1 shows, there are similarities in price developments during Period 2 (blue area), compared to Period 1 (green area), and to Period 0 (white area). There are similarities in the patterns of peaks and troughs, though the length of the periods and the size of changes both increase over time. The similarities highlighted in alternative way are showed in Figure A1 in the Annex. This replication was also confirmed in our previous research, with 87% accuracy on a 3-month test set (Medzihorský, 2021). The accuracy was defined as 1 – γ, where γ is a mean error of prediction, calculated as follows:

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|Predicted\ Price_i - Actual\ Price_i\right|}{Actual\ Price_i}.$$

As the predictions on the test set were calculated in that research only from historical prices by a simple equation, such high accuracy supports the assumption about replication. This is consistent with the results of Uras et al. (2020), who confirm the existence of time regimes for selected cryptocurrencies. The regimes differ from a random walk: for the best forecasts they recommend taking 200-day sequences. In addition, Ozdamar et al. (2021) confirmed high correlations between expected returns on crypto-currencies and daily maxima during the previous month. So crypto-markets are not purely stochastic. They exhibit some level of a determinism.

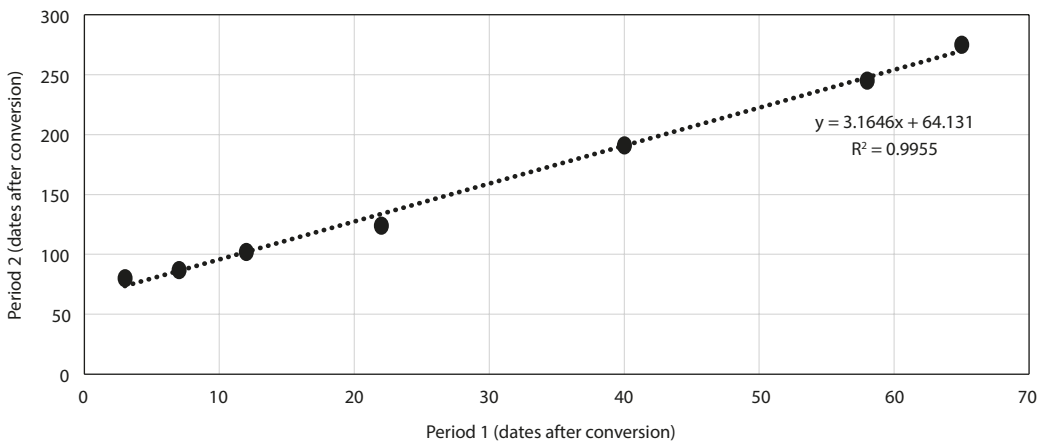
If we assume that price developments during different periods follow the same general pattern (see arrows on Figure 1), then combining points from Periods 1 and 2 we can produce the time series regression in Figure 2. The selected points represent local minimums, maximums, or high daily yields. This model will help predict the timing of tweets. The final step of the analysis is a simple projection of the dates of tweets on the same regression line that was created in the time series model of price development.

Figure 1 Logarithmic price of Dogecoin with highlighted replication of the shapes



Note: Logarithmic scale (using natural logarithm) is used for clearer illustration of the shapes replication that is not obvious on a figure with the linear scale.
Source: Own processing from Coindesk (2022), and Medzihorský (2021)

Figure 2 Regression of selected time points



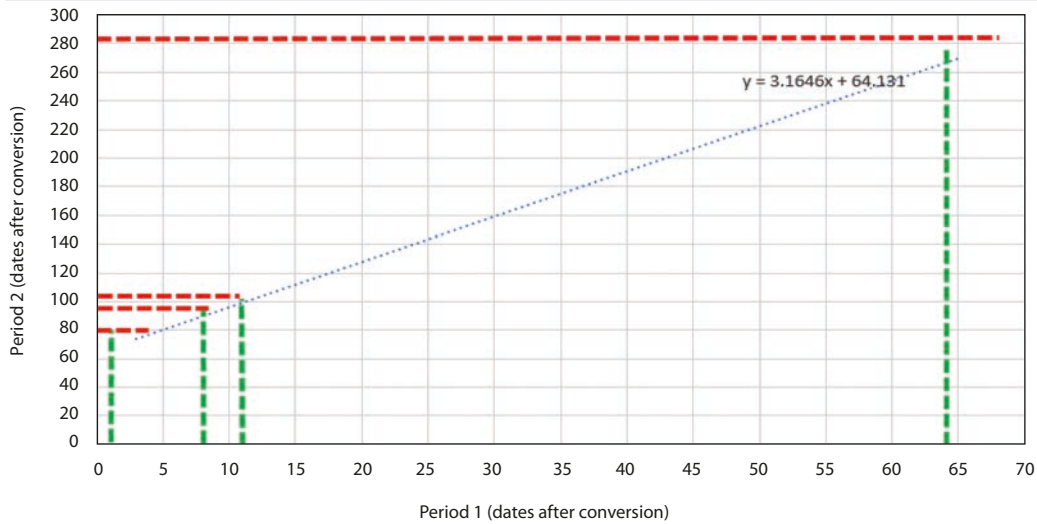
Note: Dates are converted to simple numbers. For example, 28/01/2021 = 1, and so on. The selected time points in Figure 2 are represented by the arrows in Figure 1.
Source: Own processing

2 RESULTS

There is almost a perfect correlation between the timing of selected points in Period 2 and Period 1 (see Figure 2). We can confirm that the timing of selected movements of Dogecoin price is significantly deterministic, and the lengthening of shapes is linear. However, this model is only auxiliary. We use it to find a suitable line to fit the timing of the tweets. Doing so, we also observe the determinism in the timing of the tweets (see Figure 3). As the intersections of the projections of tweet dates are approximately on the regression line, the hypotheses H1 and H2 are confirmed.

However, our approach has important limitations. Only selected tweets can be analyzed this way – there are more tweets in Period 2 than in Period 1, so some cannot be paired together. Only a limited period is studied. Finally, the exact timing cannot be calculated by a simple line – there are some inaccuracies in the order of the days – see Figure 3 and Table 1.

Figure 3 Timing of the tweets and their projection on the regression line of Dogecoin time points



Source: Own processing from Twitter (2022)

Applying the equation y = 3.1646x + 64.131 (Formula 1) to the converted dates of the tweets, produces the predictions shown in Table 1. The table shows some repeating inaccuracies. Predictions can be improved by incorporating this knowledge. We estimate the next tweet will be on 12 February 2022.

Table 1 Tweet predictions based on Formula 1

Date of original tweet	Converted date *(x)*	Future tweet estimate *(y)*	Re-converted date of estimate	Real timing of future tweet	Inaccuracy
1/28/2021	1th	67	4/4/2021	4/15/2021	11
2/4/2021	8	89	4/26/2021	4/28/2021	2
2/7/2021	11	99	5/6/2021	5/7/2021	1
4/1/2021	64	267	10/21/2021	10/27/2021	6
4/15/2021	78	311	12/4/2021	12/14/2021	10
4/28/2021	91	352	1/14/2022	1/14/2022	0
5/7/2021	100	381	2/12/2022	N/A	N/A

Note: Twitter should show date and time in the user's time zone. In our case GMT+1 or 2 (depending on summer/winter time).
Source: Own processing from Twitter (2022)

Not only timing but also the quality and price impact of the tweets play a role (see Table 2). While the permanent impact of some earlier tweets – which led to a several-fold increase in price – could be valuable for Dogecoin holders, current tweets cause only pumps and dumps. There may be several reasons for this. First, the crypto market as whole is not currently achieving yields as high as it was, up to, say,

May 2021, when Dogecoin recorded its all-time-high. Second, Dogecoin tweets are more common. They are not novel. Third, alternative dog-based memecoins have been recently created. Perhaps surprisingly, even though recent Dogecoin tweets have included more economic context, this has not changed market reactions. In fact, memecoins depend on market sentiment. So, a tweet that includes some genuinely informed economic analysis does not necessarily change prices more than one without it. Also, one has to wonder if the supply of new Dogecoin fans has been exhausted – and existing fans` demands are satiated. On the other hand, demand for Dogecoin or any cryptocurrency can be positively influenced by the growing inflation as the inflation negatively affects, especially, cash holdings (Pintér and Mešťan, 2020).

There has also been a change in the form of the tweets. Pictures are no longer used. Putting words like 'doge' or 'Dogecoin' directly in the text can be more profitable for traders using bots than for others.

Table 2 Qualitative value and impact of the tweets

Date	Quality of information	Form	Price impact	Duration of impact
1/28/2021	N-E	Text in picture	867%	Permanent
2/4/2021	N-E	Multiple tweets	90%	Permanent
2/7/2021	N-E	Picture	86%	Temporary
4/1/2021	D	Text	31%	Permanent
4/9/2021	N-E	Picture	8%	Permanent
4/15/2021	D	Text + picture	387%	Permanent
4/28/2021	N-E	Text	30%	Temporary
5/7/2021	N-E	Picture	15%	Pump-and-dump
5/11/2021	E	Text	-2%	Pump-and-dump
5/14/2021	E	Text	31%	Pump-and-dump
5/20/2021	N-E	Text + picture	0%	Pump-and-dump
6/2/2021	N-E	Picture	37%	Pump-and-dump
7/2/2021	N-E	Text in picture	4%	Nearly zero effect
7/25/2021	E	Text in picture	21%	Pump-and-dump
9/22/2021	E	Text	7%	Pump-and-dump
10/27/2021	D	Text	28%	Pump-and-dump
10/31/2021	E	Text	2%	Nearly zero effect
12/14/2021	E	Text	26%	Pump-and-dump
12/23/2021	E	Text	14%	Pump-and-dump
1/14/2022	E	Text	33%	Pump-and-dump

Note: N-E – non-economic information; E – tweets with serious economic information like an acceptance of Dogecoin by a merchant; D – economic value depends on wider context. For example, putting literal Dogecoin on the literal Moon would not be economic relevant. However, it actually is relevant, as lunar cargo for DOGE-1 mission is financed by Dogecoin. Permanent price impact means that price has not declined lower than it was before the tweet; Temporary impact means that the price remained higher than before the tweet for one or more weeks. Pump-and-dump represents a quick decline within a week of the pump. Price impact is a yield, calculated as follows: Yield = max {Daily high price $_t$, Daily high price $_{t+1}$, Daily high price $_{t+2}$} / Closing price $_{t-1}$ – 1; where *t* represents the date of the tweet.
Source: Own processing from Twitter (2022), and Coindesk (2022)

An important limitation of the calculation of price impact (see Table 2) – using a comparison with the closing price of the previous day – lies in price changes shortly before a tweet. An example is the tweet on 14 January 2022, when there was a significant rise in price one hour before the tweet. Contrary negative examples are the tweets on 11 and 20 May, when a decline of price before the tweets distorted the calculations. The actual effects of these tweets were, of course, positive.

Our results raise several questions. Only selected tweets from Period 2 can be paired with the tweets from Period 1. These tweets from Period 2 can be paired with most recent tweets. However, do the rest of the tweets from Period 2 – which cannot be paired with the tweets in Period 1 – have the same predictive power? Will there ever again be a tweet that causes a yield of more than 100%, with a permanent price impact. Or will we only observe pump-and-dumps, with no more than 30% yields. Can past inaccuracies in our model be used to achieve more precise predictions, or to analyse a wider range of cases? What are the reasons, if any, for the price increase before the tweet on 14 January 2022? As these questions remain unanswered, the need for continuing research is clear.

CONCLUSION

It is obvious that Elon Musk influences the price of Dogecoin. So, we have been able to reverse analyse his behaviour – to use hard financial data (prices) to analyse the human behaviour (tweets) that preceded and influenced financial data. If there had been no repeating patterns in Dogecoin prices, or in the timing of the tweets, we would have been unable to model the timing of human behaviour by a simple line. However, what motivation, if any, might lie behind the timing of the tweets, remains hidden. Pumps do not require such a high level of long-term deterministic timing. Nor is it clear that the pattern of the timing of the tweets could partly determine the long-term price development of Dogecoin. Our contribution is only a first step in this analysis. Therefore, the paper is intentionally structured using a single-issue approach, as we expect wider discussion and further research of this issue in the future. The aim of more complex studies should include the analysis and prediction of the behaviour of ordinary traders from price movements, the analysis of other market influencers, and searching for any deterministic trends in such areas where stochasticity would be usually expected.

ACKNOWLEDGMENT

The authors wish to thank Dr. Colin W. Lawson from University of Bath for valuable comments and corrections of the manuscript.

References

AIPPERSPACH, R., COHEN, E., CANNY, J. (2006). Modelling Human Behaviour from Simple Sensors in the Home [online]. In: FISHKIN, K. P., SCHIELE, B., NIXON, P., QUIGLEY, A. (eds.) *Pervasive Computing*, Pervasive, Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, Vol. 3968. <https://doi.org/10.1007/11748625_21>.

CARY, M. (2021). Down with the #Dogefather: Evidence of a Cryptocurrency Responding in Real Time to a Crypto-Tastemaker [online]. *Journal of Theoretical and Applied Electronic Commerce Research*, 16: 2230–2240. <https://doi.org/10.3390/ jtaer16060123>.

COINDESK. (2022). *Dogecoin* [online]. <https://www.coindesk.com/price/dogecoin>.

KAMPS, J., KLEINBERG, B. (2018). To the moon: defining and detecting cryptocurrency pump-and-dumps [online]. *Crime Science*, 7(18): 1–18. <https://doi.org/10.1186/s40163-018-0093-5>.

LIEDER, F., GRIFFITHS, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources [online]. *Behavioral and Brain Sciences*, 43(1): 1–60. <https://doi.org/10.1017/S0140525X1900061X>.

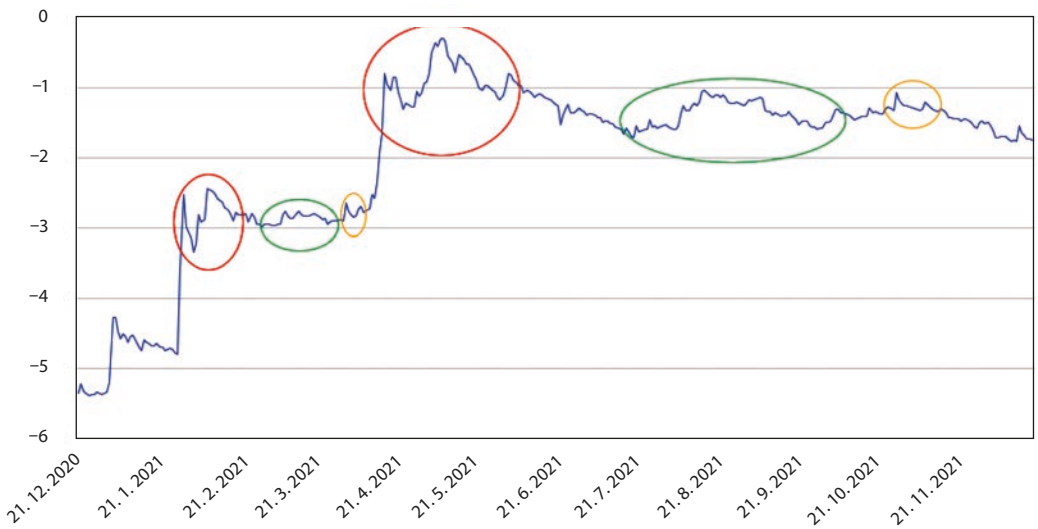
LYÓCSA, Š., VÝROST, T. (2018). Scale-free distribution of firm-size distribution in emerging economies [online]. *Physica A*, 508: 501–505. <https://doi.org/10.1016/j.physa.2018.05.088>.

MEDZIHORSKÝ, J. (2021). Dogecoin price prediction – can be a determinism supposed? [online]. *Journal of Economics and Social Research* 22(2): 67–81. <https://doi.org/10.24040/eas.2021.22.2.67-81>.

MORGIA, L., MEI, A., SASSI, F., STEFA, J. (2021). *The Doge of Wall Street: Analysis and Detection of Pump and Dump Cryptocurrency Manipulations* [online]. Ithaca, NY: Cornell University. <https://arxiv.org/abs/2105.00733v1>.
OZDAMAR, M., AKDENIZ, L., SENSOY, A. (2021). Lottery like preferences and the MAX effect in the cryptocurrency market [online]. *Financial Innovation,* 7(74): 1–27. <https://doi.org/10.1186/s40854-021-00291-9>.
PENTLAND, A. (2007). Automatic mapping and modeling of human network [online]. *Physica A*, 378: 59–67. <https://doi.org/10.1016/j.physa.2006.11.046>.
PINTÉR, Ľ., MEŠŤAN, M. (2020). *Kolektívne investovanie*. 1st Ed. Banská Bystrica: Belianum.
TANDON, C., REVANKAR, S., PALIVELA, H., PARIHAR, S. (2021). How can we predict the impact of the social media messages on the value of cryptocurrency? Insights from big data analytics [online]. *International Journal of Information Management Data Insights*, 1: 100035. <https://doi.org/10.1016/j.jjimei.2021.100035>.
TWITTER. (2022). *Elon Musk* [online]. <https://twitter.com/elonmusk>.
URAS, N., MARCHESI, L., MARCHESI, M., TONELLI, R. (2020). Forecasting Bitcoin closing price series using linear regression and neural networks models [online]. *Peer J. Computer Science*, 6: e27. <https://doi.org/10.7717/peerj-cs.279>.
XU, J., LIVSHITS, B. (2019). *The Anatomy of a Cryptocurrency Pump-and-Dump Scheme* [online]. Ithaca, NY: Cornell University. <https://arxiv.org/abs/1811.10109>.
ZWANG, M., SOMIN, S., PENTLAND, A., ALTSHULER, Y. (2018). *Detecting Bot Activity in the Ethereum Blockchain Network* [online]. Ithaca, NY: Cornell University. <https://arxiv.org/abs/1810.01591>.

ANNEX

Figure A1 Logarithmic price of Dogecoin with alternatively highlighted replication of the shapes



Note: Logarithmic scale (using natural logarithm) is used for clearer illustration of the shapes replication that is not obvious on a figure with linear scale.
Source: Own processing from Coindesk (2022)

New Randomized Response Technique for Estimating the Population Total of a Quantitative Variable

Jaromír Antoch[1] | *Charles University, Prague, Czech Republic*
Francesco Mola[2] | *Università di Cagliari, Cagliari, Italy*
Ondřej Vozár[3] | *Prague University of Economics and Business, Prague, Czech Republic*

Received 10.2.2022 (revision received 30.3.2022) **Accepted** (reviewed) 22.4.2022 **Published** 17.6.2022

Abstract

A new randomized response technique for estimating the population total, or the population mean of a quantitative variable is proposed. It provides a high degree of protection to the respondents because they never report their data. Therefore, it may be favorably perceived by them and increase their willingness to cooperate. Instead of revealing the true value of the characteristic under investigation, the respondent only states whether the value is greater (or smaller) than a number which is selected by him/her at random and is unknown to the interviewer. For each respondent, this number, a sort of individual threshold, is generated as a pseudorandom number. Furthermore, two modifications of the proposed technique are presented. The first modification assumes that the interviewer also knows the generated random number. The second modification deals with the issue that, for certain variables, such as income, it may be embarrassing for the respondents to report either high or low values. Thus, depending on the value of the fixed threshold (unknown to the respondent), the respondent is asked different questions to avoid being embarrassed. The suggested approach is applied in detail to the simple random sampling without replacement, but it can be, after a straightforward modification, applied to many sampling schemes, including cluster sampling, two-stage sampling, or stratified sampling. The results of the simulations illustrate the behavior of the proposed technique.

Keywords	DOI	JEL
Survey sampling, population total, Horvitz-Thompson's estimator, randomized response techniques, simple random sampling	*https://doi.org/10.54694/stat.2022.11*	*C83, J30*

[1]Charles University, Fac. of Mathematics and Physics, Sokolovská 83, CZ – 186 75 Prague 8 – Karlín, Czech Republic and Prague University of Economics and Business, Fac. of Informatics and Statistics, W. Churchill Sq. 4, CZ – 130 67 Prague 3, Czech Republic.

[2]Università di Cagliari, Fac. di Economia, viale S. Ignazio da Laconi 17, I – 09123 Cagliari, Italy.

[3]Prague University of Economics and Business, Fac. of Informatics and Statistics, W. Churchill Sq. 4, CZ – 130 67 Prague 3, Czech Republic and Czech Statistical Office, Na padesátém 3268/81, CZ – 100 82 Prague 10, Czech Republic.

Editorial note: The article was written in TEX and graphically prepared for print by the authors.

INTRODUCTION

A steady decline in response rates has been reported in many surveys in most countries around the world, see, e.g., Steeh (2001) or Stoop (2005). This decline is observed regardless of the mode of the survey, e.g., face-to-face survey, paper/electronic questionnaire, internet survey, or telephone interviewing. Furthermore, this trend has continued despite additional procedures aimed at reducing refusal and increasing contact rates; see Brick (2013) among others.

The growing concern about "invasion of privacy" therefore represents an important challenge for statisticians. Quite naturally, a respondent may be hesitant or even evasive in providing any information which may indicate a deviation from a social or legal norm and/or which he/she feels that might be used against him/her some time later. Therefore, if we ask sensitive or pertinent questions in a survey, conscious reporting of false values would often occur, see Särndal et al. (1992:547). Unfortunately, standard techniques such as reweighting or model-based imputation cannot usually be applied; for a detailed discussion see Särndal et al. (1992:547) or Särndal and Lundström (2005). On the other hand, this issue can be resolved, at least partially, using randomized response techniques (RRT). Comprehensive information on the broad scope of methods and theoretical foundations of RRT can be found in Chaudhuri (2017), Chaudhuri and Christofides (2013), Chaudhuri and Mukerjee (1988), Fox (2016) or Chaudhuri et al. (2016) among others.

For all of the reasons mentioned above, different RRTs have been developed with the goal of reducing the nonresponse rate and obtaining unbiased estimates. These techniques began with a seminal paper Warner (1965), aimed at estimating the proportion of people in a given population with sensitive characteristics, such as substance abuse, unacceptable behavior, criminal past, controversial opinions, etc. In Eriksson (1973) and in Chaudhuri (1987) the authors modified Warner's method to estimate the population total of a quantitative variable. However, in our opinion, these "standard RRTs" aimed at estimating the population total are rather complicated and demanding on both the respondents and the survey statisticians for various real-life applications; see also the discussion in Chaudhuri (2017). They require "nontrivial arithmetic operations" from respondent within the Chaudhuri's approach, while the survey statistician must expend a lot of effort related with the design of suitable randomization devices to be used for masking the sensitive variables in the Eriksson's approach.

Despite their advantages, practically all RRTs suffer from larger or smaller limitations, especially in the following.

- Lack of reproducibility.
- Lack of trust from respondents because the randomization device is controlled by the interviewer.
- Higher cost and higher variance of the estimators due to the use of random devices.

To avoid at least partially these limitations, already long time ago the statisticians suggested other approaches not requiring any random devices, These so-called *non-randomized response (NRR)* techniques are typically based on auxiliary questions, instead on random devices, and their alternative designs include, but are not limited to, unrelated question design, contamination design, multiple trials, and quantitative data design. Recently, researchers revitalized these ideas; see a series of papers by Tang, Tian, Wu, and their followers. To the best of our knowledge, they concentrated mainly on estimating proportions, not the totals. The NRR techniques are presented in detail in the monograph Tian and Tang (2014).

In any case, when suggesting any randomization device, we should always keep in mind that the main issue is not whether the in-person interviewer or telephone interviewer knows

the random numbers or the outcome of other random mechanisms used, but whether the random number is given back to the researcher evaluating the survey or to the survey sponsor. Personally, we prefer that the interviewer checks the methodology, not the realization itself.

Finally, we would like to point out that the question of credibility is not only a matter for statisticians, but more and more a task for psychologists. While statisticians must suggest procedures that are "sufficiently random" in their eyes, psychologists must find and offer ways to convince the respondents that they are not cheated. Unfortunately, a detailed discussion of this topic would go beyond the scope of this paper.

In this paper, we propose a method which is simpler in comparison with those proposed previously and which is practically applicable. The respondent is only asked whether the value of a sensitive variable reaches at least a certain random lower bound. This technique and its modifications are developed in detail and applied to simple random sampling without replacement. Their pros and cons are thoroughly discussed and illustrated using simulations.

The main advantages of the suggested method include the ease of implementation, simple use by the respondent, and practically acceptable precision. Moreover, respondents' privacy is well protected because they never report the true value of the sensitive variable. Unlike in Chaudhuri's or Eriksson's approach, there is no issue with the physical random device design. A disadvantage may be, from a certain point of view, a lower degree of confidence in anonymity due to the extrinsic device/technique used for generating random numbers.

The paper is organized as follows. In Section 1, selected issues of the RRTs for the estimation of the population total, or population mean, are concisely discussed. In Section 2, a new randomized response technique and its two modifications are proposed, their properties studied and the goals for future work summarized. Section 3 illustrates the suggested ideas with the aid of a simulation study. The main conclusions of the paper follow.

1 SELECTED REMARKS ON RRT INTENDED TO ESTIMATE POPULATION TOTAL AND THEIR PROPERTIES

Consider a finite population $U = \{1, \ldots, N\}$ of N identifiable units, where each unit can be unambiguously identified by its label. Let Y be a sensitive quantitative variable. The objective of the survey is to estimate the population total $t_Y = \sum_{i \in U} Y_i$ or, alternatively, the population mean $\overline{t}_Y = t_Y/N$, of the variable surveyed. To do this, we use a random sample s selected with probability $p(s)$, described by a sampling plan with a fixed sample size n. Let us denote by π_i the probability of inclusion of the i^{th} element in the sample, that is, $\pi_i = \sum_{s \ni i} p(s)$, and by ξ_i the indicator of inclusion of the i^{th} element in the sample s, i.e., $\xi_i = 1$ if $s \ni i$ and $\xi_i = 0$ otherwise. We do not introduce all notions from scratch and refer the reader to Särndal et al. (1992:547) or the more rigorous monograph Tillé (2006).

As argued above, in practice it is often impossible to obtain the values of the surveyed variable Y in sufficient quality due to its sensitivity. Therefore, statisticians try to obtain from each respondent at least a randomized response Z that is correlated to Y. This randomization of the responses must be carried out independently for each population unit in the sample and independently of the sampling plan $p(s)$.

In such a case, the survey has two phases. First, a sample s is selected from U and then, given s, responses Z_i are realized using the selected RRT. We denote the corresponding probability distributions by $p(s)$ and $q\big(r\,|s\big)$. In this setting, the notions of expected value, unbiasedness, and variance are tied to a two-fold averaging process.

- Over all possible samples s that can be drawn using the selected sampling plan $p(s)$.
- Over all possible response sets r that can be realized given s under the response distribution $q\big(r\,|s\big)$.

In the sequel, we follow the literature and, where appropriate, denote the expectation operators with respect to these two distributions by E_p and E_q, respectively.

In a direct survey, the population total t_Y is usually estimated from the observed values Y_i using a linear estimator $t_s = \sum_{i \in s} b_{si} Y_i$, where the weights b_{si} follow the unbiasedness constraint $\sum_{s \ni i} p(s) b_{si} = 1,\ i = 1, \ldots, N$. If $\pi_i > 0\ \forall i \in U$, then Horvitz-Thompson's estimator

$$t_s^{HT} = \sum_{i \in s} \frac{Y_i}{\pi_i} \tag{1}$$

is a linear unbiased estimator with weights $b_{si} = 1/\pi_i$, and $\mathrm{E}_p\left(t_s^{HT}\right) = t_Y$, see Horvitz and Thompson (1952) or Section 2.8 in Tillé (2006) for details.

If the survey is conducted using RRT, the true values of Y_i for the sample s are unknown and, instead of them, the values of random variables Z_i correlated to Y_i are collected. Variables Z_i are usually further transformed into another variables R_i, which are more suitable for the construction of the desired estimator, and then the population total is typically estimated using a Horvitz-Thompson's type estimator

$$t_s^{HT,R} = \sum_{i \in s} \frac{R_i}{\pi_i}. \tag{2}$$

Suppose now that we have an estimator (a formula, or a computational procedure) for estimating the population total t_Y or population mean $\overline{t}_Y$; we denote it by t_Y^R and $\overline{t}_Y^R$, respectively. The subscript R emphasizes that the estimator is based on the values of R_i, i.e., on randomized responses. Furthermore, we assume that the randomized responses R_i follow a model for which it holds $\mathrm{E}\left(R_i\right) = Y_i$, $\mathrm{Var}\left(R_i\right) = \phi_i\ \forall i \in U$, and $\mathrm{Cov}\left(R_i, R_j\right) = 0\ \forall i \neq j,\ i, j \in U$. Note that the variance function ϕ_i of a randomized response R_i is a function of Y_i.

Recall that the estimator t_Y^R of the population total t_Y *is conditionally unbiased*, if the conditional expectation of t_Y^R given the sample s is equal to the current estimator t_s that would be obtained if no randomization took place, that is, if $\mathrm{E}_q\left(t_Y^R \mid s\right) = t_s$. The subscript s indicates that the "usual" estimator based on the nonrandomized sample, for example the Horvitz-Thompson's one, is used, and $\mathrm{E}_q\left(t_Y^R \,\middle|\, s\right)$ stands for the conditional expectation of t_Y^R given the sample s with respect to the distribution induced by the randomization of responses. For the estimator $\overline{t}_Y^R$ of the population mean, we proceed analogously.

If t_Y^R is conditionally unbiased and t_s is unbiased, then t_Y^R is also unbiased, since $\mathrm{E}\left(t_Y^R\right) = \mathrm{E}_p\big(\mathrm{E}_q(t_Y^R \mid s)\big) = \mathrm{E}_p\big(t_s\big) = t_Y$. Analogously, it holds $\mathrm{E}\left(\overline{t}_Y^R\right) = \overline{t}_Y$. Moreover, by a standard formula of the probability theory, we get the variance of t_Y^R in the form

$$\begin{aligned} \mathrm{Var}\left(t_Y^R\right) &= \mathrm{E}_p\Big(\mathrm{Var}_q\big(t_Y^R \,\big|\, s\big)\Big) + \mathrm{Var}_p\Big(\mathrm{E}_q\big(t_Y^R \,\big|\, s\big)\Big) \\ &= \mathrm{E}_p\Big(\mathrm{Var}_q\big(t_Y^R \,\big|\, s\big)\Big) + \mathrm{Var}_p\big(t_s\big). \end{aligned} \tag{3}$$

The second term on the right-hand side of (3) is, obviously, the variance of the estimator that would apply if no randomization of responses was deemed necessary, while the first term represents the increase of the variance produced by the randomization. In other words, the two terms on the right-hand side of (3) represent, respectively, *contribution by randomized response technique used* and *contribution by sampling variation* to the total variance of t_Y^R. When treating $\overline{t}_Y^R$, we proceed analogously.

Because the design-based expression for the variance of t_s is known for the most common sampling procedures, we can focus on the contribution of randomization and study it in more detail. For example, for the estimator $t_s^{HT,R}$ given by (2), we have

$$\begin{aligned}\mathrm{Var}\left(t_s^{HT,R}\right) &= \mathrm{E}_p\Big(\mathrm{Var}_q\big(t_s^{HT,R}\,\big|\,s\big)\Big) + \mathrm{Var}_p\Big(\mathrm{E}_q\big(t_s^{HT,R}\,\big|\,s\big)\Big)\\ &= \mathrm{E}_p\Big(\sum_{i\in U}\frac{\phi_i\xi_i}{\pi_i^2}\Big) + \mathrm{Var}_p\big(t_s^{HT}\big) = \sum_{i\in U}\frac{\phi_i}{\pi_i} + \mathrm{Var}\left(t_s^{HT}\right). \end{aligned} \tag{4}$$

Several techniques for estimating the population total were suggested in the literature. The papers Eriksson (1973) and Chaudhuri (1987) were at the origin, and became a benchmark for many following approaches. Both techniques have been further developed and improved by other researchers; see, e.g., interesting papers Arnab (1995, 1998) or Gjestvanga and Singh (2009). The ideas and a representative review of further research are presented in a monograph Chaudhuri (2017). Another type of randomization technique was suggested in a series of papers by Dalenius and his colleagues, e.g., Bourke and Dalenius (1976) or Dalenius and Vitale (1979). Among recent papers on the topic of sensitive questions in population surveys, we would like to mention, for example, papers by Kirchner (2015) and Trappmann (2014). In both of them, long lists of relevant references can be found. Finally, recall that probably the most comprehensive account of developments in sample survey theory and practice can be found in Pfeffermann and Rao (2009a,b), or in the more recent monographs Arnab (2017), Tian and Tang (2014), Tillé (2020) or Wu and Thompson (2020).

2 NEW RANDOMIZED RESPONSE TECHNIQUE

In this section, we suggest a completely different approach. Assume that the studied sensitive variable Y is non-negative and bounded from above, i.e., $0 \le Y \le M$. First, let us assume the upper bound M of the variable Y is known. Each respondent performs, independently of the others, a random experiment generating a pseudorandom number Υ from the uniform distribution on interval $(0, M)$, while the interviewer does not know this value. The respondent can generate the pseudorandom number Υ using, for example, a laptop online/offline application; for some other possibilities, see Section 2.4. The respondent then answers a simple question: "*Is the value of Y greater than Υ?*" (e.g.: "*Is your monthly income greater than Υ?*").

For certain sensitive variables, such as the total amount of alcohol consumed within a certain period, it is better to use a question: "*Is the value of Y lower than Υ?*" In such a case we recode the response $Z_{i,(0,M)}$ to $Z^{\star}_{i,(0,M)} = 1 - Z_{i,(0,M)}$, and apply the suggested RRT to $Z^{\star}_{i,(0,M)}$.

The response of the i^{th} respondent follows the alternative distribution with the parameter Y_i/M, that is

$$Z_{i,(0,M)} = \begin{cases} 1 \text{ with probability } \frac{Y_i}{M}, & \text{if } \Upsilon_i < Y_i,\\ 0 \text{ with probability } 1 - \frac{Y_i}{M}, & \text{otherwise.}\end{cases} \tag{5}$$

Therefore, $\mathrm{E}\left(Z_{i,(0,M)}\right) = P\big(\Upsilon_i < Y_i\big) = Y_i/M$ and $\mathrm{Var}\left(Z_{i,(0,M)}\right) = \big(Y_i/M\big)\big(1 - Y_i/M\big)$. Therefore, we transform $Z_{i,(0,M)}$ to $R_{i,(0,M)} = MZ_{i,(0,M)}$, for which we have

$$\mathrm{E}\left(R_{i,(0,M)}\right) = Y_i \quad \text{and} \quad \mathrm{Var}\left(R_{i,(0,M)}\right) = Y_i\big(M - Y_i\big). \tag{6}$$

2.1 Application to the simple random sampling

Consider now the situation in which the sampling plan $p(s)$ is a simple random sampling without replacement with a fixed sample size n. Denote, as in Section 1, by $\overline{t}_Y = \frac{1}{N}\sum_{i\in U} Y_i$ the population mean, by $S_Y^2 = \frac{1}{N-1}\sum_{i\in U}\left(Y_i - \overline{t}_Y\right)^2$ the population variance. In this case, the inclusion probabilities are constant, that is, $\pi_i = P(\xi_i = 1) = n/N \ \forall i \in U$.

Let the population total t_Y be estimated using the Horvitz-Thompson's type estimator

$$t_{(0,M)}^{HT,R} = \frac{N}{n}\sum_{i\in s} R_{i,(0,M)}. \tag{7}$$

It follows from (6) that this estimator is unbiased, so let us calculate its variance. For this purpose (4) can be used effectively. First, note that in the case considered $\pi_i = n/N$, and due to (6) $\phi_i = Y_i(M - Y_i)$. Second, taking into account variance of the simple random sampling without replacement, see Section 4.4 in Tillé (2006) for details, we obtain after a straightforward calculation

$$\mathrm{Var}\left(t_{(0,M)}^{HT,R}\right) = \frac{N^2}{n}\Big(\overline{t}_Y(M - \overline{t}_Y) - \frac{N-1}{N}S_Y^2\Big). \tag{8}$$

To characterize the variance of the suggested estimators more deeply and to get a more transparent understanding of the variance of the suggested RRT, we introduce two auxiliary characteristics termed *measures of concentration*. More precisely, let us denote

$$\Gamma_{Y,M} = \frac{1}{N}\sum_{i\in U}\frac{Y_i}{M}\Big(1 - \frac{Y_i}{M}\Big) = \underbrace{\frac{1}{MN}\sum_{i\in U} Y_i}_{\frac{1}{M}\overline{t}_Y} - \underbrace{\frac{1}{M^2N}\sum_{i\in U} Y_i^2}_{\frac{1}{M^2}\overline{Y^2}} = \frac{\overline{t}_Y}{M} - \frac{\overline{Y^2}}{M^2} \tag{9}$$

and

$$\Gamma_{\overline{Y},M} = \frac{\overline{t}_Y}{M}\frac{(M - \overline{t}_Y)}{M} = \frac{\overline{t}_Y}{M} - \frac{\overline{t}_Y^2}{M^2}. \tag{10}$$

In the sequel, we call $\Gamma_{Y,M}$ *the mean relative concentration measure*, and $\Gamma_{\overline{Y},M}$ *the proximity measure of the population mean* $\overline{t}_Y$ *to* $M/2$.

If Y_i are iid random variables with finite variance σ^2 and an expectation μ, then, by the law of large numbers, both $\Gamma_{Y,M}$ and $\Gamma_{\overline{Y},M}$ converge, as $N \to \infty$, with probability 1 to

$$\Gamma_{Y,M,as} = \frac{\mu}{M}\Big(1 - \frac{\mu}{M}\Big) - \frac{\sigma^2}{M^2} \quad \text{and} \quad \Gamma_{\overline{Y},M,as} = \frac{\mu}{M}\Big(1 - \frac{\mu}{M}\Big). \tag{11}$$

We call $\Gamma_{Y,M,as}$ *the asymptotic mean relative concentration measure*, and $\Gamma_{\overline{Y},M,as}$ *the asymptotic proximity measure of the population mean* $\overline{t}_Y$ *to* $\frac{M}{2}$. Note that both $\Gamma_{Y,M,as}$ and $\Gamma_{\overline{Y},M,as}$ exist if $0 \le Y_i \le M \ \forall i \in U$.

Let us focus on $\Gamma_{Y,M}$ and $\Gamma_{\overline{Y},M}$ in more detail. First, note that in our setting both are population characteristics, not random variables. Second, both take their values in the interval $[0, 1/4]$, and are equal to zero only in pathological cases when either $Y_i = 0 \ \forall i \in U$ or $Y_i = M \ \forall i \in U$. The higher these measures, the higher the variance of $t_{(0,M)}^{HT,R}$. The mean relative concentration measure $\Gamma_{Y,M}$ reaches its maximum $1/4$ when all values are at the center of the interval $(0, M)$, that is, if $Y_i = M/2 \ \forall i \in U$. The proximity measure $\Gamma_{\overline{Y},M}$ of the population mean to the center of the interval $(0, M)$ reaches its maximum $1/4$ only if the population mean is at the center of the interval, that is, $\overline{t}_Y = M/2$. This case occurs, e.g.,

when random variable Y is symmetric around the center of the interval $M/2$; this feature is certainly true for the uniform distribution on $(0, M)$.

For a fixed value of the upper bound M, population size N and sample size n, the contribution of the suggested RRT to the variance of $t_{(0,M)}^{HT,R}$ depends, up to a multiplicative constant, on $\Gamma_{Y,M}$, because it holds

$$\mathrm{E}_p\Big(\mathrm{Var}_q\big(t_{(0,M)}^{HT,R} \mid s\big)\Big) = \frac{M^2N^2}{n} \underbrace{\frac{1}{N}\sum_{i\in U}\frac{Y_i}{M}\Big(\frac{M-Y_i}{M}\Big)}_{\Gamma_{Y,M}} = \frac{M^2N^2}{n}\Gamma_{Y,M}. \tag{12}$$

Analogously, this contribution can also be expressed, up to multiplicative constants, by $\Gamma_{\overline{Y},M}$ and S_Y^2, because it holds

$$\mathrm{E}_p\Big(\mathrm{Var}\,\big(t_{(0,M)}^{HT,R} \mid s\big)\Big) = \frac{M^2N^2}{n}\Gamma_{\overline{Y},M} - \frac{N(N-1)}{n}S_Y^2. \tag{13}$$

Thus, both $\Gamma_{Y,M}$ and $\Gamma_{\overline{Y},M}$ can help us explain how the suggested RRT increases the variance of the estimator of the population total t_Y for distributions symmetrical around $M/2$, for distributions concentrated close to the center of $(0, M)$, symmetrical around $M/2$, or uniformly distributed. Moreover, they show that the suggested approach is especially suitable for skewed distributions, provided that they are concentrated around their mean values. Let us sum up: both measures of concentration help us not only to describe the variance of the estimator used, compare (12) and (13), but also to interpret it better.

Remark 1. If the values of Y are bounded both from below and above, that is, $0 < m \leq Y \leq M$, then variance of $t_{(0,M)}^{HT,R}$ can be significantly reduced by generating pseudorandom numbers Υ_i from the uniform distribution on the interval (m, M) instead on $(0, M)$. In fact, if this is the case, we replace $Z_{i,(0,M)}$, described by (5), with

$$Z_{i,(m,M)} = \begin{cases} 1 \text{ with probability } \frac{Y_i-m}{M-m}, & m \leq \Upsilon_i < Y_i, \\ 0 \text{ with probability } 1-\frac{Y_i-m}{M-m}, & \text{otherwise,} \end{cases}$$

transform these variables to $R_{i,(m,M)} = m + (M-m)Z_{i,(m,M)}$, and estimate population total t_Y analogously to (7) using the Horvitz-Thompson's type estimator

$$t_{(m,M)}^{HT,R} = \frac{N}{n}\sum_{i\in s} R_{i,(m,M)}. \tag{14}$$

It is easy to show that the variance of $t_{(m,M)}^{HT,R}$ is smaller than that of $t_{(0,M)}^{HT,R}$, that is, by the value $\frac{N^2m}{n}\big(M-\overline{t}_Y\big)$.

The immediate question arises of what happens if the interval $[m, M]$ is not set correctly. Evidently, if some values of Y_i are outside the interval $[m, M]$, then with probability 1 it holds $Z_{i,(m,M)} = 0$ if $Y_i < m$ and $Z_{i,(m,M)} = 1$ if $Y_i > M$. The bias of the suggested estimator is equal to

$$\sum_{i\in U \mid Y_i<m}\big(Y_i-m\big) + \sum_{i\in U \mid Y_i>M}\big(Y_i-M\big). \tag{15}$$

In practice, the bounds of the variable Y are often unknown. When choosing parameters m and M, a researcher should carefully consider the trade-off between bias and privacy.

While lower bound m affects mostly bias and is not very crucial to the privacy of respondents, the choice of M affects both bias and privacy. Moreover, there is also a trade-off between bias and variance of estimates; see the results of the simulations in Tables 2 – 4 in the Annex. Therefore, a reasonable guess about the empirical quantiles of the characteristics studied is vital for setting the values of m and M properly.

Let us discuss some advantages and disadvantages of our approach compared to the other techniques suggested in the literature.

- It is simple; this fact increases respondents' confidence and cooperation, and thus reduces the estimation error.
- Respondents' privacy is well protected, because they never report the true value of the sensitive variable.
- It avoids the demanding task of designing a randomization device intended for masking the surveyed variable.
- It enables to estimate the population total at an acceptable level of accuracy, see Section 3 for details. Of course, what level is acceptable depends on the survey and selected precision requirements. According to our simulations, standard errors of the estimators described up to now are at most two times higher than those of HT-estimators, see Tables 2 – 4.
- On the other hand, due to the need of a device/technique for generation random numbers, some respondents may feel a lower degree of confidence in preserving their anonymity.

Finally, we find rather problematic any comparison of our approach with other methods because their performance strongly depends on the choice of the randomizing device used. In our opinion, it is tricky to design, e.g., a deck of cards for a continuous variable with a high range, such as the income in the Czech Republic, and a reliable estimator of this type with an acceptably small variance value would need an excessively large size.

2.2 Estimators using knowledge of Υ

A natural question arises as to whether we could improve the accuracy of the suggested method. Thus, in what follows, we discuss the two modifications of the RRTs suggested in Section 2.1 and their properties in the following subsections. The heuristics behind this approach are based on the following observations. All techniques presented up to now have assumed that the interviewer does not know the outcome of the randomization device leading to the randomized response, such as the card drawn, the value of the pseudorandom number, etc. It is plausible to ask what would happen if we also knew the outcome of that random experiment on the one hand, while protecting respondents' privacy on the other one. More precisely: *Can we modify the estimator and to increase its accuracy, that is, to decrease its variance, if we also know the values of the generated pseudorandom number*? We surmise that it is feasible and suggest one possible way of reaching this goal. However, we point out that the success of the suggested approach, to a considerable extent, depends on the statistician's insight into the problem.

Assume again that the studied sensitive variable Y is non-negative and bounded from above, that is, $0 \leq Y \leq M$. Each respondent carries out, independently of the others, a random experiment generating a pseudorandom number Υ from the uniform distribution on interval $(0, M)$, and *informs the interviewer of both its value* and *whether* $\Upsilon < Y$ *or not*. For example, the response is that the simulated number has been xxx (let say 45 000 CZK) and the respondent earns more/less. To distinguish from the situation described in Section 2.1, we further assume that the corresponding random response is now described

using a dichotomous random variable

$$Z_{i,\alpha,(0,M)} = \begin{cases} 1-\alpha+2\alpha\frac{\Upsilon_i}{M}, & \text{if } \Upsilon_i < Y_i, \\ -\alpha+2\alpha\frac{\Upsilon_i}{M}, & \text{otherwise,} \end{cases} \qquad 0 \le \alpha < 1,\ i = 1,\dots,n, \tag{16}$$

where α is a tuning parameter. Its value is a priori set by the interviewer, is fixed and unknown to the respondent. This proposal is a linear combination of our initial proposal $Z_{i,(0,M)}$ given by (5) and $2\Upsilon_i/M$. Higher the value α, more weight is put on the term using the pseudorandom number Υ. For $\alpha = 0$ we have the initial method described in Section 2.1. The rule for an optimal choice of α is given later in this section.

The response of the respondent to $Z_{i,\alpha,(0,M)}$ is transformed not by the respondent, but by the interviewer off-line. The discussion about the choice of α is postponed here and will be done later.

Since $P\big(Z_{i,\alpha,(0,M)} = 1-\alpha+2\alpha\frac{\Upsilon_i}{M}\big) = P\big(\Upsilon_i < Y_i\big)$, we have

$$\mathrm{E}\left(Z_{i,\alpha,(0,M)}\right) = \frac{1}{M}\int_0^{Y_i}\Big(1-\alpha+2\alpha\frac{u}{M}\Big)\,du + \frac{1}{M}\int_{Y_i}^{M}\Big(-\alpha+2\alpha\frac{u}{M}\Big)\,du = \frac{Y_i}{M},$$

$$\mathrm{Var}\left(Z_{i,\alpha,(0,M)}\right) = \frac{1-2\alpha}{M^2}Y_i\big(M-Y_i\big) + \frac{\alpha^2}{3}.$$

Therefore, the random responses $Z_{i,\alpha,(0,M)}$ are further transformed to $R_{i,\alpha,(0,M)} = MZ_{i,\alpha,(0,M)}$, and the desired estimator of the population total t_Y is constructed analogously to (7) and (14). More precisely, we suggest using again the Horvitz-Thompson's type of estimator in the form

$$t^{HT,R}_{\alpha,(0,M)} = \frac{N}{n}\sum_{i\in s} R_{i,\alpha,(0,M)}. \tag{17}$$

It is evident that $\mathrm{E}\left(R_{i,\alpha,(0,M)}\right) = Y_i$, so the estimator (17) is unbiased. Moreover, the contribution of randomization to its variance is

$$\mathrm{E}_p\Big(\mathrm{Var}_q\big(t^{HT,R}_{\alpha,(0,M)}\,\big|\,s\big)\Big) = \frac{M^2N^2}{n}\sum_{i\in U}\Big[\,\frac{1}{N}\big(1-2\alpha\big)\frac{Y_i}{M}\Big(1-\frac{Y_i}{M}\Big) + \frac{\alpha^2}{3N}\,\Big]. \tag{18}$$

An easy calculation shows that (18) has a global minimum at $\alpha = 3\Gamma_{Y,M} \in [0, 3/4]$. If we set $\alpha_{opt} = 3\Gamma_{Y,M}$ and substitute it back to (18), then the contribution of randomization to the variance of (17) for this choice of α is

$$\begin{aligned}\mathrm{E}_p\Big(\mathrm{Var}_q\big(t^{HT,R}_{\alpha_{opt},(0,M)}\,\big|\,s\big)\Big) &= \frac{M^2N^2}{n}\sum_{i\in U}\Big[\,\big(1-6\Gamma_{Y,M}\big)\frac{1}{N}\frac{Y_i}{M}\Big(1-\frac{Y_i}{M}\Big) + \frac{3\Gamma^2_{Y,M}}{N}\,\Big] \\ &= \frac{M^2N^2}{n}\Gamma_{Y,M}\big(1-3\Gamma_{Y,M}\big).\end{aligned} \tag{19}$$

If we compare (19) with (12), we see that the knowledge of pseudorandom numbers Υ_i and the use of α_{opt} considerably decrease the variability, of course, depending on the suggested RRT. It is worth highlighting that our simulations summarized in Section 3 confirm these findings.

The conclusion that the knowledge of Υ leads to a smaller variance of the estimator is expected; see above. The reason is clear and is based on the well-known inverse relationship

that exists between the disclosure of personal information and the efficiency of estimates, that is, *the more the privacy is jeopardized the lower the variance*. For a discussion, see Chaudhuri and Mukerjee (1988), among others. In our case, the assumption that the interviewer knows Υ means that the privacy of the respondent is less protected and, consequently, we get better estimates.

Parameter α should be set to its optimal value $\alpha_{opt} = 3\Gamma_{Y,M}$, where the mean relative concentration measure $\Gamma_{Y,M}$ is introduced in Section 2, Formula (9). If the interviewer has some prior information about the mean μ and variance σ^2 values for the theoretical distribution of the surveyed variable Y, he/she should rather apply the asymptotic concentration measure (11), which can be estimated using a plug-in moment estimator. More precisely, the population mean $\overline{t}_Y$ should be replaced by μ, and the population variance S_Y^2 by σ^2. Since the population second moment $\overline{Y^2}$ can be expressed as $\frac{N-1}{N}S_Y^2 + \overline{t}_Y^2$, it is sufficient to substitute μ and σ^2 into this expression. Moreover, recall that the prior information is often available for regular surveys in official statistics, such as EU-SILC, because in such a case we can either use results from previous years updated by inflation, or we can rely on the expert opinion. If no prior information is available, we recommend choosing small values of α, such as 0.5.

Notice that if a nonnegative surveyed random variable Y is bounded not only from above but also from below, that is, $0 < m \leq Y \leq M$, we generate Υ_i from the uniform distribution on the interval (m, M), modify $Z_{i,\alpha,(0,M)}$ given by (16) to

$$Z_{i,\alpha,(m,M)} = \begin{cases} 1 - \alpha + 2\alpha\frac{\Upsilon_i - m}{M-m}, & \text{if } \Upsilon_i < Y_i, \\ -\alpha + 2\alpha\frac{\Upsilon_i - m}{M-m}, & \text{otherwise,} \end{cases} \qquad 0 \leq \alpha < 1,$$

transform $Z_{i,\alpha,(m,M)}$ to $R_{i,\alpha,(m,M)} = m + (M-m)Z_{i,\alpha,(m,M)}$, and form an estimator of the population total t_Y of the Horvitz-Thompson's type, parallel to (17), as

$$t^{HT,R}_{\alpha,(m,M)} = \frac{N}{n}\sum_{i\in s} R_{i,\alpha,(m,M)}. \tag{20}$$

Using analogous arguments as above, it is straightforward to show that $\mathrm{E}\left(R_{i,\alpha,(m,M)}\right) = Y_i$, so that the estimate (20) is again unbiased regardless of the value of the parameter α.

We must firmly emphasize that neither the information about the value of pseudorandom number Υ nor the value α enables us to guess the exact value of the sensitive variable Y, except for the case $Y = M$. In other words, knowing them does not intrude on the respondent's privacy.

The heuristics behind the proposed modification are the following:

- If the response is *YES*, then a high value of the pseudorandom number Υ implies a high value of the studied variable Y, because $Y > \Upsilon$, and these observations "considerably" increase the value of the estimator.
- However, if the response is *NO*, then a low value of the pseudorandom number Υ implies a low value of Y, because $Y \leq \Upsilon$, and these observations "considerably" decrease the value of the estimator.

Unfortunately, in both situations, that is, when the value of the response is either (too) low or (too) high, the respondent may be more prone to fabricate his/her response.

As we can see, $Z_{i,\alpha,(m,M)}$ can occasionally attain negative values, which is an obvious drawback. On the other hand, using a guess about the distribution of Y, it is possible to estimate (at least roughly) the probability of such an event. For illustration, in the case of

practical application of our approach described in Section 3, the probability of obtaining a negative $Z_{i,\alpha,(m,M)}$ is of the order 10^{-5}, and during our extensive simulations, we never met such a case. In this paper, we do not study the effect on bias and variance when setting negative values to zero.

2.3 Estimator using switching questions

We emphasize that for some characteristics, such as the monthly income of a household or alcohol consumption, it can be sensitive for respondents to report either high or low values. This led us to modify the suggested RRT approach in the following way.

Assume that a nonnegative surveyed random variable Y is bounded both from below and above, that is, $0 < m \leq Y \leq M$. First, we set a proper fixed threshold T, $m < T < M$, unknown to the respondent. Second, we generate Υ from the uniform distribution on (m, M) and, depending on whether the pseudorandom number Υ does or does not exceed this fixed threshold T, we ask one of the following questions:

i. If $\Upsilon \leq T$: "*Is the value of* Y *greater than* Υ*?*",
ii. If $\Upsilon > T$: "*Is the value of* Y *smaller or equal than* Υ*?*".

Third, for the ith respondent, we form a random variable

$$Z_{i,T,(m,M)} = \begin{cases} 1, & \text{if } \Upsilon_i \leq T \ \& \ \Upsilon_i \leq Y_i, \\ 0, & \text{if } \Upsilon_i \leq T \ \& \ \Upsilon_i > Y_i \quad \text{or} \quad \Upsilon_i > T \ \& \ \Upsilon_i \leq Y_i, \\ -1, & \text{if } \Upsilon_i > T \ \& \ \Upsilon_i > Y_i. \end{cases}$$

If we know both the response concerning the value of Y and the question asked, that is whether $\Upsilon_i \leq T$ or not, then it is easy to show that $\mathrm{E}\left(Z_{i,T,(m,M)}\right) = (T+Y_i-m-M)/(M-m)$. This advises to transform $Z_{i,T,(m,M)}$ to $R_{i,T,(m,M)} = (M-m)Z_{i,T,(m,M)}+m+M-T$, because then $\mathrm{E}\left(R_{i,T,(m,M)}\right) = Y_i$. Thus, the Horvitz-Thompson's type estimator of the population total t_Y of the form

$$t^{HT,R}_{T,(m,M)} = \frac{N}{n} \sum_{i \in s} R_{i,T,(m,M)}. \tag{21}$$

is evidently also unbiased.

As concern variance of $R_{i,T,(m,M)}$, we must distinguish between $Y_i > T$ and the complementary inequality. After a bit of tedious calculation we get, as expected, that it is always higher than the variance of $R_{i,(m,M)}$. Worse still is the fact that negative values of $Z_{i,T}$ may occur quite frequently, leading to negative values of the corresponding $R_{i,T,(m,M)}$. Looking at the results of our simulations, we observe that $t^{HT,R}_{T,(m,M)}$ can return inadmissibly low or even negative values, which is a major drawback. Moreover, we cannot find the way how to set optimal value of the threshold T minimizing $\mathrm{Var}\left(R_{i,T,(m,M)}\right)$, being another drawback.

An unbiased estimator of the population mean $\overline{t}_Y$ can be constructed in parallel. On the other hand, if we know only the response concerning the value of Y but not the question asked, in this case it is not possible to construct an estimator of the population total t_Y, respectively of the population mean $\overline{t}_Y$.

Thus, the seemingly appealing idea described in this section seems to be interesting from a theoretical point of view. We cannot recommend it for practical use automatically without prior information on the population studied, which is also illustrated by the simulations presented in Section 3.

2.4 Random number generation

In all RRTs, the choice of randomization device is probably the trickiest point. If we assume direct face-to-face interviewing, the following points describe several possibilities that might be used in our approach.

- We allow the respondent to select the random number according to some standard, e.g. the European ISO 28640:2010(en) Standard ISO. We are convinced that the existence of a standard can increase the credibility of the survey and willingness of respondents to respond truthfully. The selected random number is then used according to the RRT used.
- To those respondents who feel like "experts in the field of randomness", the reviewer can offer them the option to select a random number from the uniform distribution using their own method.
- Another possibility is, for example, to use a large deck of cards, but it would require additional calculations to find the bias of such an approach.

2.5 Open problems

There are several relevant related research issues, not treated in this paper due to its current length. We aim to concentrate on them in subsequent papers. They include, but are not limited to, the following points:

- To generalize suggested estimators to more complex sampling plans as cluster sampling, two-stage sampling, stratified sampling, etc. Moreover, it has been repeatedly emphasized during the discussions, e.g. after the presentation of our results, that the median and other quantiles are important statistics for many applications, sometimes even more important than the mean. Similarly, the question has been raised whether parallel methodology could be used in any type of regression analysis combining it, e.g., with ideas from Antoch and Janssen (1989), Pfeffermann and Rao (2009a,b) or Tillé (2020).
- To modify, where appropriate, suggested estimators to the case when pseudorandom numbers are generated not from the uniform distribution, but from the distribution that mimics the surveyed variable Y. To prepare a numerical study illustrating the effect of the distribution from which we simulate random numbers on the on the possible improvements in the performance of estimators. In the case of income covered in our simulation example, the log-logistic or log-normal distribution might be used.
- To study more profoundly effects of tuning parameters on the bias, variance, and privacy jeopardy, as well as the trade-off among the parameters and pseudorandom numbers generated from different distributions. To suggest rules of thumb for the choice of parameters m, M and α and to study optimal choice of parameters with respect to the minimization of the mean square error.
- To derive unbiased estimators of variance and to study the impact on the corresponding confidence intervals. To study the effect on bias and variance of the suggested procedures when treating possible negative values as zeros.
- To compare our proposal with that of the unrelated question model suggested originally in Greenberg (1971).
- To find approximate formulae for sample sizes with required margin of error.

3 SIMULATION STUDY

In many countries, income is recognized as private and (highly) sensitive information. Respondents often refuse to respond at all or provide strongly biased responses. This in particular happens if their income is (very) high or (very) low. This leads us to assess the

performance of the proposed RRTs through a simulation study using Czech wage data from the Average Earnings Information System (IPSV) of the Ministry of Labor and Social Affairs of the Czech Republic.

Based on the extensive analysis of monthly wage statistics provided by IPSV for the years 2004 – 2014, Vrabec and Marek (2016) recommended to model wages in the Czech Republic using a three-parameter log-logistic distribution with the density

$$f(y;\tau,\sigma,\delta)=\begin{cases}\frac{\tau}{\sigma}\left(\frac{y-\delta}{\sigma}\right)^{\tau-1}\left(1+\left(\frac{y-\delta}{\sigma}\right)^{\tau}\right)^{-2}, & y\geq\delta>0,\ \tau>0,\ \sigma>0,\\ 0, & \text{otherwise,}\end{cases} \tag{22}$$

where $\tau>0$ is a shape parameter, $\sigma>0$ is a scale parameter and δ is a location parameter.

Vrabec and Marek (2016) also calculated the estimates of the parameters of (22) for the data of 2^{nd} quarter 2014 and obtained

$$\widehat{\tau}=4.0379,\ \widehat{\sigma}=21,687\ \text{ and }\ \widehat{\delta}=250. \tag{23}$$

The estimates (23) are based on aggregated data (frequencies by wage intervals with constant width 100 CZK) of roughly 2.1×10^6 observations, covering practically half of the overall relevant population. The corresponding estimated average monthly income is 24, 290 CZK (approximately 950 EUR).

The probability histogram of the data with bin width 500 (CZK), and density of the log-logistic distribution (22) with the unknown parameters replaced by their estimates (23), are presented in Figure 1. In addition to that, the corresponding sample distribution function is presented in Figure 2. Both the histogram and the sample distribution function were constructed from the same aggregated data from the 2^{nd} quarter 2014 used for estimation of parameters of the model. Point out that all calculation and simulations were conducted by the statistical freeware R, version 3.5.1, see R Core Team (2021).

It is interesting to look at both the lower and upper sample quantiles of the data used. While 7 000 CZK corresponds to the 0.0003 sample quantile, 8 000 CZK corresponds to the 0.01 sample quantile, whis is the reason why we set $m=7\,000$. Analogously, 40 000 CZK corresponds to the 0.91 sample quantile, 60 000 CZK to the 0.97 sample quantile and, finally, 80 000 CZK to the 0.98 sample quantile, see Figure 2.

It is obvious from Figure 1 that the original data are highly skewed. Therefore, it is not surprising that the mean relative concentration measure $\Gamma_{Y,M}=0.198$ is close to its attainable maximum, so that the estimator $t^{HT,R}_{\alpha,(m,M)}$ based on the knowledge of Υ_i's and "almost-optimal" choice of the parameter $\alpha\approx3\Gamma_{Y,M}$ should have smaller variance than $t^{HT,R}_{(m,M)}$ (corresponding to $\alpha=0$). Moreover, it follows from (4) that the variance of the estimators using the suggested RRTs will be higher than for the Horvitz-Thompson's estimator based on the nonrandomized data. All this is confirmed by our simulations, compare the results of Tables 2 – 4.

Neither the real population nor the real sample is available to us, because files with microdata from ISPV survey are not available to researchers. Therefore, the populations U are generated using the model wage distribution (log-logistic). More precisely, 1000 replications of populations sized $N=200$, or $N=400$, are simulated from model (22), in which the unknown parameters have been replaced with their estimates (23), using the package *flexsurv*, see Jackson (2016). Let us point out that the population sizes $=200$ and $N=400$ are commonly used sizes of a stratum in business statistics or surveyed community (village, group of students, etc.). It is worth to emphasize that the simulation results virtually do not change after 100 replications of the population; the differences begin at the third significant digit.

Moreover, 1000 replications from the log-logistic distribution are generated using the package *flexsurv*, see Jackson (2016). Point out that the simulation results virtually do not change after 100 replications of the population; the differences begin at the third significant digit. All simulations and calculations are conducted by statistical freeware R, version 3.5.1, see R Core Team (2021).

From each replication of the population, we draw, without replacement, 1000 random samples of the size $n = 20$, or $n = 50$. Such sample sizes are standard for separate strata in business sampling surveys, and also in the social statistical surveys, such as the EU Statistics of Income Living Condition. Let us take a closer look at average sample size per stratum in more detail. In such a survey, for a medium sized country like the Czech Republic with a population of 10^7 inhabitants and approximately $4.3 \cdot 10^6$ households, the samples approximately include 9 500 households surveyed in a two-dimensional stratification (region and size of municipality), giving $78 \times 4 = 312$ strata. The average sample size is then about 30 per stratum. In EU-SILC, detailed results are presented for eight income groups, leading on average to the population size of approximately $N = 1.25 \cdot 10^6$ inhabitants per one income group. The setting of the simulation was based on the real sample and population sizes of the EU-SILC of a medium size EU country. For a more detailed description of the stratification, strata, sample sizes, and sampling design, see GESIS (2016).

For each sample, both t_Y and $\overline{t}_Y$ are estimated using the techniques described in Section 2. Estimates of the total mean values, instead of population totals, are presented to enable a more easy comparison between the results obtained for populations with different sizes N and different sample sizes n.

In simulations, we are especially interested in the impact of "tuning parameters" m, M, T, α and α_{opt} on estimates. Taking into account the type and nature of the data that we simulate, we set the parameters as described in Table 1. The values of α_{opt} were set using the formulae for the optimal variance described in Section 2. Other parameters were chosen with regard to our experience, in particular, the monthly salary that can be perceived to be high. Since practically all available data are larger than 7 000 CZK, we set the lower bound of the interval for generating pseudorandom numbers Υ_i to $m = 7\,000$.

The results are summarized[4] in Tables 2 – 4 and in Figure 3 – 5. They show that for larger population size N and larger sample sizes n the accuracy improves substantially. The original proposal without knowledge of pseudorandom numbers seems to be also promising for real life applications. Even the method of switching questions might be applicable for large samples from large populations if prior information is available. However, more simulations using different shapes of population distributions are needed to support these hypotheses.

The reason for the lower standard deviation of $\overline{t}^{HT,R}_{\alpha,(m,M)}$, and especially $\overline{t}^{HT,R}_{\alpha_{opt},(m,M)}$, compared to $\overline{t}^{HT,R}_{(m,M)}$ and $\overline{t}^{HT,R}_{T,(m,M)}$ is that these estimators efficiently use the information on the generated numbers of Υ. Recall that we used the moment plug-in estimate for the optimal value of α.

As expected, the values of variance of the suggested estimators are higher than those of Horvitz-Thompson's estimator based on the non-randomized data. The precision of our basic proposal is practically acceptable because, according to simulations, the corresponding sample standard deviation of the estimates increased by a mere 60 % in comparison with the Horvitz-Thompson estimate for $M = 60\,000$. This result is quite reasonable, taking into account that Y is a very sensitive variable and high nonresponse (even 50 % and more

[4] In Tables 2 – 4 both the sample averages (means) and sample standard deviation (sd) of the estimates from the simulations are presented. For simplicity, we omit "HT" in the descriptions of the estimators analyzed in all figures and tables because all the estimators we compare here are of the Horvitz – Thompson's type.

in everyday practice) for direct questioning. However, note that the modification using knowledge of the values of Υ_i leads to a substantial reduction in variance. Thus, while mildly relaxing respondents' privacy on the one hand but still keeping secret the true response because the true value of the sensitive variable is never reported, this modification provides estimates whose precision is comparable with directly surveying under zero nonresponse. On the other hand, the high variability of the estimates, even the presence of negative estimates for the mean wages, shows that the modification of the switching questions described in Section 2.3 is only a theoretical exercise and cannot be recommended for practical use. Its improvement remains an open question.

Comparing in all tables the simulation results for optimal value α_{opt} of the parameter α and fixed values $\alpha = 0.75$, we see that the mean has practically not changed; however, the expected decrease occurs in the variability of the estimate. This decrease of approximately 9% of the standard deviation (sd) shows that it pays "to tune up" the procedure and its parameters according to the given problem and available data.

Both the results of Section 2.1 and the simulations show that the variance of the estimators can be greatly reduced by choice of bounds m and M. We see that for low values of the upper bound $M = 40\,000$ the proposed estimators are competitive even with the Horvitz-Thompson estimator. It follows from (15) that approximately unbiased estimators with low variance can be constructed if we use prior information on population quantiles for choice of bounds m and M. The optimal choice of bounds with respect to the minimization of the mean square error is a field of further research.

CONCLUSIONS

The paper introduces a new randomized response model and two variants of it, intended to gather information on a (positive) sensitive quantitative variable and to estimate the population total (population mean). The idea underlying the proposal is seemingly very easy and, unlike many scrambled response methods present in the literature, does not require demanding arithmetic operations from the respondents nor the use of complicated randomization devices.

It possesses three attractive properties, namely:

1. Although a quantitative estimate is the final end, the respondent is only asked for a qualitative response.
2. It is simple to use.
3. It provides a high level of anonymity to the respondent.

In the first model, respondents are first asked to generate a random number (a sort of random threshold) from a continuous uniform distribution. Then, without revealing the generated number to the interviewer, the survey participants are asked to declare whether the true value of the sensitive variable is greater than the generated number. Under this model, the privacy of the respondents is completely protected. The two variants of the model discuss the case where the generated number is also known to the interviewer, and therefore privacy is less protected. Consequently, the use of the two variants in real analyses is not recommended, since they are prone to produce misreporting and untruthful response. They have a value only from a theoretical point of view.

A disadvantage of the discussed method may, for some respondents, be a feeling of infringement on their privacy due to an extrinsic device/technique being used for generating random numbers. This problem is mainly psychological in nature and can, at least partially, be resolved by a proper explanation of the approach of the interviewer. Unfortunately, all currently used RRT procedures suffer, to some extent, from the same problem, see the thorough discussion in Chaudhuri (2017), Chaudhuri and Christofides (2013), among others.

For all suggested RRT procedures, we show their unbiasedness and derive the corresponding variance for the Horvitz-Thompson's type estimator under simple random sampling without replacement. The optimal values of the tuning parameters that enable us to minimize the variance of the suggested procedures are also discussed.

As a technical tool, two auxiliary measures are proposed. With the aid of them we can explain why and especially how the suggested RRTs increase the variance of the estimators of t_Y and $\overline{t}_Y$ for symmetrical distributions, distributions closely concentrated around their centers, or uniform distribution.

ACKNOWLEDGEMENTS

The work of the first author was partially supported by GA ČR under the Grant number P403/22/19353S. The work of the third author was prepared under Institutional Support to Long-Term Conceptual Development of Research Organization, the Faculty of Informatics and Statistics of the University of Economics, Prague. The authors are grateful to the associated editor and two unknown reviewers for their valuable comments that considerably improved the contents of this paper.

REFERENCES

ANTOCH, J., JANSSEN, P. (1989). Nonparametric regression M-quantiles. *Statistical and Probability Letters*, 8: 355–362. <https://doi.org/10.1016/0167-7152(89)90044-8>.

ARNAB, R. (1995). Optimal estimation of a finite population total under randomized response surveys. *Statistics*, 27: 175–180. <https://doi.org/10.1080/02331889508802520>.

ARNAB, R. (1998). Randomized response surveys. Optimum estimation of a finite population total. *Statistical Papers*, 39: 405–408. <https://doi.org/10.1007/BF02927102>.

ARNAB, R. (2017). *Survey Sampling, Theory and Applications.* London: Elsevier. ISBN 978-0-81148-1.

BOURKE, P., DALENIUS, T. (1976). Some new ideas in the realm of randomized inquiries. *Int. Statistical Review*, 44: 219–221. <https://doi.org/10.2307/1403280>.

BRICK, M. (2013). Unit nonresponse and weighting adjustments: A critical review. *J. Official Statistics*, 29: 329–353. <https://doi.org/10.2478/jos-2013-0026>.

CHAUDHURI, A. (1987). Randomized response surveys of a finite population: A unified approach with quantitative data. *J. Statistical Planning and Inference*, 15: 157–165. <https://doi.org/10.1016/0378-3758(86)90094-7>.

CHAUDHURI, A. (2017). *Randomized Response and Indirect Questioning Techniques in Surveys.* New York: Chapman and Hall/CRC. ISBN 978-11-3811542-2.

CHAUDHURI, A., CHRISTOFIDES, T. (2013). *Indirect Questioning in Sample Surveys.* Heidelberg: Springer. ISBN 978-3642-36275-0.

CHAUDHURI, A., MUKERJEE, R. (1988). *Randomized Response, Theory and Techniques.* New York: Marcel Dekker.

CHAUDHURI, A., CHRISTOFIDES, T., RAO, C. (2016). *Handbook of Statistics 34. Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques.* Amsterdam: Elsevier. ISBN 978-0444-63570-9.

DALENIUS, T., VITALE, R. (1979). A new randomized response design for estimating the mean of a distribution. In: JUREČKOVÁ, J., (eds.) *Contributions to Statistics*, 54–59. Praha: Academia.

ERIKSSON, S. (1973). A new model for randomized response. *Int. Statistical Review*, 41: 101–113. <https://doi.org/10.2307/1402791>.

FOX, J. (2016). *Randomized Response and Related Methods: Surveying Sensitive Data*, 2nd Ed. London: Sage. ISBN 978-1483-38103-9. <https://doi.org/10.4135/9781506300122>.

GESIS (2016). *EU-SILC 2016, Metadata for Official Statistics.* Mannheim: Gesis Missy. <https://www.gesis.org/en/missy/metadata/EU-SILC/2016/>.

GJESTVANGA, C., SINGH, R. (2009), An improved randomized response model: Estimation of mean. *J. Applied Statistics*, 36: 1361–1367. <https://doi.org/10.1080/02664760802684151>.

GREENBERG, B. (1971). Application of the randomized response technique in obtaining quantitative data. *J. American Statistical Association*, 66: 243–250. <https://doi.org/10.1080/01621459.1971.10482248>.

HORVITZ, D., THOMPSON, D. (1952). A generalization of sampling without replacement from a finite universe. *J. American Statistical Association*, 47: 663–685. <https://doi.org/10.1080/01621459.1952.10483446>.

ISO (2010). *ISO 28640:2010 (en). Random variate generation methods.* Geneve: ISO. <https://www.iso.org/obp/ui/#iso:std:42333:en>.

JACKSON, C. (2016). flexsurv: A platform for parametric modelling in R. *J. of Statistical Software*, 70: 1–33. <https://cran.r-project.org/web/packages/flexsurv/index.html>.

KIRCHNER, A. (2015). Validating sensitive questions: A comparison of survey and register data. *J. Official Statistics*, 31: 31–59. <https://doi.org/10.1515/jos-2015-0002>.

PFEFFERMANN, D., RAO, R.C. (2009a). *Handbook of Statistics 29A. Sample Surveys: Design, Methods and Application.* Amsterdam: Elsevier. ISBN 978-0444-53124-7.

PFEFFERMANN, D., RAO, R.C. (2009b). *Handbook of Statistics 29B. Sample Surveys: Inference and Analysis.* Amsterdam: Elsevier. ISBN 978-0444-53438-5.

R CORE TEAM (2021). *R: A language and environment for statistical computing.* Austria, Vienna: R Foundation for Statistical Computing. <https://www.R-project.org>.

SÄRNDAL, C., LUNDSTRÖM, S. (2005). *Estimation in Surveys with Nonresponse.* Chichester: J. Wiley and Sons. ISBN 978-0470-01133-1.

SÄRNDAL, C., SWENSSON, B., WRETMAN, J. (1992). *Model Assisted Survey Sampling.* Heidelberg: Springer. ISBN 978-0387-40620-6.

STEEH, C. (2001). Are they really as bad as they seem? Nonresponse rates at the end of the twentieth century. *J. Official Statistics*, 17: 227–247.

STOOP, I. (2005). *The Hunt for the Last Respondent: Nonresponse in Sample Surveys.* The Hague: Social and Cultural Planning Office of the Netherlands. ISBN 90-377-0215-5.

TIAN, G.-L., TANG, M.-L. (2014). *Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys.* Boca Raton: Chapman & Hall/CRC. ISBN 978-1439-85533-1.

TILLÉ, Y. (2006). *Sampling Algorithms.* New York: Springer. ISBN 978-0387-30814-2.

TILLÉ, Y. (2020). *Sampling and Estimation from Finite Populations.* New York: J. Wiley and Sons. ISBN 978-0470-68205-0.

TRAPPMANN, M. (2014). A new technique for asking quantitative sensitive questions. *J. Survey Statistics and Methodology*, 2: 58–77. <https://doi.org/10.1093/jssam/smt019>.

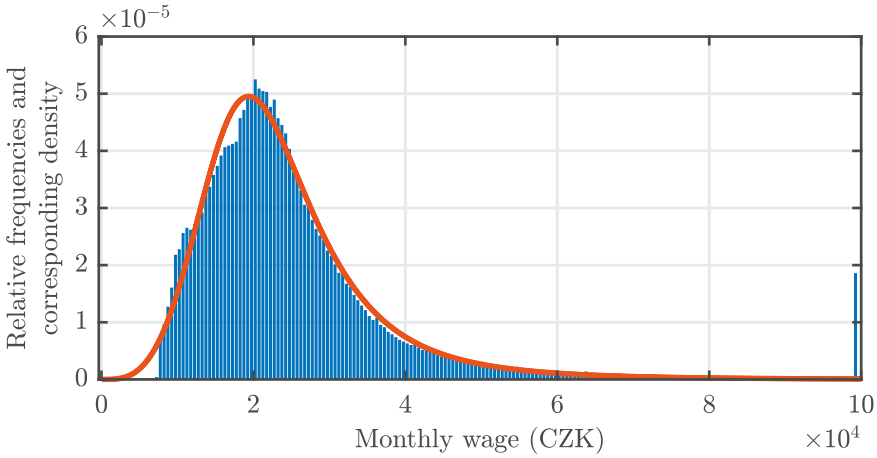
VRABEC, M., MAREK, L. (2016). Model of distribution of wages. *AMSE 2016, 19th Symp. Applications of Mathematics and Statistics in Economics*, Banská Štiavnica, 378–396. <https://amsesite.wordpress.com>.

WARNER, S. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. American Statistical Assoc.*, 60: 63–69. <https://doi.org/10.2307/2283137>.

WU, C., THOMPSON, M. (2020). *Sampling Theory and Practice.* Basel: Springer.

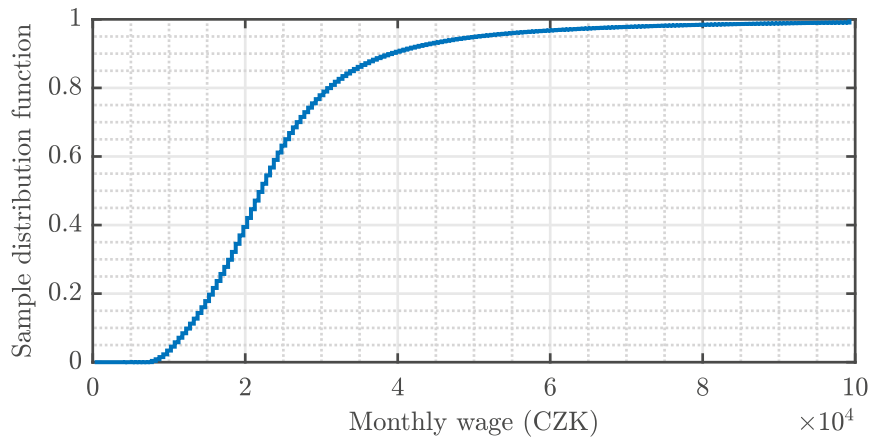
ANNEX

Figure 1 Probability histogram of monthly wages in the Czech Republic in the 2nd quarter of 2014, and the density (in red) of approximating model (22) with the parameters estimated by (23)



Source: Own construction

Figure 2 The sample distribution function of monthly wages in the Czech Republic in the 2nd quarter of 2014



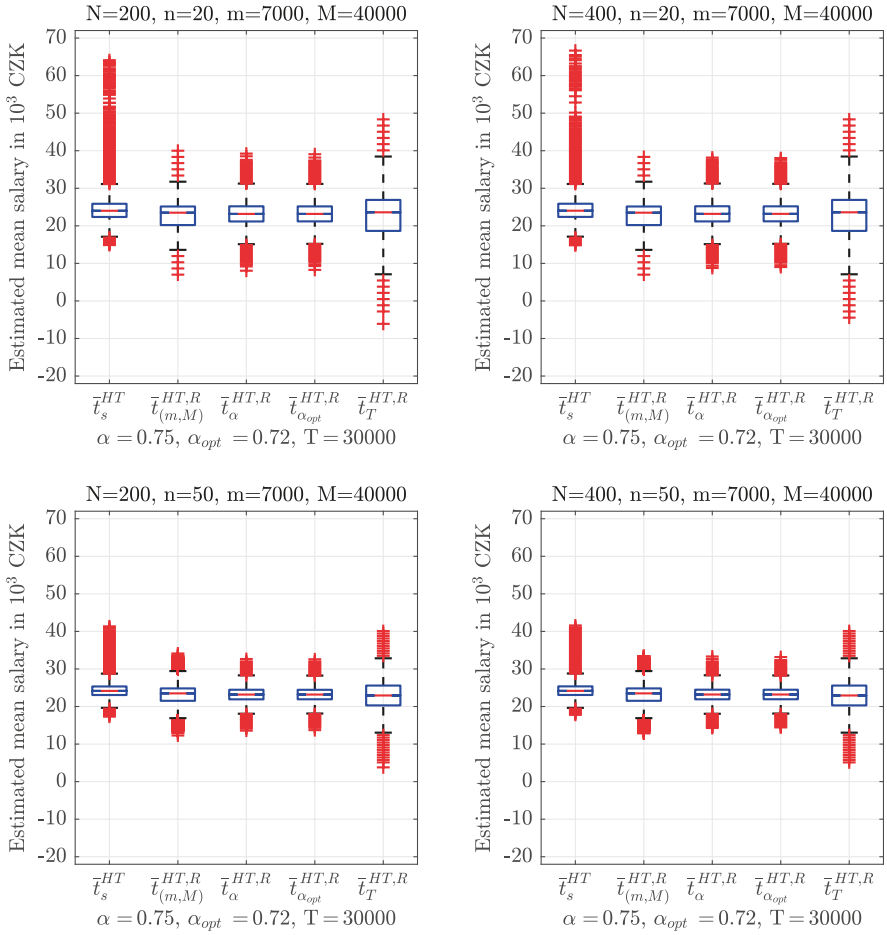
Source: Own construction

Table 1 Choice of tuning parameters for the simulations

m	M	T	α	α_{opt}
7 000	40 000	30 000	0.75	0.72
7 000	60 000	45 000	0.75	0.59
7 000	80 000	45 000	0.75	0.52

Source: Own construction

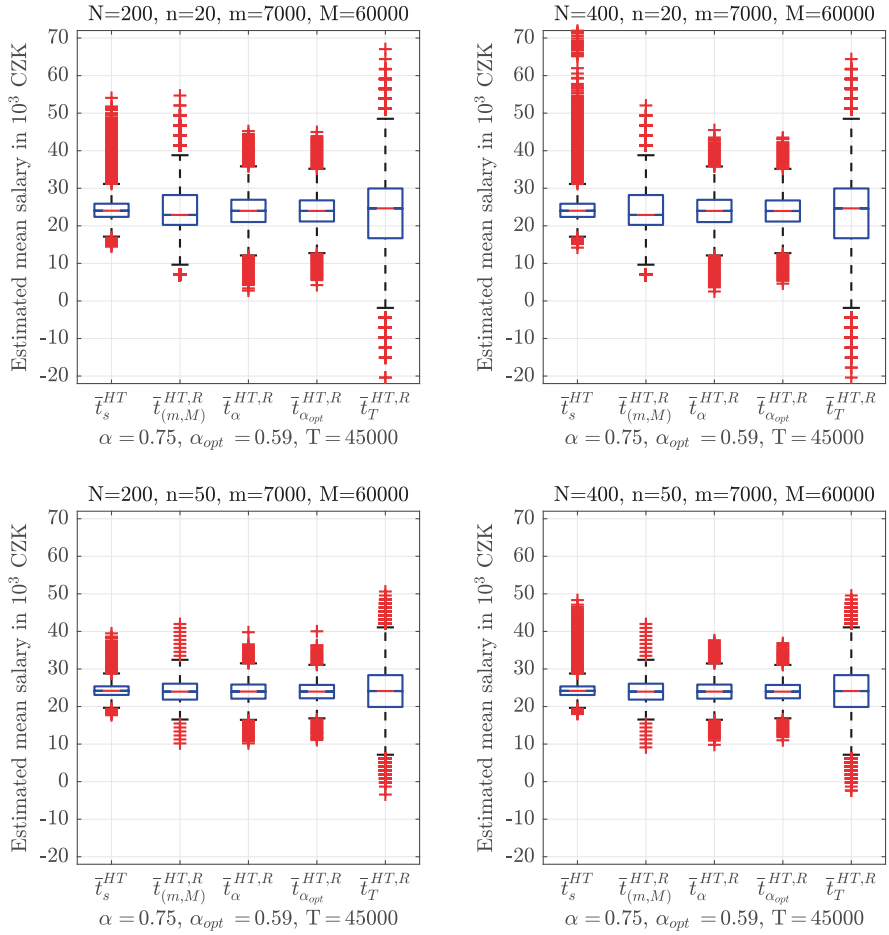
Figure 3 Behavior of considered estimators applied to different population and sample sizes



Parameters of the simulation, population N and sample sizes n : $(m, M) = (7\,000; 40\,000)$, $T = 30\,000$, $\alpha = 0.75$ and $\alpha_{opt} = 0.72$. To increase readability, we use $\overline{t}_{\alpha}^{HT,R}$, $\overline{t}_{\alpha_{opt}}^{HT,R}$ and $t_T^{HT,R}$ instead of $\overline{t}_{\alpha,(m,M)}^{HT,R}$, $\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$ and $t_{T,(m,M)}^{HT,R}$ in description of boxplots.

Source: Own construction

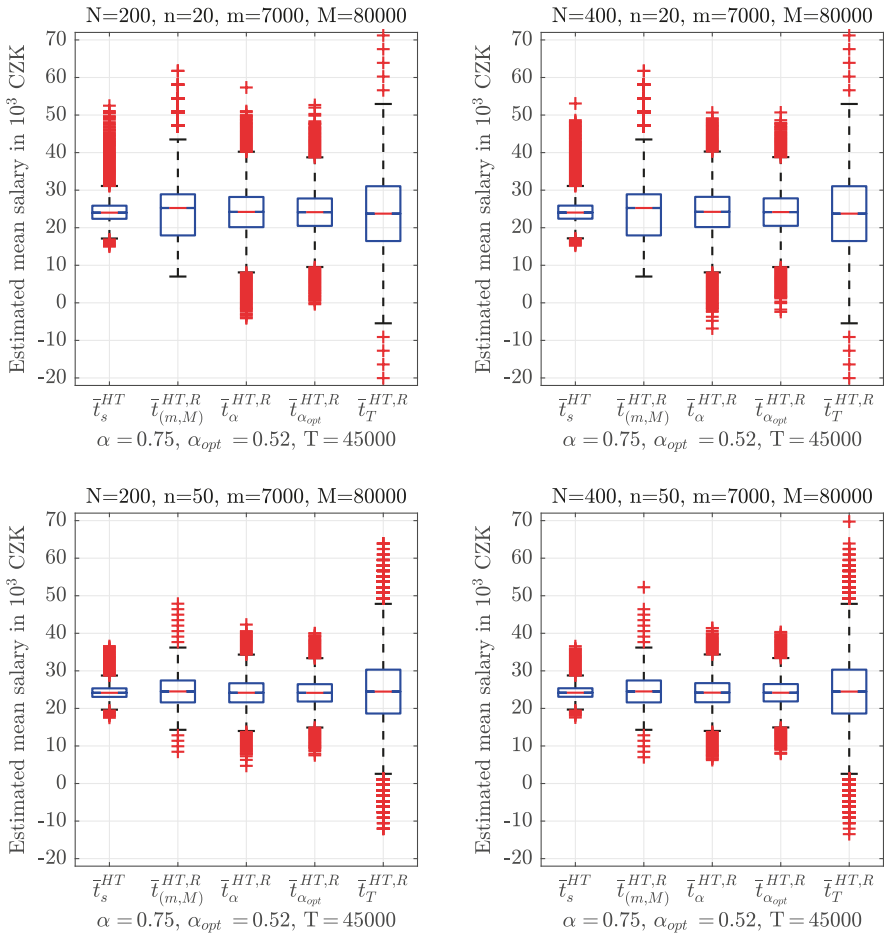
Figure 4 Behavior of considered estimators applied to different population and sample sizes



Parameters of the simulation, population N and sample sizes $n : (m, M) = (7\,000; 60\,000)$, $T = 45\,000$, $\alpha = 0.75$ and $\alpha_{opt} = 0.59$. To increase readability, we use $\overline{t}_{\alpha}^{HT,R}$, $\overline{t}_{\alpha_{opt}}^{HT,R}$ and $t_T^{HT,R}$ instead of $\overline{t}_{\alpha,(m,M)}^{HT,R}$, $\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$ and $t_{T,(m,M)}^{HT,R}$ in description of boxplots.

Source: Own construction

Figure 5 Behavior of considered estimators applied to different population and sample sizes



Parameters of the simulation, population N and sample sizes n : $(m, M) = (7\,000; 80\,000)$, $T = 45\,000$, $\alpha = 0.75$ and $\alpha_{opt} = 0.52$. To increase readability, we use $\overline{t}_{\alpha}^{HT,R}$, $\overline{t}_{\alpha_{opt}}^{HT,R}$ and $t_T^{HT,R}$ instead of $\overline{t}_{\alpha,(m,M)}^{HT,R}$, $\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$ and $t_{T,(m,M)}^{HT,R}$ in description of boxplots.

Source: Own construction

Table 2 Numerical results of simulations

		$N = 200$		$N = 400$	
Estimator		$n = 20$	$n = 50$	$n = 20$	$n = 50$
$\overline{t}_s^{HT}$	mean	24.270	24.272	24.287	24.288
	sd	2.782	1.757	2.773	1.758
$\overline{t}_{(m,M)}^{HT,R}$	mean	23.189	23.192	23.203	23.205
	sd	3.687	2.333	3.690	2.336
$\overline{t}_{\alpha,(m,M)}^{HT,R}$	mean	23.192	23.194	23.206	23.207
	sd	3.000	1.897	3.001	1.902
$\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$	mean	23.192	23.194	23.206	23.207
	sd	2.965	1.875	2.966	1.880
$\overline{t}_{T,(m,M)}^{HT,R}$	mean	23.185	23.189	23.199	23.202
	sd	6.066	3.836	6.068	3.837

The mean estimated salaries (in 10^3 CZK) and the corresponding sample standard deviations (in 10^3 CZK) for different population sizes N and sample sizes n. Random numbers Υ_i are generated from the uniform distribution on the interval $[m, M] = [7\,000; 40\,000]$, $T = 30\,000$, $\alpha = 0.75$, $\alpha_{opt} = 0.72$, 1000 simulated populations, 1000 replications of each. Means and standard deviations (sd) were averaged over 1000×1000 random samples.

Table 3 Numerical results of simulations

		$N = 200$		$N = 400$	
Estimator		$n = 20$	$n = 50$	$n = 20$	$n = 50$
$\overline{t}_s^{HT}$	mean	24.297	24.301	24.288	24.290
	sd	2.773	1.758	2.813	1.779
$\overline{t}_{(m,M)}^{HT,R}$	mean	23.983	23.984	23.965	23.974
	sd	5.530	3.501	5.529	3.495
$\overline{t}_{\alpha,(m,M)}^{HT,R}$	mean	23.974	23.976	23.956	23.965
	sd	4.401	2.786	4.398	2.780
$\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$	mean	23.976	23.977	23.958	23.967
	sd	4.164	2.637	4.161	2.631
$\overline{t}_{T,(m,M)}^{HT,R}$	mean	23.991	23.992	23.973	23.982
	sd	9.066	5.729	9.067	5.726

The mean estimated salaries (in 10^3 CZK) and the corresponding standard deviations (in 10^3 CZK) for different population sizes N and sample sizes n. Random numbers Υ_i are generated from the uniform distribution on the interval $[m, M] = [7\,000; 60\,000]$, $T = 45\,000$, $\alpha = 0.75$, $\alpha_{opt} = 0.59$, 1000 simulated populations, 1000 replications of each. Means and standard deviations (sd) were averaged over 1000×1000 random samples.

Table 4 Numerical results of simulations

		$N = 200$		$N = 400$	
Estimator		$n = 20$	$n = 50$	$n = 20$	$n = 50$
$\overline{t}_s^{HT}$	mean	24.275	24.273	24.299	24.299
	sd	2.765	1.739	2.753	1.737
$\overline{t}_{(m,M)}^{HT,R}$	mean	24.138	24.140	24.158	24.168
	sd	6.911	4.372	6.921	4.378
$\overline{t}_{\alpha,(m,M)}^{HT,R}$	mean	24.145	24.146	24.165	24.174
	sd	5.962	3.770	5.950	3.767
$\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$	mean	24.143	24.145	24.163	24.173
	sd	5.404	3.417	5.398	3.417
$\overline{t}_{T,(m,M)}^{HT,R}$	mean	24.136	24.137	24.156	24.165
	sd	13.018	8.236	13.036	8.244

Numerical results of simulations. The mean estimated salaries (in 10^3 CZK) and the corresponding standard deviations (in 10^3 CZK) for different population sizes N and sample sizes n. Random numbers Υ_i are generated from the uniform distribution on the interval $[m, M] = [7\,000; 80\,000], T = 45\,000$, $\alpha = 0.75$, $\alpha_{opt} = 0.52$, 1000 simulated populations, 1000 replications of each. Means and standard deviations (sd) were averaged over 1000×1000 random samples.

Source of Tables 2 – 4: Own construction

Recent Events

Conferences

The ***18th IAOS (International Association for Official Statistics) Conference*** took place ***from 26th to 28th April 2022 in Kraków, Poland.*** More at: <*https://www.iaos2022.pl*>.

The ***10th Q2022 Conference (European Conference on Quality in Official Statistics)*** was held ***during 8–10 June 2022 in Vilnius, Lithuania.*** More at: <*https://q2022.stat.gov.lt*>.

The ***24th AMSE Scientific Conference (Applications of Mathematics and Statistics in Economics)*** will take place ***from 31st August to 4th September 2022 in Velké Losiny, Czechia.*** More at: <*http://www.amse-conference.eu*>.

Central Statistical Library – invitation

The book collection of the Central Statistical Library (in the main CZSO building in Prague 10, Skalka, Czech Republic) has been built for more than 200 years.

It is the only professional statistical library in the Czech Republic. It offers a wide selection of current statistical literature, statistical periodicals (from the Czech Republic and abroad), unique historical statistical publications, yearbooks, maps or lexicons of municipalities, and also sale of statistical publications and periodicals.

Users can find all information about our publications in the online catalog.

Czech Statistical Office – Central Statistical Library
Adress: Na padesátém 81 | 100 82 Prague 10 | Czech Republic
E-mail: knihovna@czso.cz | <*https://www.czso.cz/csu/czso/central_statistical_library*>

Papers
We publish articles focused at theoretical and applied statistics, mathematical and statistical methods, conception of official (state) statistics, statistical education, applied economics and econometrics, economic, social and environmental analyses, economic indicators, social and environmental issues in terms of statistics or economics, and regional development issues.

The journal of *Statistika* has following sections:

The *Analyses* section publishes complex and advanced analyses based on the official statistics data focused on economic, environmental, social and other topics. Papers shall have up to 12 000 words or up to 20 1.5-spaced pages.

Discussion brings the opportunity to openly discuss the current or more general statistical or economic issues, in short what the authors would like to contribute to the scientific debate. Contribution shall have up to 6 000 words or up to 10 1.5-spaced pages.

In the *Methodology* section we publish articles dealing with possible approaches and methods of researching and exploring social, economic, environmental and other phenomena or indicators. Articles shall have up to 12 000 words or up to 20 1.5-spaced pages.

Consultation contains papers focused primarily on new perspectives or innovative approaches in statistics or economics about which the authors would like to inform the professional public. Consultation shall have up to 6 000 words or up to 10 1.5-spaced pages.

Book Review evaluates selected titles of recent books from the official statistics field (published in the Czech Republic or abroad). Reviews shall have up to 600 words or 1–2 1.5-spaced pages.

The *Information* section includes informative (descriptive) texts, information on latest publications (issued not only by the Czech Statistical Office), or recent and upcoming scientific conferences. Recommended range of information is 6 000 words or up to 10 1.5-spaced pages.

Language
The submission language is English only. Authors are expected to refer to a native language speaker in case they are not sure of language quality of their papers.

Recommended Paper Structure
Title – Authors and Contacts – Abstract (max. 160 words) Keywords (max. 6 words / phrases) – Introduction – 1 Literature survey – 2 Methods – 3 Results – 4 Discussion –Conclusion – Acknowledgments – References – Annex (Appendix). Tables and Figures (for print at the end of the paper; for the review process shall be placed in the text).

Authors and Contacts
Rudolf Novak,[1] Institution Name, City, Country
Jonathan Davis, Institution Name, City, Country
[1] Address. Corresponding author: e-mail: rudolf.novak @domainname.cz, phone: (+420) 111 222 333.

Main Text Format
Times 12 (main text), 1.5 spacing between lines. Page numbers in the lower right-hand corner. *Italics* can be used in the text if necessary. *Do not* use **bold** or underline in the text. Paper parts numbering: 1, 1.1, 1.2, etc.

Headings
1 FIRST-LEVEL HEADING (Times New Roman 12, bold)
1.1 Second-level heading (Times New Roman 12, bold)
1.1.1 Third-level heading (Times New Roman 12, bold italic)

Footnotes
Footnotes should be used sparingly. Do not use endnotes. Do not use footnotes for citing references.

References in the Text
Place references in the text enclosing authors' names and the year of the reference, e.g. „…White (2009) points out that…", „…recent literature (Atkinson and Black, 2010a, 2010b, 2011; Chase et al., 2011: 12–14) conclude…". Note the use of alphabetical order. Between the names of two authors please insert „and", for more authors we recommend to put „et al.". Include page numbers if appropriate.

List of References
Arrange list of references alphabetically. Use the following reference styles: [book] HICKS, J. (1939). *Value and Capital: an Inquiry into Some Fundamental Principles of Economic Theory.* 1st Ed. Oxford: Clarendon Press. [chapter in an edited book] DASGUPTA, P. et al. (1999). Intergenerational Equity, Social Discount Rates and Global Warming. In: PORTNEY, P., WEYANT, J. (eds.) *Discounting and Intergenerational Equity,* Washington, D.C.: Resources for the Future. [on-line source] CZECH COAL. (2008). *Annual Report and Financial Statement 2007* [online]. Prague: Czech Coal. [cit. 20.9.2008]. <http://www.czechcoal.cz/cs/ur/zprava/ur2007cz.pdf>. [article in a journal] HRONOVÁ, S., HINDLS, R., ČABLA, A. (2011). Conjunctural Evolution of the Czech Economy. *Statistika: Statistics and Economy Journal,* 91(3): 4–17. [article in a journal with DOI]: Stewart, M. B. (2004). The Employment Effect of the National Minimum Wage [online]. *The Economic Journal,* 114(494): 110–116. <http://doi.org/10.1111/j.0013-0133.2003.0020.x>.
Please **add DOI numbers** to all articles where appropriate (prescribed format = link, see above).

Tables
Provide each table on a separate page. Indicate position of the table by placing in the text "insert Table 1 about here". Number tables in the order of appearance Table 1, Table 2, etc. Each table should be titled (e.g. Table 1 Self-explanatory title). Refer to tables using their numbers (e.g. see Table 1, Table A1 in the Annex). Try to break one large table into several smaller tables, whenever possible. Separate thousands with a space (e.g. 1 528 000) and decimal points with a dot (e.g. 1.0). Specify the data source below the tables.

Figures
Figure is any graphical object other than table. Attach each figure as a separate file. Indicate position of the figure by placing in the text "insert Figure 1 about here". Number figures in the order of appearance Figure 1, Figure 2, etc. Each figure should be titled (e.g. Figure 1 Self-explanatory title). Refer to figures using their numbers (e.g. see Figure 1, Figure A1 in the Annex).

Figures should be accompanied by the *.xls, *.xlsx table with the source data. Please provide cartograms in the vector format. Other graphic objects should be provided in *.tif, *.jpg, *.eps formats. Do not supply low-resolution files optimized for the screen use. Specify the source below the figures.

Formulas
Formulas should be prepared in formula editor in the same text format (Times 12) as the main text and numbered.

Paper Submission
Please email your papers in *.doc, *.docx or *.pdf formats to statistika.journal@czso.cz. All papers are subject to double-blind peer review procedure. Articles for the review process are accepted continuously and may contain tables and figures in the text (for final graphical typesetting must be supplied separately as specified in the instructions above). Please be informed about our Publication ethics rules (i.e. Authors responsibilities) published at: <*http://www.czso.cz/statistika_journal*>.

Managing Editor: Jiří Novotný
Phone: (+420) 274 054 299 | **fax:** (+420) 274 052 133
E-mail: statistika.journal@czso.cz | **web:** www.czso.cz/statistika_journal
Address: Czech Statistical Office | Na padesátém 81 | 100 82 Prague 10 | Czech Republic

Subscription price (4 issues yearly)
CZK 66 per copy + postage.
Printed copies can be bought at the Publications Shop of the Czech Statistical Office (CZK 66 per copy).
Address: Na padesátém 81 | 100 82 Prague 10 | Czech Republic

Subscriptions and orders
Czech Statistical Office | Information Services
Na padesátém 81 | 100 82 Prague 10 | Czech Republic
Phone: (+420) 274 052 733, (+420) 274 052 783
E-mail: objednavky@czso.cz

Design: Toman Design
Layout: Ondřej Pazdera
Typesetting: Václav Adam
Print: Czech Statistical Office

102nd year of the series of professional statistics and economy journals of the State Statistical Service in the Czech Republic: ***Statistika*** (since 1964), ***Statistika a kontrola*** (1962–1963), ***Statistický obzor*** (1931–1961) and ***Československý statistický věstník*** (1920–1930).

Published by the Czech Statistical Office
ISSN 1804-8765 (Online)
ISSN 0322-788X (Print)
Reg. MK CR E 4684