STATISTIKA

STATISTICS
AND ECONOMY
JOURNAL

VOL. **104** (1) 2024

CZECH
STATISTICAL
OFFICE

EDITOR-IN-CHIEF

EDITORIAL BOARD

EXECUTIVE BOARD

MANAGING EDITOR

CONTENTS

ANALYSES

METHODOLOGY

CONSULTATION

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: Statistics and Economy Journal is professional double-blind peer reviewed open access journal included e.g. in the *Web of Science Emerging Sources Citation Index* (since 2016) of the **Web of Science Core Collection** database (**Impact Factor** 2022: **0.2),** in the international citation database of peer-reviewed literature **Scopus** (since 2015, SJR 2022 = 0.179, CiteScore 2022 = 0.6), and others. Since 2011, Statistika has been published quarterly in English only.

Publisher

ThThe 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

EU Countries Tax Evasion Clustering

Mária Stachová[1] | *Matej Bel Universtity, Banská Bystrica, Slovakia*
Kamil Ščerba[2] | *Matej Bel Universtity, Banská Bystrica, Slovakia*

Received 12.7.2023, **Accepted** (reviewed) 4.10.2023, **Published** 15.3.2024

Abstract

The problem of tax evasion is becoming more and more prominent, and it is causing large losses in individual state budgets. However, European policies have taken significant steps to eliminate them over the last two decades. National policies also contribute to reducing the estimated extent of tax evasion, as measured by Schneider's MIMIC model. However, these tendencies are not the same in individual countries.

Therefore, in our paper, we try to find pattern in tax evasion time trajectories of European Union countries and thus to create homogenous clusters that include countries with similar tax evasion situation. To meet our goal, we use panel data clustering method on our data with information about tax evasion of EU countries from 2000 to 2019. As a next step, we compare created cluster from the perspective of quality indicators of public institutions.

Keywords	***DOI***	***JEL code***
Tax evasion, clustering, panel data	*https://doi.org/10.54694/stat.2023.34*	*H26, C38*

INTRODUCTION

The shadow economy is known under different names, namely the hidden economy, the gray economy, the black economy, or the economy of scarcity, or the informal economy. These synonyms point to some characteristics of the shadow economy. However, we will stick with the label shadow economy and try to approximate this term. Schneider and Enste (2000, p. 78) quote a definition attributed to several authors, which defines the shadow economy as "economic activity that is unregistered but would otherwise contribute to increasing the gross national product." Orviská (2005, p. 7) clarifies that "these are activities that are illegal, such as illegal production, sale and smuggling of drugs, but also activities that are not inherently illegal, but which individuals try to conceal for various reasons, e.g. due to the avoidance of paying taxes." According to Medina and Schneider (2018), all the reasons that lead to hiding economic activities can be divided into monetary, regulatory, and institutional. Among the monetary reasons, they include avoiding paying taxes and all social security contributions. Regulatory reasons mainly include avoiding government bureaucracy and institutional reasons include corruption, low quality of institutions and weak rule of law (Medina and Schneider, 2018).

[1] 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, phone: (+421)484466620.
[2] Department of Finance and Accounting, Faculty of Economics, Matej Bel University, Tajovského 10, 975 90 Banská Bystrica, Slovakia.

Although the shadow economy and tax evasion are not the same, in most cases activities in the shadow economy involve the evasion of direct or indirect taxes. Therefore, the authors advocate the opinion that the factors affecting tax evasion are most likely to affect the shadow economy as well (Schneider and Buehn, 2016). In addition, the size of the shadow economy is a basic input for estimating the extent of tax evasion, and thus also for decisions about its adequate control (Medina and Schneider, 2018). Šimonová (2017, p. 6) defines tax evasion as "the resulting effect of unconscious or purposeful reduction of tax obligations by taxpayers." Bánociová and Vravec (2015, p. 2) define them as "the result of illegal economic behavior of tax subjects oriented towards minimization, or complete elimination of tax liability towards the state, or obtaining excessive VAT deductions." Kiabel and Nwokah (2009, p. 53) argue that "tax evasion is a direct and dishonest step by which a taxpayer seeks to reduce his tax liability by illegal means." They also add that tax evasion is achieved by deliberate by an action that constitutes a criminal act according to the law. Tax evasion is the cause of large public deficits by causing a lack of financial resources and the state is unable to cover its costs.

1 DATA AND METHODOLOGY

Our data consists of 27 EU countries and their characteristics: tax evasion index over the years 2000–2019 (Schneider, 2022) measured as a percentage of GDP and quality indicators of public institutions (Table 1). Because we used time trajectories of tax evasion indices as cluster variables, we needed to deal with longitudinal character of data. For this purpose, we select the K-means algorith applied in the "kml" package (Genolini et al., 2015) of statistical system R, that is modified for panel data. This algorithm is based on the original K-means clustering (MacQueen, 1967).

This method minimizes the utility function iteratively for the time t, N objects according to an assumption of C clusters. The utility function can be expressed as follows:

$$\min \sum_{i=1}^{N} \sum_{c=1}^{C} u_{ict} d_{ict}^2, \tag{1}$$

where u_{ict} is a degree of appropriateness of the i-th object into the K-th cluster in the time t with conditions:

$$\sum_{c=1}^{C} u_{ict} = 1, \forall i, t, \tag{2}$$

$$\forall u_{ict} : u_{ict} = \begin{cases} 1 & \| \mathbf{x}_t - \mathbf{h}_c \| = \arg\min_i \| \mathbf{x}_t - \mathbf{h}_c \|. \\ 0 & \text{elsewhere} \end{cases}$$

We used the Euclidean distance $d_{ict} = ||\mathbf{x}_{it} - \mathbf{h}_{ct}||$ between i-th vector of objects $\mathbf{x}_{it} = (x_{i1t}, \ldots, x_{ijt}, \ldots, x_{iJt})'$ and K-th centroid $\mathbf{h}_{ct} = (h_{c1t}, \ldots, h_{cjt}, \ldots, h_{cJt})'$ in the time t. We applied the algorithm to the standardized values of variables.

When measuring the estimated extent of the shadow economy/tax evasion described by tax evasion index, we use the MIMIC model (multiple indicators, multiple causes estimation) created by Schneider (Schneider and Enste, 2000). The model considers the shadow economy as a latent variable that is influenced by the variables entering the model as the main factors that cause the growth or decline of the shadow economy. According to Orviská (2005), the factors used in the MIMIC model include the unemployment rate, tax morale, disposable income per inhabitant, tax and social burden, and the like. On the other hand, the shadow economy has an impact on the official economy, in the form of, for example, changes in real GDP, in the rate of participation in the labor market, in the number of hours worked, and others.

The clusters created by their tax evasion index trajectories were analyzed in the next step in our study. We added variables that fall into the group of non-economic indicators and evaluate the quality of institutions. They are indicators/indexes that are captured by the research institute of the World Bank

and we know them under the term Global Governance Indicators (WGI). These indicators include the right to express oneself, political stability and the absence of violence, government efficiency, the quality of regulatory measures, the rule of law and the control of corruption. According to Hiwatari (2014), these are aggregate indicators based on many underlying assets taken from a wide range of existing data. The data reflects views on quality from various research institutes, non-governmental organizations, international organizations, and private sector companies around the world. In the following Table 1 is list of these indicators with short description and its expected impact on tax evasion.

Table 1 Global Governance Indicators

Indicator	Short description	Expected impact	Source
Voice and accountability (democracy)	It determines the degree to which the citizens of the country are able to participate in the choice of their government, but also the degree of freedom of speech, freedom of association and freedom of the media. The index is measured on a scale from –2.5 to 2.5.	Positive	The World Bank (2021)
Political stability and absence of violence/terrorism	It evaluates the provision of freedom in the country, the provision of security, including the elimination of threats of civil discontent through protests, up to terrorism. It is rated on a scale from –2.5 to 2.5.	Positive	The World Bank (2021)
Government effectiveness	It captures the quality of public services, the civil service and its degree of independence from political pressures, the quality of policy implementation and the credibility of the government. It is measured on a scale from –2.5 to 2.5.	Positive	The World Bank (2021)
Regulatory quality	It assesses perceptions of the government's ability to formulate and implement policies and regulations that enable and support private sector development. It is expressed on a scale from –2.5 to 2.5.	Positive	The World Bank (2021)
Rule of law	It captures perceptions of the extent to which law enforcement officers trust and follow society's rules, particularly regarding the quality of contract enforcement, property rights, the police and courts, and the likelihood of crime and violence. It is rated on a scale from –2.5 to 2.5.	Positive	The World Bank (2021)
Control of corruption	It assesses perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as the pursuit of private interests. It is measured on a scale from –2.5 to 2.5.	Positive	The World Bank (2021)

Source: Authors' work according to World Bank (2021)

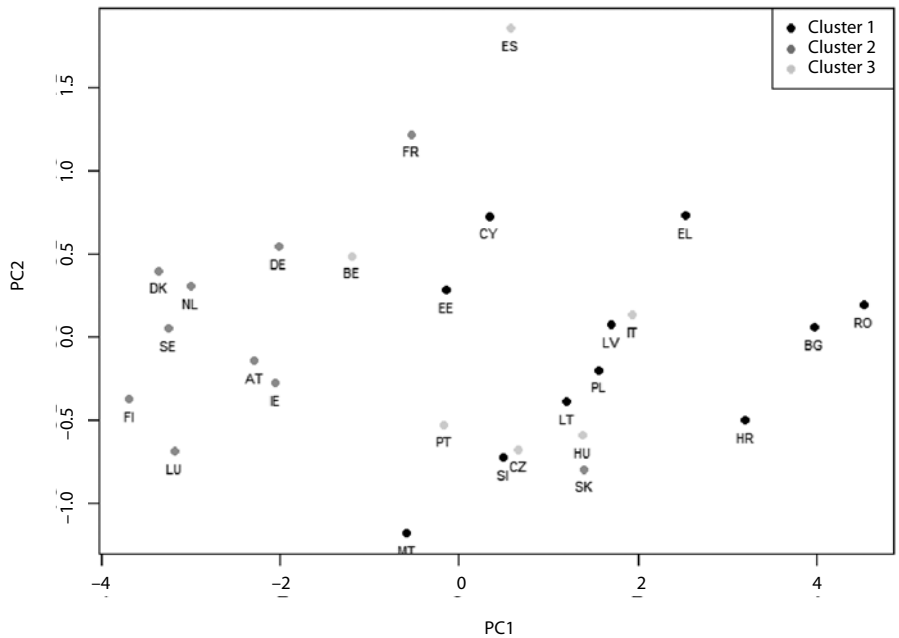
Bruinshoofd (2016) examined the quality of these indicators, but also the indicator of the difficulty of doing business in the European Union compared to the rest of the world. According to the scheme, European institutions outperform the world average in all seven components. The primacy is somewhat less pronounced in terms of political stability and absence of violence. We already mentioned above that the EU average for this indicator is the lowest among all indicators of the quality of institutions.

2 RESULTS AND DISCUSSION

Based on Calinski and Harabatz criterium, the best number of clusters is 3. These clusters contain countries with a similar range of the estimated range of tax evasion for the monitored period from 2000 to 2019.

To visualize the results of clustering we used principal components analysis and created the scatterplot of the first two principal components of quality indices of EU countries. In Figure 1 can be seen the classification of EU countries into the clusters.

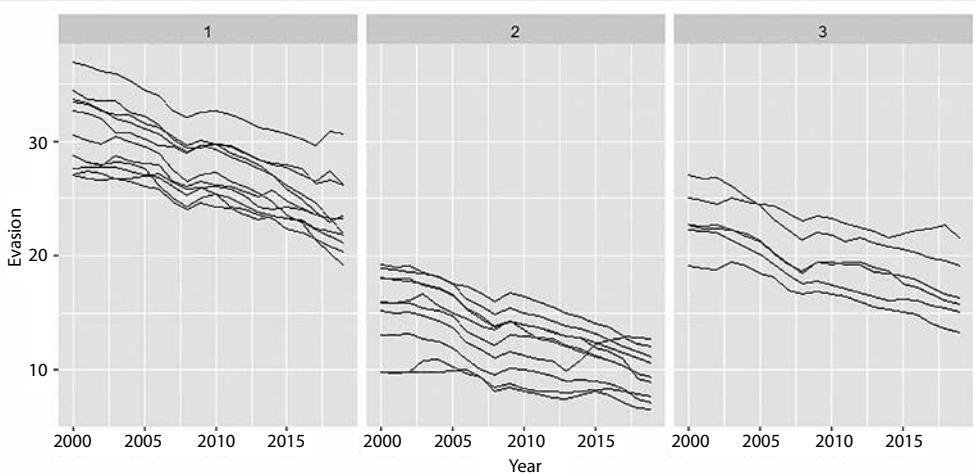
Figure 1 Classification of EU countries into the clusters



Source: Authors' work

The first cluster includes 11 countries, the second one 10 countries and the third one includes 6 countries. The time trajectories of evasion index of created clusters are shown in Figure 2.

Figure 2 The time trajectories of evasion index in created clusters



Source: Authors' work

We can see the reducing trend for the entire observed period in all countries. We know from the data that the exception is the year 2009, when there was an increase in this value for all monitored countries, probably due to the global crisis. However, most countries (except Romania, Bulgaria, Greece, Latvia, and Lithuania belonging to one cluster) saw a decrease in the estimated rate of tax evasion already in the following year 2010. We evaluate this as evidence of a relatively quick recovery from the economic crisis.

If we focus on individual member states in 2019, we see that the estimated rate of tax evasion is lower than the EU average, mainly in the states that are among the older members of the Union (except Slovakia and the Czech Republic). This suggests that tax evasion is a bigger problem in countries that have been in the EU for a shorter period. When following the older members of the EU, we can also notice a lower estimated range of tax evasion for states located further north (e.g. the Netherlands, Denmark, Ireland) than for states located in the south of Europe (e.g. Greece, Portugal, Spain). A good example from the north are the Scandinavian countries (Denmark, Sweden, and Finland). They are among the countries with the lowest estimated range of tax evasion and have even significantly improved in this value over the monitored period. The tax systems of Scandinavia are characterized by a high level of taxation. However, according to Ďurinova (2013), this tax burden is mitigated by many tax reliefs and exemptions, as well as by the provision of various services, especially in the social field. Thanks to the good social system, they are also considered as so-called welfare states. In addition, they have developed economies and a high standard of living of the inhabitants. Conversely, the highest estimated rate of tax evasion can be observed in Croatia (26.22%), Romania (26.15%) and Bulgaria (30.59%). These are the states that joined the EU as the last ones, and they are in the south of Europe.

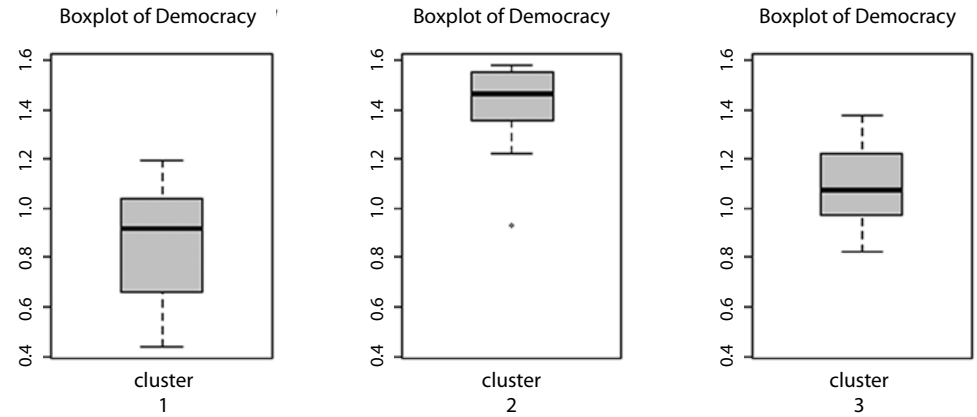
We see the most significant improvement over the monitored period for the Baltic countries (Latvia up to 13.07 p.p.), while one of the lowest improvements for Hungary (by 3.59 p.p.). It is Hungary that has recently recorded an increase in the estimated scope of tax evasion since 2015. The author Benyik (2019) attributes this fact to the strong position of the Orbán government, which is characterized by a high level of corruption that undermines democracy and economic growth.

In the next step of our study, we analyze created clusters based on non-economic indicators that evaluate the quality of institutions.

The first indicator describes the level of democracy in the country. In connection with this indicator, the authors Feld and Tyran (2002) investigated the possibility to participate in the electoral process and its connection with compliance with tax regulations. Based on research, they state that participation in the electoral process is positively associated with compliance with tax regulations in society. They therefore agree with the view that the greater the right to express oneself, the lower the level of tax evasion. The authors' claims are also supported by our analysis, as we see in Figure 3 that cluster 2 shows the lowest values of the estimated extent of tax evasion and, on the other hand, the highest values of the democracy index with the smallest dispersion and values that move around the median. The opposite is shown by the first cluster, with the highest percentage of tax evasion, where democracy or the right to express oneself is the lowest and with greater dispersion in individual countries.

Second indicator, political stability, and absence of violence, is described by Hiwatari (2014) as the probability that a government will be overthrown by unconstitutional or violent means, including politically motivated violence and terrorism. Katz and Owen (2013) argue that the political uncertainties faced by taxpayers' stem from not knowing which political party will take control and not knowing whether the new government would catch them in tax evasion. Yamen et al. (2018) adds that based on how taxpayers perceive these political uncertainties; they may try to protect their wealth by concealing their true income and tax evasion. These authors expect tax evasion to decrease with political stability. Again, this statement is confirmed, as it can be seen in Figure 4, to the greatest extent in cluster two, where we can assume

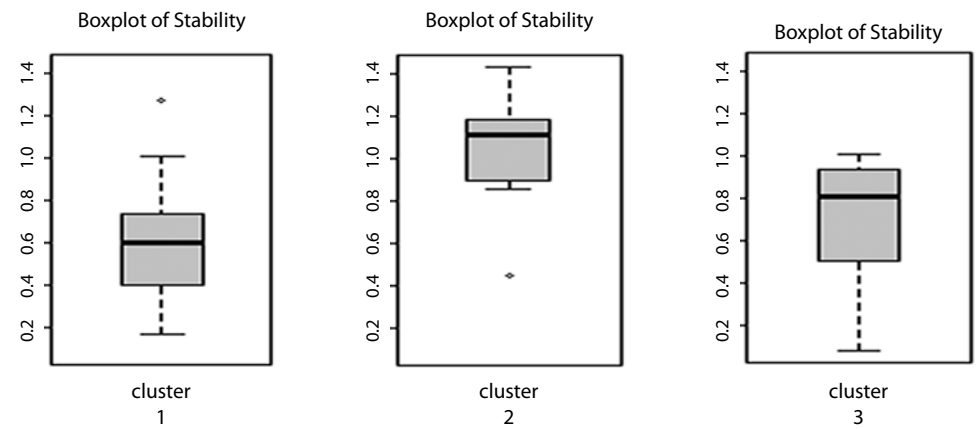
Figure 3 Boxplots of level of Democracy index in the created clusters



Source: Authors' work

that political stability and the absence of violence largely influenced the estimated extent of tax evasion, and the indicator values are the highest (from 0.9 to 1.2). Relatively worst is cluster one, where the estimated range of tax evasion is the highest and stability is the lowest, which, on the other hand, does not show such a large variance from the median, as, for example, in cluster 3.

Figure 4 Boxplots of level of Stability index in the created clusters



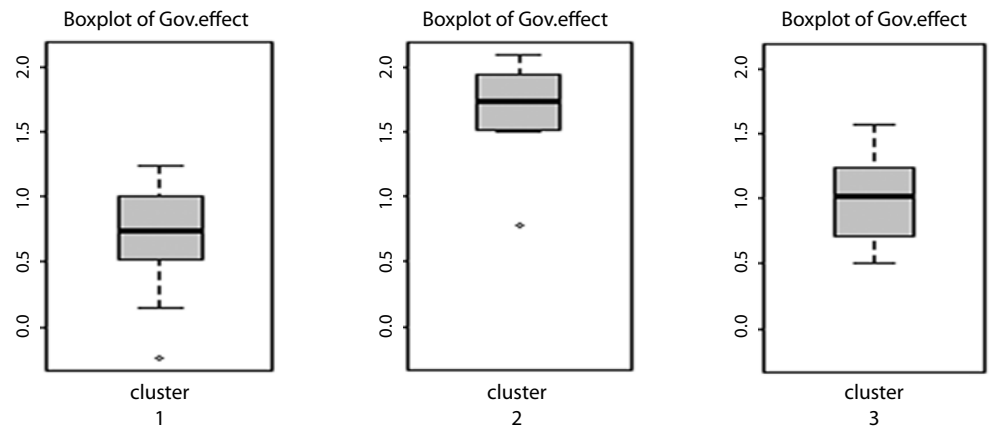
Source: Authors' work

Another indicator is the efficiency of the government, which is reflected in the quality of public services. Arsic et al. (2015) have the opinion that the quality of public services is negatively correlated with the extent of tax evasion, and a higher quality of public services means a greater willingness of the public to pay taxes, as these taxes go towards financing goods and services that meet their needs. Authors Hanousek and Palda (2004) consider tax evasion as a form of citizen dissatisfaction with the relevant government and the quality of government services. Therefore, they recommend that honest and effective governments that want to increase tax compliance pay special attention to informing their subjects about what the government is doing for them. Schneider and Buehn (2016) are also inclined

to the opinion that tax subjects often turn to illegal behavior precisely because of the inefficient provision of trust goods. They advise governments to make more efforts to strengthen institutions and create a fiscal policy that is closer to voter preferences.

In Figure 5, this index shows us the possible dependence of the relationship between tax evasion and government efficiency. The difference in the individual clusters is also noticeable, where the values of the indicator range from 0.5 to 1.2 for the first and third clusters, which is a relatively good value, but there is still a room for improvement. However, as we can see, these clusters still show relatively high values of the estimated range of tax evasion. However, we see the opposite in the second cluster, where the authors' assertions are confirmed and the government efficiency indicator shows high values, which we assume, affects the extent of the estimated scope of tax evasion. We know from the study that Finland, Sweden, Denmark, Germany, and other countries from the second cluster show high efficiency, but also the quality of the government, which was also confirmed by the results of the analysis.

Figure 5 Boxplots of level of Government efficiency index in the created clusters

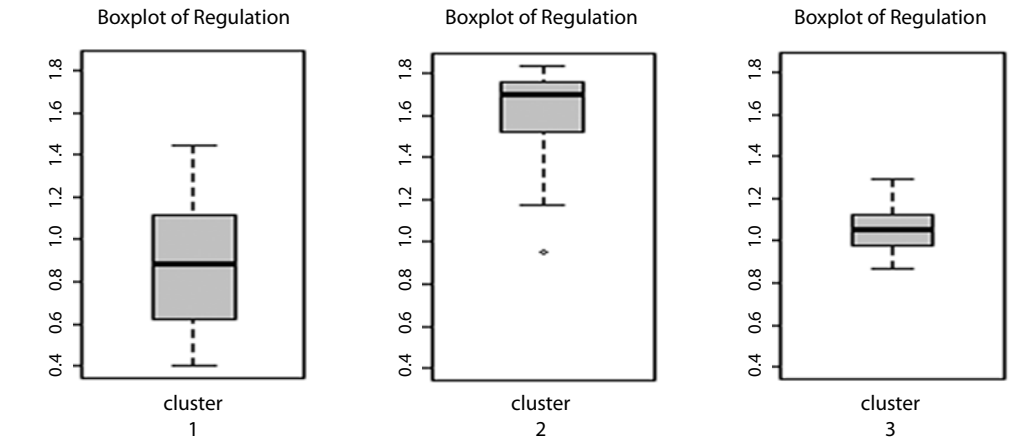


Source: Authors' work

We are also interested in the quality of regulatory measures, i.e., the government's ability to formulate and implement reliable regulations. According to Yamen et al. (2018), governments should establish clear and fair regulations to establish and maintain a good relationship with taxpayers. This could contribute to reducing the negative attitude towards the government. Also, even a simple tax system increases compliance with tax regulations. According to this, quality regulation should also lead to lower tax evasion. We see in Figure 6, quite a large heterogeneity in the results, either between the clusters or within the clusters (mainly in the 1st cluster). This index largely reflects the government's ability to create an environment in which tax evasion is better fought. As a verification of this statement and at the same time of our results, we see the results of cluster 1. This cluster includes countries such as Romania, Bulgaria, Croatia, and others, which in several studies came out as countries where governments take very few measures, or no measures to combat tax evasion. Of course, this fact changes over time, as the EU itself has begun to fight this disease to a large extent, which adopts measures across the board and countries must (or should) apply them. Whether it was the EU action plan in the fight against tax evasion, or the BEPS measures themselves from 2015 (OECD, 2019), which should eliminate tax evasion.

Another indicator is government laws. Hiwatari (2014) describes government law as the measure to which law enforcement officers' trust and follow the rules of society, particularly regarding the quality of contract enforcement, property rights, the police and courts, and the likelihood of crime and violence. Kirchler et al. (2008) emphasize in this context that, in addition to audits and the use of power, measures

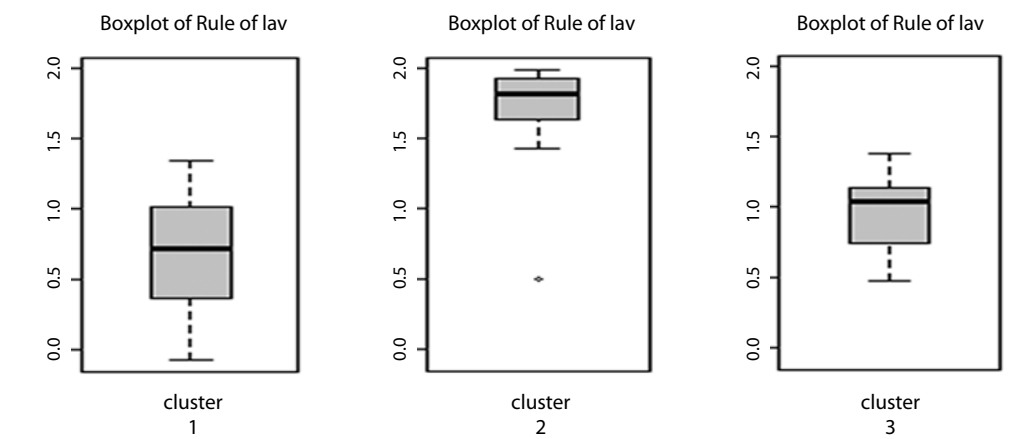
Figure 6 Boxplots of level of Tax regulations index in the created clusters



Source: Authors' work

to build trust are also necessary. These increase voluntary tax evasion compliance, which is more cost-effective than enforced compliance. Moreover, trust creates an environment in which authorities and citizens accept each other. This group of authors therefore holds that tax compliance can be achieved by increasing trust and the level of government laws. At the same time, we can connect this indicator with the previous one, where we can notice the similarity of the results, as well as the impact on the estimated scope of tax evasion. Both indicators are linked to laws, whether adoption, compliance, or implementation, which is often a problem for countries that we call "new" (they joined the EU later) and that show worse economic results compared to other countries. The large heterogeneity of index values in cluster 1, as can be seen in Figure 7, only confirms this.

Figure 7 Boxplots of level of Government laws index in the created clusters



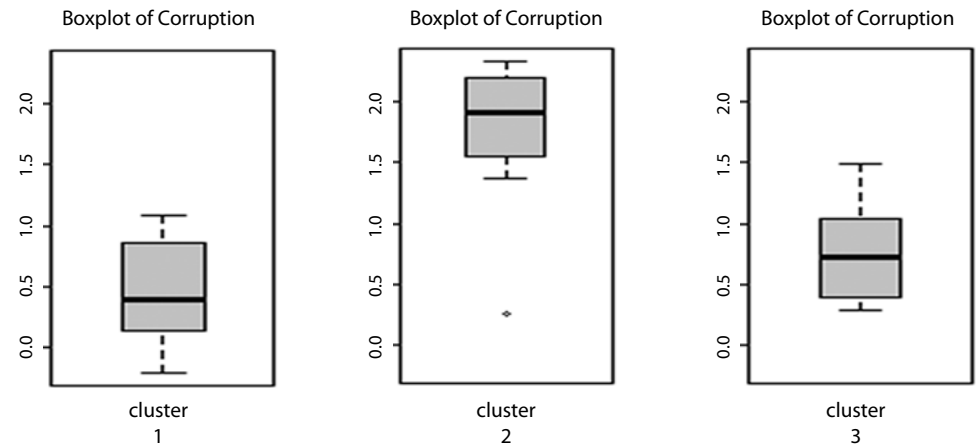
Source: Authors' work

In the next part, we focus also on corruption. Arsic et al. (2015) argue that a high level of corruption discourages taxpayers from paying taxes, because corruption gives the impression that these taxes will not be used to adequately finance the public sector but will rather lead to the private profit of certain people

or politicians. Likewise, Akinyom and Okpala (2013) draw attention to the fact that if citizens cannot be sure, due to the high level of corruption, whether the taxes they pay are used to finance public goods and services, they are unwilling to pay and are more likely to avoid tax obligations. Thus, they state that, although corruption and tax evasion are distinct and separate problems, they can easily be interconnected and reinforced. A society that has a higher level of corruption may allow more tax evasion as corrupt officials seek more income through bribes. On the other hand, a higher level of tax evasion can lead to corruption by offering more bribes (Alm, Martinez-Vazquez, McClellan, 2016).

A high level of corruption and little regulation, this is characteristic of cluster 1. This is confirmed by our analysis as well as by the authors who draw attention to this very fact. Of course, with decreasing corruption, we can assume that the estimated scope of tax evasion can also decrease. We can confirm this statement. In the cluster 3 we can see, in Figure 8, a decreasing tendency of both corruption and tax evasion, and cluster 2 shows high values of the corruption index, which means little corruption in the country, and on the other hand, we also know that countries in this cluster have a low value of the estimated extent of tax evasion.

Figure 8 Boxplots of level of Corruption index in the created clusters



Source: Authors' work

CONCLUSION

The main goal of our study was to find pattern in tax evasion time trajectories of European Union countries and thus to create homogenous clusters that include countries with similar tax evasion situation. Our next goal was to compare created cluster from the perspective of quality indicators of public institutions. For this purpose, we used K-means algorithm adjusted for panel data and based on Calinski and Harabatz criterium, we decided to use 3 clusters.

From our results can be seen that the time trajectories of tax evasion index show a decreasing trend for all countries over the observed period, with a temporary increase in 2009, likely due to the global economic crisis. In 2019, the estimated rate of tax evasion is lower than the EU average in older member states, while newer member states and countries located in the south of Europe tend to have higher estimated rates of tax evasion. Scandinavian countries (Denmark, Sweden, and Finland) show the lowest estimated range of tax evasion and perform significant improvement over the monitored period. Croatia, Romania, and Bulgaria, as the newest EU member states located in the south of Europe, have the highest estimated rates of tax evasion. The quality of institutions, including democracy, political stability, government efficiency, regulatory measures, government laws, and corruption, have a significant impact

on the estimated scope of tax evasion. Cluster 2, characterized by high values in indicators of democracy, political stability, government efficiency, and low corruption, exhibits the lowest estimated extent of tax evasion. Conversely, cluster 1 shows the highest tax evasion and lowest values in these indicators. Our study suggests a positive association between democratic participation, political stability, efficient government services, quality regulation, government laws, and lower levels of tax evasion. Corruption and lack of regulation are prominent factors associated with higher tax evasion rates, as observed in cluster 1. Overall, the findings highlight the importance of institutional quality and the role of effective governance in combating tax evasion. Countries with strong democratic systems, political stability, efficient governments, and low corruption tend to have lower levels of tax evasion.

ACKNOWLEDGMENT

This work has been supported by the project VEGA No. 1/0150/21.

References

AKINYOMI, O. J., OKPALA, K. E. (2013). Appraisal of Factors Influencing Tax Avoidance and Evasion in Nigeria. *International Journal of Research in Commerce and Management,* 4(5): 107–111.

ALM, J., MARTINEZ-VAZQUEZB, J., MCCLELLAN, CH. (2016). Corruption and firm tax evasion [online]. *Journal of Economic Behavior & Organization*, 124: 146–163. <https://doi.org/10.1016/j.jebo.2015.10.006>.

ARSIĆ, M., ARANDARENKO, M., RADULOVIĆ, B., RANĐELOVIĆ, S., JANKOVIĆ, I. (2015). Causes of the Shadow Economy [online]. In: KRSTIĆ, G., SCHNEIDER, F. (eds.) *Formalizing the Shadow Economy in Serbia*, Contributions to Economics, Springer, Cham. <https://doi.org/10.1007/978-3-319-13437-6_4>.

BÁNOCIOVÁ, A., VRAVEC, J. (2015). Analýza efektívnosti výberu DPH na Slovensku. *Finančné trhy*, 4.

BENYIK, M. (2019). *Growing socialine qualities* [online]. [cit. 18.3.2021]. <https://www.socialwatch.org/node/18290>.

BRUINSHOOFD, A. (2016). *Institutional quality and economic performance* [online]. [cit. 28.11.2020]. <https://economics.rabobank.com/publications/2016/january/institutional-quality-and-economic-performance/>.

ĎURINOVÁ, I. (2013). Daňové systémy krajín Škandinávie. *Biatec: odborný bankový časopis* 21(1): 21–25.

FELD, L. P., TYRAN, J. R. (2002). Tax evasion and voting: An experimental analysis [online]. *Kyklos*, 55(2): 197–221. <https://doi.org/10.1111/1467-6435.00183>.

GENOLINI, C., ALACOQUE, X., SENTENAC, M, ARNAUD, C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*, 65(4): 1–34.

HANOUSEK, J., PALDA, F. (2004). Quality of Government Services and the Civic Duty to Pay Taxes in the Czech and Slovak Republics, and other Transition Countries [online]. *Kyklos,* 57(2): 237–252. <https://doi.org/10.1111/j.0023-5962.2004.00252.x>.

HIWATARI, N. (2014). *The Quality of Governance and Tax Effort: Evidence from Developed and Developing Countries* [online]. [cit. 25.2.2021]. <http://www.pp.u-tokyo.ac.jp/graspp-old/courses/2014/documents/graspp2014-5140143-5a.pdf≥.

KATZ, B. G., OWEN, J. (2013). Exploring tax evasion in the context of political uncertainty [online]. *Economic Systems*, 37(2): 141–154. <https://doi.org/10.1016/j.ecosys.2012.12.005>.

KIABEL, B. D., NWOKAH, N. G. (2009). Curbing Tax Evasion and Avoidance in Personal Income Tax Administration: A Study of the South-South States of Nigeria. *European Journal of Economics, Finance and Administrative Sciences,* 15: 51–61.

KIRCHLER, E., HOELZL, E., WAHL, I. (2008). Enforced versus voluntary tax compliance: The "slippery slope" Framework [online]. *Journal of Economic Psychology,* 29(2): 210–225. <https://doi.org/10.1016/j.joep.2007.05.004>.

MACQUEEN, J., B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, 281–297.

MEDINA, L., SCHNEIDER, F. (2018). Shadow Economies Around the World: What Did We Learn Over the Last 20 Years? [online]. *IMF Working Paper No. 18/17*. <https://ssrn.com/abstract=3124402>.

OECD. (2019). *What is BEPS?* [online]. <https://www.oecd.org/tax/beps/about>.

ORVISKÁ, M. (2005). *Sociálno-ekonomická a inštitucionálna analýza tieňovej ekonomiky a daňových únikov.* Banská Bystrica: FF UMB.

SCHNEIDER, F., ENSTE, D. H. (2000). Shadow Economies: Size, Causes, and Consequences [online]. *Journal of Economic Literature,* 38(1): 77–114. <https://sci-hub.se/10.1257/jel.38.1.77>.

SCHNEIDER, F., BUEHN, A. (2016). Estimating the Size of the Shadow Economy: Methods, Problems and Open Questions [online]. *IZA Discussion Paper No. 9820*. <http://ftp.iza.org/dp9820.pdf>.

SCHNEIDER, F. (2022). New COVID-related results for estimating the shadow economy in the global economy in 2021 and 2022 [online]. *Int. Econ. Policy*, 19(2): 299–313. <https://doi.org/10.1007/s10368-022-00537-6>.

ŠIMONOVÁ, J. (2017). *Daňové úniky v Slovenskej republike.* Bratislava: Wolters Kluwer.
YAMEN, A., ALLAM, A., BANI-MUSTAFA, A., UYAR, A. (2018). Impact of institutional environment quality on tax evasion: A comparative investigation of old versus new EU members [online]. *Journal of International Accounting, Auditing and Taxation,* 32: 17–29. <https://doi.org/10.1016/j.intaccaudtax.2018.07.001>.

Diffusion of Mobile Technologies in the Area of Financial and Insurance Activities in Poland

Marcin Salamaga[1] | *Krakow University of Economics, Krakow, Poland*

Received 12.10.2023, **Accepted** (reviewed) 8.1.2024, **Published** 15.3.2024

Abstract

The diffusion of innovation in the area of banking and certain financial services is a topic eagerly taken up in empirical research, but in the case of Poland the study of this phenomenon in the area of finance and banking that utilizes the econometric models in question is practically non-existent. This article attempts to fill this research gap to some extent. The main objective of the article is to present a comparative analysis of the rate of mobile technologies diffusion among 3 groups of enterprises in the financial and insurance sectors: banking and lending, insurance and reinsurance, and brokerage and other activities. The occurrence of diffusion of mobile technologies and its rate was studied with the help of a logistic function. Results show that in companies operating in the insurance industry, the phase of rapid growth of mobile innovation diffusion lasted longer than in other groups of enterprises, but the highest dynamics of mobile innovation diffusion in the first phase of this phenomenon was achieved by companies conducting brokerage and other activities.

Keywords	***DOI***	***JEL code***
Innovation diffusion, logistic function, banking, insurance, Poland	*https://doi.org/10.54694/stat.2023.47*	*O30, C01, C22*

INTRODUCTION

In recent years the process of industrial production and the ways of providing services have undergone an enormous transformation, and the scale of these changes is so large that the accompanying ubiquitous progress has the hallmarks of the fourth industrial revolution (Yang and Gu, 2021). Its effects are visible in virtually all spheres of human activity. The essence of the fourth industrial revolution is the full integration of the physical production environment with the Internet, where communication between devices based on mobile technologies plays a fundamental role. Mobile technologies include the production of mobile devices and software dedicated to them, but also products and services that can be created

[1] Department of Statistics, Krakow University of Economics, Rakowicka 27, 31-510 Krakow, Poland. E-mail: salamaga@uek.krakow.pl, phone: (+48)122935231. ORCID: <https://orcid.org/0000-0003-0225-6651>.

and/or provided through these devices. The creation of innovations based on such technologies usually requires a large scientific and technical base, high-class specialists, significant financial resources invested in research and development, as well as providing appropriate conditions by the state to conduct such innovative activity. Most countries do not have an appropriate potential to create mobile innovations, so for countries such as Poland the main channel for the development of these technologies is their diffusion. The diffusion of mobile technologies can be understood as the process of market penetration by new products and services using remote access to the Internet network, GSM network, GPS geolocation technology or NFC proximity technology. The spread of this type of technology is stimulated by social factors (Peres et al., 2010), and whether an innovation is appreciated by the audience depends on their perceived use value of the solution, which is an outcome of the benefits and costs associated with the innovation. Mobile technology should be considered an unprecedented civilizational leap taking place at an express pace and affecting almost every area of human life. But it is also an important factor in giving businesses a competitive advantage. In themselves, these technologies stimulate innovation in enterprises and directly affect the development of many industries and economic sectors, consequently strengthening their economic potential.

The main advantages of mobile technologies include connectivity and immediacy that allow users to utilize them anywhere and at any time, personalization options allowing for the exchange of personal information in the target group of users, wide configuration options enabling system settings in a manner tailored to user needs, as well as data portability allowing for the possibility of storing and managing them (Łysik and Kutera, 2013).

The recipients of mobile technologies are both individual customers, enterprises and organizations. Improving mobile techniques within organizations and companies enables improvements in the production process and strengthens ties between the service provider – producer – consumer. This is visible in the sphere of financial and insurance services, where quick and direct contact with the customer during the provision of the appropriate service may be of key importance (e.g. in brokerage activities). In banking sector, internal mobile applications enable the company employees to easily access customer and account information, as well as supporting transaction processing. It is worth noting that mobile technologies allow for easy, safe and quick verification of customer accounts and data, as well as for carrying out various types of financial operations. The financial and insurance sectors are certainly one of those areas of economic activity where mobile technologies find particularly fertile ground and spread quite quickly. In the area of insurance, mobile technologies equip employees with the ability to quickly access information about customers and policies while away from the office, as well as to report claims of damage, handle them and process compensation claims. As a result, customers can be served faster and more effectively, which translates into increased customer satisfaction and loyalty. Thanks to such benefits, mobile technologies within banking and insurance organizations are becoming a business standard, since they increase work efficiency and flexibility, improve communication, facilitate customer service and increase the level of satisfaction with the use of financial and insurance services.

The diffusion of innovations in the area of banking, insurance and related financial services is the subject of numerous analyses and empirical studies. International literature offers a number of research results on the diffusion of mobile technologies using econometric models, however, the study of this phenomenon in the area of finance and banking with the use of these models is rare in Poland. The existing research gap in this area makes it difficult to reliably assess the diffusion of mobile technologies and the pace of this phenomenon. This article attempts to fill this gap to some extent. The main objective of undertaken analysis is to investigate the extent and rate of diffusion of mobile technologies in the area of financial and insurance services. The occurrence of diffusion of mobile technologies and its rate was studied with the help of a logistic function.

1 LITERATURE REVIEW

Technology diffusion is studied in various contexts, including a geographical perspective (Haynes et al., 1977; Brown, 1981), marketing (Mahajan et al., 1990), economics (Gurbaxani, 1990), and sociology (Rogers, 1995). A leading trend in mobile technology research is the diffusion analysis of mobile telephony innovations, since the mobile phone smartphone is the most recognizable example of the technology in question. Almost every consumer has a smartphone, and its wide functionality allows for using many functions and mobile applications. Hence, research on the diffusion of mobile innovations in this area is very popular. The analyses focus on the rate of innovation diffusion, as well as its determinants. Dekimpe et al. (1998), studying the diffusion of innovation between countries, showed that its determinants include the social system and the adoption ceiling. The research also draws attention to the deregulation of the telecommunications market and the possibilities of competing within it. Gruber and Verboven (2001a) showed, for example, that the regulatory effect upon the market in question, and competition within it, is less than the technological effect. In their research on mobile communications, they found that technological development and market deregulation, when creating the right conditions for competition through licensing, allows for accelerating the spread of mobile telephony in fifteen European countries. Similar conclusions were reached by Petrazzini (1996), Wallsten (2001), and Buys et al. (2009). Perkins and Neumayer (2005) found that the openness of trade significantly positively affected the speed of innovation diffusion, but did not prove the positive effect of FDI on the rate of diffusion of new technologies.

Gruber and Verboven (2001a) have concluded that a country's GDP is not a significant factor in diffusion of innovation, and countries that were the first to grant licenses to operate in the mobile telephony market tend to be slower to implement the *catch-up effect*. Research also identifies the telecommunications infrastructure as a factor in innovation diffusion. Gruber and Verboven (2001b) and Frank (2004) showed, for example, that the number of recipients of new mobile technologies was influenced by the signal coverage of telecommunications networks, and the level of development and availability of fixed-line telephony were not important factors in the diffusion of innovation. Liikanen et al. (2004) provided an interesting analysis of the impact of the first generation (1G) mobile networks on the diffusion of innovations related to the next generation network (2G) and vice versa. They concluded that the development of the first generation (1G) mobile network was an important factor in the diffusion of the second generation cellular network, but not the other way around. The level of economic development, the wealth of the country and cultural conditions as factors in the diffusion of mobile innovations, have been highlighted by authors such as Sundqvist et al. (2005), who conducted their research among 25 countries. Rouvinen (2006) showed that the diffusion rate of mobile innovation did not differ significantly between developed and developing countries. Chu et al. (2009) showed that market competition significantly influenced the diffusion of innovation in this market through downward pressure on the price of telecommunications services. They also proved that, in the case of Taiwan, the level of economic development, the number of mobile network operators did not affect the speed of diffusion of mobile innovations. Comer and Wikle (2008) examined mobile telephony diffusion between 1995 and 2005 and concluded that GDP per capita explained more than 75 percent of the variability of this phenomenon, Another important research strand analyses the spread of innovations by modeling their life cycles (Peres et al., 2010), with results in the form of mathematical models describing innovation growth curves with estimated parameters. The aim of this research is generally to provide accurate images of diffusion processes over time, so that managers can forecast sales, develop appropriate strategies, and act accordingly (Mahajan, Muller and Bass, 1990). When researching the process of diffusion of mobile innovations, the logistic function is most often used (Gruber and Verboven, 2001a, 2001b; Frank, 2004; Liikanen et al., 2004; Lee and Cho, 2007), and less often the Bass model (Dekimpe et al., 1998; Sundqvist, Frank and Puumalainen, 2005) or the Gompertz function (Rouvinen, 2006). It should be noted that in most cases, different countries are studied, characterized by different levels of technological development, GDP, and cultural customs,

so the research results are not always consistent. What in some countries may be an important determinant of the diffusion of mobile innovations, in another country may not have a significant impact on this process. The methods used to study this phenomenon are also an open question. A model that may work well in describing diffusion in one country does not necessarily correctly describe it in another country. For this reason, further study of this phenomenon is still relevant. In Polish literature, studies of the diffusion of innovation can be found in the works of Klincewicz (2011), Firlej and Żmija (2014), Wiśniewska (2004), Gwarda-Gruszczyńska (2017). Apart from a few exceptions (Kolarz, 2006), the use of econometric tools in this type of research in the case of Poland is rare, which makes it difficult to properly identify this process in general. An even greater challenge is the area of financial and insurance services, which, due to their specificity, show great ease in absorbing new mobile technologies.

2 METHODOLOGY

The authors of many studies have shown that the innovation diffusion rate varies according to an S-shaped curve (Sharif and Kabir, 1976; Desiraju et al., 2004; Cyclist, 2006). The process of spreading innovation is generally predictable: it is usually characterized first by a slow growth rate, then in the middle phase this rate increases rapidly, then finally decreases, the S-shaped curve flattens, and the level of innovation stabilizes (the growth dynamics of innovation diffusion fades). Therefore, in research on the dynamics of innovation diffusion, econometric models that reflect this kind of course of the phenomenon are most often used. Popular tools in modeling this type of phenomena include the logistic function, Gompertz model and Bass model. This article centers on the logistics model, whose effectiveness in the study of mobile technologies has been confirmed in other studies (Gruber and Verboven, 2001a, 2001b; Frank, 2004; Cramer, 2004; Liikanen et al., 2004; Lee and Cho, 2007). The logistic curve belongs to a more general family of S-shaped curves, which are represented by the Richards function (Richards, 1959; Lei and Zhang, 2004):

$$y(t) = \delta + \frac{\alpha - \delta}{\left(\theta + \beta exp(-\gamma t)\right)^{1/\nu}}, \tag{1}$$

where: t – time variable, α, β, γ, δ, θ, ν – parameters of the Richards function.

In Formula (1), the parameter γ expresses the growth rate, ν is the growth parameter with the maximum asymptote, and the parameters α and δ represent the upper and lower asymptote of the Richards function, respectively.

If we assume that $\delta = 0$, $\theta = \nu = 1$, then we get the logistic function (2):

$$y(t) = \frac{\alpha}{1 + \beta exp(-\gamma t)}. \tag{2}$$

In order to improve flexibility of the model that will be adapted to the data with limited supply, for the purposes of this study, the model (2) has been supplemented with the free term φ, which allows the modified logistic function to be written as follows:

$$y(t) = \frac{\alpha}{1 + \beta exp(-\gamma t)} + \varphi. \tag{3}$$

In modeling the diffusion of mobile technologies, it is important to determine the rate of change of this process. The growth rate of the logistic function can be calculated from the formula:

$$GR = \frac{dy}{dt}\frac{1}{y}. \tag{4}$$

According to the course of the logistic function, it is possible to distinguish the area in which it has an increasing growth rate, and then the area with decreasing growth dynamics, whereby at infinity the phenomenon tends to the saturation level expressed by the asymptote $y = (\alpha + \varphi)$.

The point separating the area of rapid growth rate from the area with decreasing growth rate is the *inflection point* with coordinates $\left(\frac{ln\beta}{\gamma}, \frac{\alpha}{2} + \varphi\right)$. The logistics function and its properties is used in the empirical part of the article to analyze the process of mobile innovations diffusion in various areas of financial and insurance activity in Poland.

3 EMPIRICAL RESEARCH RESULTS AND DISCUSSION

The study of innovation diffusion was based on statistical data on the number of companies belonging to the following financial industry sectors (according to the Polish Classification of Activities; PKD, 2007): banking and lending, insurance and reinsurance, and brokerage and other activities, which provide their employees with access to mobile devices enabling mobile access to the Internet, such as portable computers and smartphones.[2] Access to this type of equipment among employees in these industries is important because the modern financial sector is increasingly based on the latest technologies and digital solutions. Access to mobile devices, and, through them, to the Internet, is vital for the effective functioning of these industries, since many financial sector employees work remotely or need access to information and systems outside the office, and mobile devices enable them to work in the field, during meetings with clients or on business trips. Use of mobile equipment in the researched industries is necessary for:

- Communication with customers – financial sector employees need to keep in touch with customers in order to obtain additional information, explanations or to sign contracts. Mobile devices enable quick communication via e–mails, text messages or phone calls.
- Viewing customer data – employees need to access customer data, transaction history, or credit information in a variety of places and situations. Mobile devices enable real–time access.
- Analysis, calculations and decisions (credit, insurance premium, etc.) – when evaluating loan applications or determining insurance premiums, employees often have to analyze a variety of data, information and parameters. Mobile devices enable viewing this data in any location, which can speed up the decision–making process.
- Signing electronic documents – increasingly, documents can be signed electronically. Mobile devices allow employees and their customers to sign the necessary documents online.
- Accessing documents – financial sector employees must have access to documents such as policies, contracts and documents related to claims, loan agreements. Mobile devices enable storage and viewing of these documents on–line
- Market and trend monitoring – many financial firms follow markets and trends to make appropriate investment decisions. Mobile devices enable access to current market information and trend analysis.
- Banking and client applications – banks and credit companies often offer their customers mobile applications for account management, loan repayment, etc. Employees must be familiar with these applications to help customers use them.

Data for 2011–2020, obtained from the CSO Local Data Bank, were used to model the phenomenon in question jointly, and separately in each of the highlighted areas. Data includes time series of the number of companies in Poland in mentioned areas. Thanks to the characteristics of the logistics function, it is possible to compare the areas of banking, insurance and insurance and brokerage activities in terms of the rate of diffusion of mobile technologies, their ability to absorb innovations and develop patterns

[2] *<https://bdl.stat.gov.pl/bdl/start>.*

of diffusion of innovation in each of these areas. In the first periods, the number of enterprises in the financial and insurance sectors decreased (this may be due to changes in the way companies are classified), and in subsequent periods it systematically increased. First, using the Gauss-Newton algorithm, the parameters of the logistics function (2) were estimated separately for all enterprises from the financial and insurance sector. The estimation results of parameters are included in Table 1 and the course of the corresponding logistic curve for this type of enterprises is presented in Figure 1.

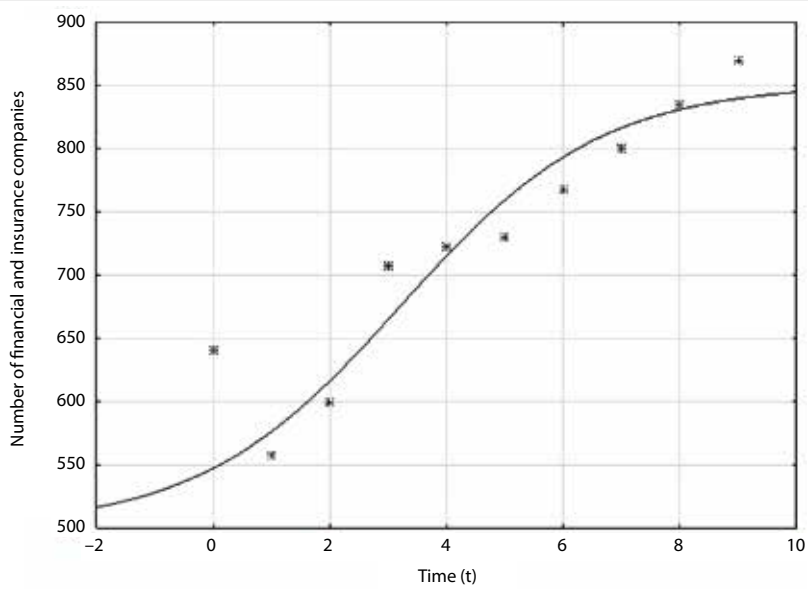
Table 1 Parameters of the logistic regression function among all enterprises operating in all analyzed sectors

Parameter	α	β	γ	φ
Coefficient	351.08476	6.45637	0.47121	501.21220
Standard error	54.61674	2.30892	0.14428	133.25470
t-Stat	6.42815	2.79628	3.26589	3.76131
p	0.00036	0.02667	0.01375	0.00706

Source: Own calculations

As can be seen in Figure 1, it was possible to match the logistic function to empirical data, which is confirmed by the process of diffusion of mobile technology innovations in this group of Polish enterprises. The parameters of the model are statistically significant, and the level of fit measured by the coefficient of determination is satisfactory. The inflection point of this logistic function is 3.96, which means that the phase of rapid diffusion growth in enterprises operating in the financial and insurance sectors lasts approximately 48 months. Also, the growth rate (GR) of the logistic function calculated at this point is 6.11%. The theoretical number of companies in the financial and insurance industry supplying employees with mobile devices with the Internet access will stabilize at around 853 in the long term.

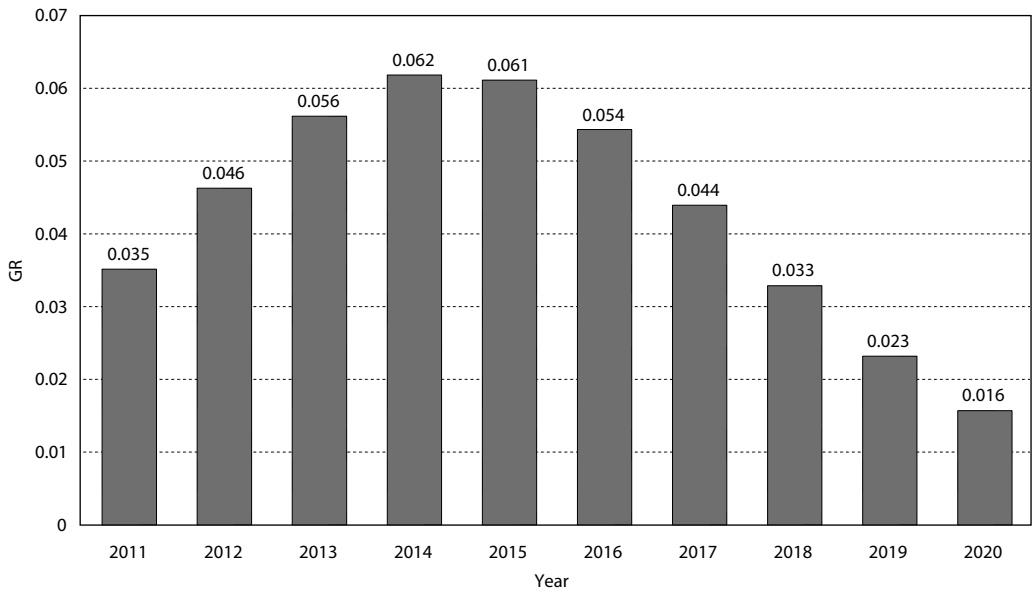
Figure 1 The course of logistic curve for all enterprises operating in all analyzed sectors



Source: Own elaboration

The growth rate in the logistics function in the examined group of companies was also calculated in subsequent periods. The results are illustrated in Figure 2. It can be seen here that the diffusion rate systematically increases in the first 4 periods from 3.5% to 6.2%, and then systematically decreases from 6.1% to 1.6%.

Figure 2 Growth rate of the logistics function for all enterprises operating in the financial and insurance sectors



Source: Own elaboration

Among Polish companies active in the researched sectors, a similar survey was carried out separately for companies engaged in banking and lending activities, insurance and reinsurance, and brokerage and other activities (following the Polish Classification of Activities; PKD, 2007). In each of these three cases, logistic functions were matched with statistically significant parameters (except parameter β in the model for brokerage and other activities sector). Therefore, it can be argued that the diffusion of mobile technologies occurs in each group of companies considered. The inflection points of the logistics curves in each of these 3 groups of companies (banking, insurance and brokerage activities) were respectively: 5.83, 6.94 and 3.94. This means that the insurance sector had the longest phase of rapid growth in the diffusion of innovation, where it lasted about 84 months. On the other hand, the shortest period of dynamic growth of 47 months had the diffusion of mobile innovations in companies conducting brokerage activities.

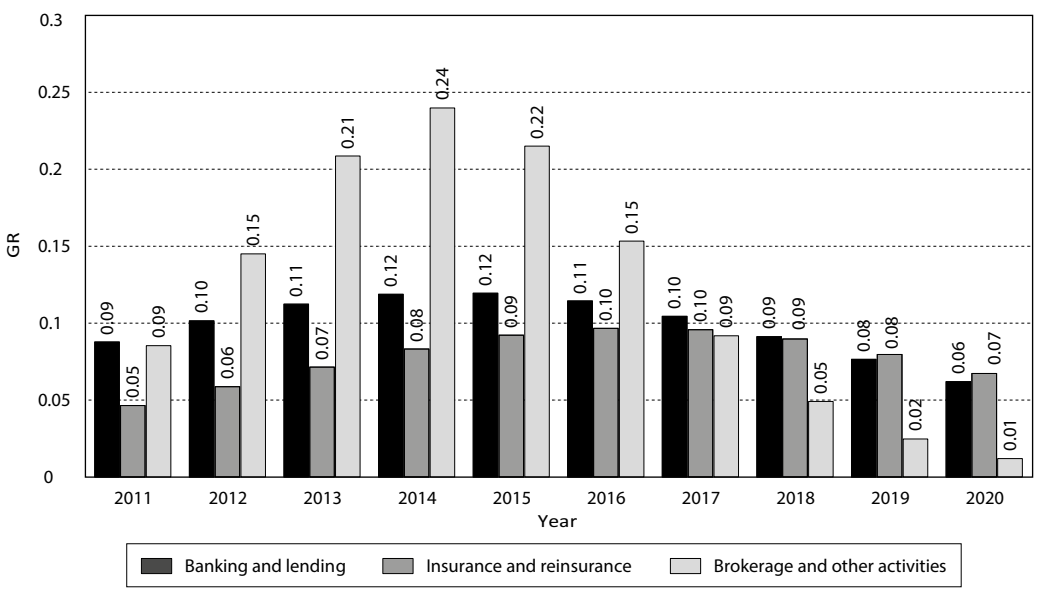
At the same time, in this group of companies, the growth rate at the inflection point was the highest among all groups of companies (the growth rate of the logistics function amounted to approx. 21.77%). On the other hand, the lowest growth dynamics were characterized by companies conducting insurance activities, as evidenced by the growth rate (GR) of the logistics curve of 9.01%. A slightly higher growth rate of the diffusion of mobile innovations was recorded by enterprises conducting banking or lending activities, where the growth rate was 10.64%. The growth rate in the logistics function in each of the 3 groups of companies, was also calculated in 10 consecutive periods. The results are illustrated in Figure 3, which shows that the rate of diffusion systematically increases in all groups of enterprises, and then decreases after passing through the inflection point, which is different in each group of enterprises.

Table 2 Parameters of the logistic regression function for enterprises operating separately in all analyzed sectors

Parameter	α	β	γ	φ
Banking and lending				
Coefficient	360.87530	7.12562	0.33668	105.1247
Standard error	145.91257	2.49353	0.13534	39.3251
t-Stat	2.47323	2.85764	2.48764	2.67322
p	0.04263	0.02442	0.04174	0.03185
Insurance and reinsurance				
Coefficient	45.36380	11.54123	0.35227	21.6361
Standard error	16.83706	3.28445	0.09972	8.8723
t-Stat	2.69428	3.51390	3.53239	2.43861
p	0.03089	0.00981	0.00957	0.04485
Brokerage and other activities				
Coefficient	82.21831	19.33760	0.75109	29.78168
Standard error	7.70495	13.51639	0.20822	7.30219
t-Stat	10.67084	1.43068	3.60717	4.07846
p	0.00001	0.19561	0.00866	0.00470

Source: Own calculations

Figure 3 Growth rate of logistics function of enterprises operating separately in all analyzed sectors



Source: Own elaboration

In the period 2011–2015, the highest growth dynamics of the logistics curve was achieved by companies conducting brokerage and other activities. On the other hand, the dynamics of diffusion of mobile innovations after 2015 in this group of companies clearly weakened and was lower than in other groups of companies. The lowest dynamics of diffusion in the first 6 years was shown by companies conducting insurance activities, while in the last two years for this group of companies the dynamics of diffusion growth was the highest among all enterprises. It may be surprising that, in terms of the analyzed rate of innovation diffusion, companies engaged in banking and lending activities are weaker compared to other groups of enterprises. Banks, in particular, generally have considerable investment capital to implement specific innovation strategies. However, it should be remembered that in this article, the measurement of innovation is limited to portable devices enabling mobile Internet access among employees. Simultaneously, the innovativeness of the banking sphere concerns various processes and product innovations, focused not on equipment but rather on a specific type of service provided by banks, in the form of development work initiated by customer needs, implementation of information processing and security technologies or development of marketing and sales support tools. Depending on the adopted innovation strategy, the aim of a specific innovative activity may include the intent to achieve and/or maintain a leading position on the market, gain new customers, acquire new solutions that are no longer protected by patents. Obtained results suggest that mobile devices may be more important for employees in the insurance industry. Some of them work in the field (e.g. claims liquidation) and such devices are necessary to carry out their tasks, which is reflected in the longest period of rapid increase in the diffusion of innovations in the surveyed groups of enterprises. Equally important is the use of such devices in the work of stack market traders. Modern stock exchange transactions take place mainly electronically, and fast and uninterrupted access to data is crucial for making the correct investment decisions. Hence, the relatively deep diffusion of innovation in brokerage companies can be explained by the nature of the work performed by brokers. Often their work requires making decisions in a fairly short time. Mobile devices such as smartphones and tablets make it easy for them to monitor the market in real time, view up-to-date financial information, analyze charts and indicators, and execute trades anywhere they have access to the Internet.

CONCLUSIONS

Based on the results obtained, it can be concluded that the diffusion of mobile technologies occurs in all analyzed groups of companies in the financial and insurance industries, although the duration and intensity of this phenomenon varies and depends on the specific group of enterprises. Among companies operating in the insurance industry, the phase of rapid growth of mobile innovation diffusion lasts longer than in other groups of enterprises, but the highest dynamics of mobile innovation diffusion in the first phase of this phenomenon was achieved by companies conducting brokerage and other activities. The shortest phase of rapid growth of innovation diffusion was in brokerage companies, but the dynamics of this phenomenon in the period under review was the highest. The varied length of the phase of increase in the diffusion of mobile innovations may result from the specificity of individual industries, and its lower dynamics in the group of companies conducting insurance activity (in the first phase of diffusion) may indicate the presence of certain barriers, factors weakening the phenomenon of diffusion. Their detailed diagnosis requires continued research (e.g. by using a purpose-designed questionnaire). Previous studies of diffusion of innovations in enterprises may indicate that factors differentiating the period and pace of diffusion of innovation include the level of technological advancement of the industry, the level of internationalization, and the ownership structure of enterprises. The analyzed groups of companies belong to the same financial and insurance industry, so the differences in the level of technological advancement may be insignificant, but the impact of this factor on the length and depth of diffusion, as well as other factors, cannot be excluded without additional research. It should also be taken into

account that the research was limited to the analysis of diffusion of product innovations, namely portable equipment enabling access to the Internet, the use of which in different groups of employees may vary, depending on the specifics of work in a given area. Analysis of the length and depth of process innovations among similar groups of companies could provide different results.

References

BUYS P., DASGUPTA S., THOMAS T. S., WHEELER D. (2009). Determinants of a Digital Divide in Sub-Saharan Africa: A Spatial Econometric Analysis of Cell Phone Coverage. *World Development, Elsevier*, 37(9): 1494–1505.

BROWN, L. A. (1981). *Innovation diffusion: A new perspective*. New York: Methuen.

CHU, W., WU F., KAO, K., YEN, D. C. (2009). Diffusion of mobile telephony: An empiricalstudy in Taiwan [online]. *Telecommunications Policy*, 33(9): 506–520, <https://doi.org/10.1016/j.telpol.2009.07.003>.

COMER, J. C., WIKLE, T. A. (2008). Worldwide diffusion of the cellular telephone, 1995–2005 [online]. *The Professional Geographer*, 60(2): 252–269, <https://doi.org/10.1080/00330120701836303>.

COURI, C. L., ARBACHE, J. (2006). *Are Fixed and Cell Phones Substitutes or Compliments? The Case of Brazil* [online]. <http://dx.doi.org/10.2139/ssrn.905125>.

CRAMER, J. S. (2004). The early origins of the logit model [online]. *Studies in History Philosophy of Biological & Biomedical Science*, 35(4): 613–626, <https://doi.org/10.1016/j.shpsc.2004.09.003>.

DEKIMPE, M. G., PARKER, P. M., SARVARY, M. (1998). Staged Estimation of International Diffusion Models: An Application to Global Cellular Telephone Adoption [online]. *Technological Forecasting and Social Change*, 57(1–2): 105–132. <https://doi.org/10.1016/S0040-1625(97)00085-1>.

DESIRAJU, R., NAIR, H., CHINTAGUNTA, P. (2004). Diffusion of new pharmaceutical drugs in developing and developed Nations [online]. *International Journal of Research in Marketing*, 21(4): 341–357, <http://dx.doi.org/10.2139/ssrn.945419>.

FIRLEJ, K., ŻMIJA, D. (2014). *Transfer wiedzy i dyfuzja innowacji jak źrodło konkurencyjności przedsiębiorstw przemysłu spożywczego w Polsce* [*Knowledge transfer and diffusion of innovation as a source of competitiveness of food industry comapnies in Poland*]. Kraków: Fundacja Uniwersytetu Ekonomicznego w Krakowie.

FRANK, L. D. (2004). An analysis of the effect of the economic situation on modeling and forecasting the diffusion of wireless communications in Finland [online]. *Technological Forecasting and Social Change*, 71(4): 391–403, <https://doi.org/10.1016/S0040-1625(02)00392-X>.

GRUBER, H., VERBOVEN, F. (2001a). The diffusion of mobile telecommunications services in the European Union [online]. *European Economic Review*, 45(3): 577–588, <https://doi.org/10.1016/S0014-2921(00)00068-4>.

GRUBER, H., VERBOVEN, F. (2001b). The evolution of markets under entry and standards regulation – the case of global mobile telecommunications [online]. *International Journal of Industrial Organization*, 19(7): 1189–1212, <https://doi.org/10.1016/S0167-7187(01)00069-8>.

GURBAXANI, V. (1990). Diffusion in computing networks: The case of BITNET [online]. *Communications of ACM*, 33(12): 65–75, <https://doi.org/10.1145/96267.96283>.

GWARDA-GRUSZCZYŃSKA, E. (2017). Dyfuzja innowacji – następstwo komercjalizacji nowych technologii [Diffusion of innovation – the consequence of commercialization of new technologies]. *Organizacja i Kierowanie*, 2(176): 383–396.

HAYNES, K. E., MAHAJAN, V., WHITE, G. M. (1977). Innovation diffusion: A deterministic model of space-time integration with physical analog [online]. *Socio-Economic Planning Sciences*, 11(1): 25–29, <https://doi.org/10.1016/0038-0121(77)90043-X>.

KLINCEWICZ, K. (2011), *Dyfuzja innowacji. Jak odnieść sukces w komercjalizacji nowych produktów i usług* [*Diffusion of innovation. How to be successful in commercializing new products and services*]. Warszawa: Wydział Zarządzania UW.

KOLARZ, M. (2006). *Wpływ zagranicznych inwestycji bezpośrednich na innowacyjność przedsiębiorstw w Polsce* [*The influence of foreign direct investments on the innovativeness of companies in Poland*]. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

LEE, M-K, CHO, Y. (2007). The diffusion of mobile telecommunications services in Korea. *Applied Economics Letters*, 14(7): 477–481.

LEI, Y. C., ZHANG, S. Y. (2004). Features and Partial Derivatives of Bertalanffy – Richards Growth Model in Forestry. *Nonlinear Analysis: Modelling and Control*, 9(1): 65–73.

LIIKANEN, J., STONEMAN, P., TOIVANEN, O. (2004). Intergenerational effects in the diffusion of new technology: the case of mobile phones [online]. *International Journal of Industrial Organization*, *Elsevier,* 22(8–9): 1137–1154. <https://doi.org/10.1016/j.ijindorg.2004.05.006>.

ŁYSIK, Ł., KUTERA, R. (2013). Technologie mobilne jako determinanta rozwoju innowacyjnego społeczeństwa informacyjnego [Mobile technologies as a determinant of development of innovative information socjety]. *Zeszyty Naukowe Uniwersytetu Szczecińskiego nr. 763, Ekonomiczne Problemy Usług nr. 105*.

MAHAJAN, V., MULLER, E., BASS, F. M. (1990). New product diffusion models in marketing: A review and directions for research [online]. *Journal of Marketing*, 54(1): 1–26, <https://doi.org/10.2307/1252170>.

PERES, R., MULLER, E., MAHAJAN, V. (2010). Innovation diffusion and new product growth models: A critical review and research directions [online]. *International Journal of Research in Marketing*, 27(2): 91–106, <https://doi.org/10.1016/j.ijresmar.2009.12.012>.

PERKINS, R., NEUMAYER, E. (2005). International technological diffusion, latecomer advantage and economic globalization: A multi-technology analysis [online]. *Annals of the Association of American Geographers*, 95(4), 789–808, <https://www.jstor.org/stable/3694013>.

PETRAZZINI, B. (1996). Competition in telecoms – Implications for universal service and employment. *Public Policy for the Private Sector*, Note 96, Washington, DC: World Bank.

RICHARDS, F. J. (1959). A Flexible Growth Function for Empirical Use [online]. *Journal of Experimental Botany*, 10(2): 290–300, <http://dx.doi.org/10.1093/jxb/10.2.290>.

ROGERS, E. M. (1995). *Diffusion of innovations*. 4th Ed. New York: Free Press.

ROUVINEN, P. (2006). Diffusion of digital mobile telephony: Are developing countries different? [online]. *Telecommunications Policy*, 30(1): 46–63, <https://doi.org/10.1016/j.telpol.2005.06.014>.

SHARIF, M. N., KABIR, C. (1976). System dynamics modelling for forecasting multilevel technological substitution. *Technological Forecasting and Social Change*, 9(1–2): 89–112.

SUNDQVIST, S., FRANK, L., PUUMALAINEN, K. (2005). The effects of country characteristics, cultural similarity and adoption timing on the diffusion of wireless communications [online]. *Journal of Business Research*, 58(1): 107–110, <https://doi.org/10.1016/S0148-2963(02)00480-0>.

WALLSTEN, S. (2001). An empirical analysis of competition, privatization, and regulation in African and Latin America. *Journal of Industrial Economics*, 49(1): 1–19.

WIŚNIEWSKA, J. (2004). *Ekonomiczne determinanty dyfuzji innowacji produktowych i technologicznych w banku komercyjnym* [*Economic determinants of diffusion of product and technological innovations in a commercial bank*]. Szczecin: Wydawnictwo Naukowe Uniwersytetu Szczecińskiego.

Assessing the Importance of ICT on High-Educated Poverty in East Java Using Random Forest Methods

Dias Satria[1] | *University of Brawijaya, Malang, Indonesia*
Radityo Putro Handrito | *University of Brawijaya, Malang, Indonesia*
Girindra Mega Paksi | *University of Brawijaya, Malang, Indonesia*
Patricia Juan Aurellia | *University of Brawijaya, Malang, Indonesia*

Received 3.8.2023 (revision received 30.10.2023), **Accepted** (reviewed) 28.11.2023, **Published** 15.3.2024

Abstract

Poverty is one of the main issues in the economic development of a country, including Indonesia. This was exacerbated when the Covid-19 pandemic had an impact on economic activities, so there are educated people who are still below the poverty line. Therefore, strategies in tackling the problem of poverty need to be developed in line with the times that have entered the era of digitalization. The research aimed to examine the importance of Information and Communication Technologies (ICTs) on high-educated poverty in East Java. Data obtained from the National Survey in Indonesia is analyzed with the random forest algorithm as a classification method in machine learning. The analysis shows that social media, laptop ownership, age, higher education, and internet access are the five most important variables for high-educated poverty classification in East Java. Based on this, a number of recommendations can be made to policy makers regarding the effect of digitalization on high-educated poverty in East Java.

Keywords	***DOI***	***JEL code***
Education, digitalization, machine learning, poverty, random forest	*https://doi.org/10.54694/stat.2023.38*	*I24, I32, O12*

INTRODUCTION

The Covid-19 pandemic has exacerbated global poverty levels. Lockdowns, business closures and travel restrictions have affected the informal sector and the small and medium enterprise sector, and millions of workers in these sectors have lost their jobs. To mitigate the pandemic‘s impact on global poverty,

[1] University of Brawijaya, Jl. Veteran No 10–11, Ketawanggede, Kec. Lowokwaru, Kota Malang, Jawa Timur 65145, Indonesia. Corresponding author: e-mail: dias.satria@ub.ac.id, phone: (+62)81333828319. ORCID: <https://orcid.org/0000-0002-4068-6807>.

many countries and international organizations have increased aid and financial support, such as direct cash transfers, food subsidies, and other social assistance programs. However, these efforts are far from enough, and more support is needed to strengthen social security systems in the future and ensure that people most vulnerable to poverty are protected.

Poverty is one of the main issues affecting human development (Nili, 2019). Indonesia is one of the countries with 9.78% poverty or around 27 million people living below the poverty line (Finch and Finch, 2020). Poverty in Indonesia is high when compared to countries in Southeast Asia, such as Malaysia and Thailand, which are only 0.4% and 6.1%. The poverty rate in Indonesia is still high when compared to international standards. According to World Bank standards, a person is categorized as living in poverty if income is less than US$1.90 per day (Diop and Asongu, 2021). In this case, Indonesia still has a lot of homework to do to overcome the problem of poverty. Indonesia has 38 provinces with various characteristics. One of the provinces in Indonesia that has a considerable economic potential is East Java. As the second largest province in Indonesia after Papua, East Java has quite diverse economic sectors, such as industry, agriculture, fisheries, and tourism. Despite its huge economic potential, East Java is also one of the provinces with a significant poverty rate in Indonesia. Therefore, a study of poverty in East Java can provide an overview of the factors that cause high poverty rates.

Poverty rates must be reduced for several important reasons, including the human right to have an equal access to resources and economic opportunities. A high poverty rate indicates that a large number of individuals do not have sufficient access to the resources necessary to satisfy basic needs such as food, clothing and housing. In addition, poverty can reduce people's quality of life and social well-being. High poverty rates can also hinder economic progress, as poor individuals have fewer resources and less capacity to actively participate in the economy (Abduvaliev and Bustillo, 2019). Similarly, in terms of peace, a high level of poverty can lead to social tensions and exacerbate social conflicts, thus threatening the stability and peace of society.

In recent years, the impact of technology on poverty has gained an increased attention and interest. Technology has the potential to improve economic opportunities and access to essential services, yet there are also concerns about the unequal distribution of technology and its negative effects (Chernova et al., 2019). The relationship between technology and poverty is complex and multi-faceted, and the full impact of technology on poverty is still not well understood. This highlights the need for more comprehensive and in-depth research to better understand the complex relationship between technology and poverty, and to develop effective solutions to address poverty and promote economic growth.

Based on Central Bureau of Statistics publication data for 2020, there are 19.5% of the poor population in East Java who have completed 12 years of compulsory education. In other words, there are still people with high school, vocational, and even university education who are still below the poverty line, called high-educated poverty. Over the past few years, the proportion of poor people with secondary or higher education in East Java has also shown an increasing number. There are several potential impacts if there are people with high education but classified as poor, including: poor individuals may struggle to secure jobs that match their qualifications, or even access to business opportunities or capital. This can hinder their ability to improve their standard of living and fully leverage their skills. Poor individuals may also lack equal access to social resources such as connections, information, or family and friend support. This can impact their mental health and overall well-being. Therefore, currently many factors that can affect poverty are increasingly complex. Not only about education, but also matters related to skills and the times, namely digitalization.

The purpose of this study is to address the research gap by exploring the relationship of technology on high-educated poverty in East Java. The research question for this study is as follows: Does technology, the use of social media, e-commerce, and digital banking, have an impact on high-educated poverty in East Java? To address this research question, a literature review was conducted to examine existing

research on the relationship between technology and poverty and we focused on high-educated poverty in East Java. The literature review revealed that there have been studies examining the potential benefits of technology in alleviating poverty, such as improving access to education and financial services. However, the full impact of technology on poverty is still not well understood, and more comprehensive and in-depth research is needed to better understand the complex relationship between technology and poverty.

To analyze the relationship between technology and poverty, we will be using the machine learning method called Random Forest (RF). RF is a popular machine learning algorithm that can handle large datasets and complex relationships between variables. This method will allow us to examine the relationship between various factors, including technology use and access, socio-economic status, and poverty. RF is a robust and efficient method that can handle both continuous and categorical variables, and it can also handle missing data and outliers (Ao et al., 2019). The results generated by RF will provide insights into the factors that contribute to poverty and the importance of technology on the model of high-educational poverty in East Java. With the use of Susenas data and the RF method, this study aims to fill the research gap in understanding the complex relationship between technology and poverty and to inform policy decisions to address poverty and promote economic growth in East Java.

The policy implications of this study are important, as the findings can inform policy-making decisions related to technology and poverty. For instance, the results of this study can provide valuable insights into the need for government programs and initiatives to address the unequal distribution of technology and ensure that its benefits are distributed equitably. Additionally, the results of this study can inform the development of policies and initiatives aimed at improving digital literacy and technology access for those in poverty, to ensure that they are able to fully benefit from the opportunities that technology can provide.

Based on the backgrounds above, this study was conducted to exploring the importance of technology on high-educated poverty in East Java using machine learning method such as RF. This study is organized into five sections. Introduction explains the Covid-19 impact on Indonesia, especially high-educated poverty in East Java, and how the digitalization related to skills. Section 1 provides theoretical support and past studies related to high-educated poverty and digitalization in Indonesia and worldwide. Section 2 provides information about research analysis and tools to know the importance of digitalization on high-educated poverty in East Java using RF. Section 3 discusses the results of the analysis and explains the types of digitalization that most importance to the model of high-educated poverty in East Java. The last section presents the concluding remarks.

1 LITERATURE REVIEW

In terms of poverty, Hofmarcher (2021) found an important role of the education sector in poverty alleviation programs in Europe. Meo et al. (2020) gave a fresh idea about the non-linear relationship between unemployment, governance, and poverty in Pakistan. Long-term unemployment will create too much immoral crime in countries such as frustration, homelessness, family tension, loss of confidence, social isolation, self-esteem, and poverty (Siddiqa, 2021). In Indonesia, many studies have investigated the determinants and nature of poverty as it relates to poverty reduction, unemployment, and national economic growth. For example, Erlando et al. (2020) found a relationship between economic growth, inequality and poverty. Investment has a direct effect on poverty, while economic growth has no direct effect on poverty. Poverty in Indonesia is also caused by many other factors, including population growth, investment, education, health, market structure, and government regulation (Jacobus et al., 2019).

In this era, we know about digitalization. Digitalization is the process of transforming traditional processes or activities into digital form through the use of digital technologies such as computers, the internet, and other electronic devices. In digitalization, data that was previously stored in physical

form, such as paper or books, is recorded and stored in digital form. Digitalization enables faster, more efficient, and easily accessible access to information. Digitalization has impacted many aspects of human life such as communication, education, health, entertainment, and business. Digital technology can enable new business models in the form of digital services, digital platforms, digital tools or infrastructure (McQuire, 2021), digital artifacts (Sebastiani, 2021), or service innovations enabled through the Internet (Vakulenko et al., 2019).

E-commerce is a form of digitalization because it enables the digitization of business transactions, which previously could only be conducted physically. The rapid advancement of information and communication technologies, combined with globalization, has resulted in a significant impact on economic life through electronic commerce. E-commerce has become accessible and convenient, with applications and progress seen in all economic sectors. Its availability on the internet and instantaneous nature are the main reasons behind these developments. E-commerce's features have made it widely used in economic life, with its usage becoming increasingly diverse as more people use desktop computers and mobile phones (Guven, 2020). With mobile applications, transactions can be made at any time and place. Additionally, e-commerce also allows for the digitization of promotion, where businesses can promote their products or services through digital platforms such as social media.

Social media is an online platform that enables individuals or groups to interact, communicate, and share information over the internet (Lovari and Valentini, 2020). Before the existence of social media, social interactions were more physical and limited by geographical distance. However, with the advent of social media, social interaction can occur online, where people can interact with others who are located all around the world without being hindered by geographical distance. Social media also enable the digitalization of content, where information and content such as images, videos, or text can be uploaded and accessed online. Thus, social media allow for the digitalization of social interaction and content, which is one aspect of digitalization in general.

Digital banking refers to banking services that are provided through digital channels such as mobile applications, internet banking, and other online platforms. Digital banking has become increasingly popular in recent years due to advancements in technology and the growing trend of digitalization. Digital banking is a type of digitalization that has a strong connection to business. Banks around the globe are shifting towards digital banking services to offer their customers convenience and speed. These services involve opening accounts, transferring funds, paying bills, and more (Wewege et al., 2020). One of the main advantages of digital banking is its ease and convenience. Customers can access banking services at any time and from any location using electronic devices such as smartphones, tablets, or laptops. Furthermore, digital banking provides lower transaction costs than traditional banking services. By using digital banking, businesses can accelerate transaction processes, save time and cut costs. As a result, digital banking is increasingly becoming the go-to solution for businesses seeking to boost efficiency and productivity in their operations.

The literature for this study provides a thorough examination of the research done on the connection between ICTs access to finance, and poverty. It also offers an overview of the research on how new technologies can promote digital financial services. The growing body of literature suggests that ICTs play a significant role in promoting economic and social improvement. Research also indicates that improved access to finance can help reduce poverty and enhance household welfare. According to Kendall et al. (2010), developed countries have greater access to finance compared to developing countries. ICTs can provide access to financial services, information, and e-banking to the poor, especially in remote areas. Isukul and Tantua (2021), and Mishra and Singh Bisht (2013) found that mobile technology was an effective tool for financial inclusion in remote regions. Chatterjee (2020), and Pradhan et al. (2021) confirmed the positive impact of ICTs on economic growth through financial inclusion. Bhavnani et al. (2008) demonstrated the increased social and economic benefits of mobile phones in rural areas

and predicted the poverty alleviation effects of mobile phones. Chib et al. (2015) reviewed the ICTs literature and progress in measuring its impact and concluded that prior to 2002, researchers focused more on macro-level ICTs linkages, while there was less attention given to its role in poverty alleviation at the micro level.

Research on poverty has been conducted with various methods of analysis. Faharuddin and Endrawati (2022) classified poverty in Indonesia with statistical learning with logistic regression analysis. But, such a method has high requirements on the sample size of the data, so that modeling of poverty in a region continues to grow. Alsharkawi et al. (2021) conducted research on multidimensional poverty in Jordan with machine learning with the LightGBM algorithm and Bagged Decision Tree. Yao et al. (2023) conducted research to estimate poverty reduction in China spatially by utilizing deep learning Long Short Term Memory (LSTM) analysis, a neural network technique. Kaur and Kaur (2020) conducted a study comparing several classification methods in machine learning based on their accuracy. The results showed that Random Forest is better than Logistic Regression, Decision Tree, and K-Nearest Neighbors. Based on previous research, we employed the RF algorithm as a machine learning method to explore the factors associated with educated poverty in East Java. It is important to note that the use of RF aims to analyze relationships and associations between variables, and it does not establish causal relationships.

The literature review provides a comprehensive understanding of the relationship between ICTs, access to finance, and poverty. The studies reviewed suggest that ICTs play a significant role in promoting economic and social improvement and improved access to finance can help reduce poverty and enhance household welfare. The literature highlights the positive impact of ICTs in providing financial services, information, and e-banking to the poor. The review further highlights the shift in focus from macro-level ICTs linkages to its role in poverty alleviation at the micro level in recent years. Overall, the literature supports the idea that ICTs can have a positive impact on poverty alleviation through improved access to finance.

2 DATA AND METHODOLOGY

2.1 Data sources and description

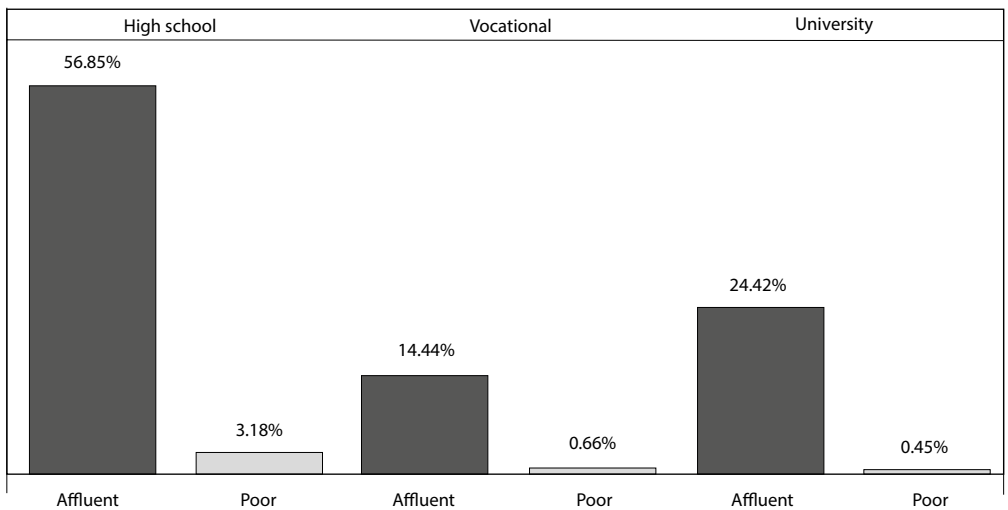
This study focuses to examine the impact of ICTs (e-commerce, laptop usage, handphone usage, internet access, social media, and digital banking) on poverty, particularly high-educated poverty, in East Java. In addition to ICTs as the independent variable, this study also involves variables that contain basic information on each respondent that can affect poverty, including age, marital status, gender, and education level. The internet access question in this study allows one to access the Internet anywhere and through personal or company-owned devices. This is fine because this variable provides information on whether a person uses the internet in their daily life. Furthermore, the ICT indicator in the form of laptop or cellphone ownership specifically requires personal ownership not belonging to the company where one works. This variable provides information that technology in the form of laptops and cellphones is needed by someone in their daily life to communicate not only at work. In order to do this, the study utilizes national survey data from the Susenas survey in 2020. The data used was obtained from a survey with community respondents in East Java with the provision of having received at least a high school education with the total of 30 719 respondents of the age of 18–60 years which is the working age in Indonesia.

East Java is one of the provinces in Indonesia with the second highest population density in Indonesia after West Java. This high population density is certainly vulnerable to economic problems, including poverty. East Java is also one of the provinces that has high economic growth potential in Indonesia (Sengaji et al., 2019). Therefore, understanding the factors that affect poverty in East Java and how to reduce it can help in designing more effective government policies and programs in reducing poverty in the province. In addition, a study of poverty in East Java can also provide greater insight into the problem of poverty in Indonesia in general, and can serve as a basis for comparing poverty rates in other provinces in Indonesia.

Poverty is a condition in which a person or family does not have the ability to satisfy minimum basic needs, such as food, clothing, housing, education, and health services. BPS measures poverty with reference to the Poverty Line or Minimum Poverty Line (MPL), which is the threshold value of income needed to fulfill these basic needs. The poverty data obtained is categorized into 2 categories, namely 0 for non-poor where a person is above the poverty line and 1 for poor where a person is below the poverty line.

The focus of this study is high-educated poverty in East java. High-educated poverty is a term that refers to the condition in which an individual or a group of people who have achieved higher education still live in poverty. High-educated poverty can occur due to several factors, such as difficulty in finding a job that matches their educational qualifications, low salaries and wages, lack of access to resources and capital, and lack of skills and experience in managing finances and businesses. High-educated poverty can have a significant impact on a person's life, such as difficulty in meeting basic needs, difficulty in obtaining adequate healthcare services, and limited access to economic and social opportunities. This can lead to a decline in the quality of life and well-being of an individual. To address high-educated poverty, efforts are needed from various stakeholders, such as the government, educational institutions, and the private sector. This research focuses on educated people in East Java who have studied high school, vocational school, and higher education with the distribution of data shown in Figure 1.

Figure 1 Information criteria and entropy of the various LCA models



Source: Own construction

Figure 1 provides information that most affluent people have a senior high school education. However, high school graduates also show the highest poverty rate compared to other graduates. When calculated at each level of education, the highest percentage of poverty occurs among senior high school graduates, which is 3.18%. Meanwhile, the poverty rate for vocational graduates is 0.66% and university graduates show the lowest poverty rate, at 0.45%. This documents that even though a person has taken a high level of education or in other words has been educated, there is still the possibility of poverty caused by other factors. The occurrence of poverty in educated communities can be caused by the use of digital technology that continues to develop in this day and age. This digital technology is the focus of this study to determine its effect on high-educated poverty in East Java.

This research also highlights the characteristics of high-educated poverty in East Java. The variables to be observed in this study are shown in Table 1.

Table 1 The variables

Variable	Category
Age	18–60 years
Gender	0: Male
	1: Female
Marital status	0: Not marriage
	1: Marriage
Internet access	0: No
	1: Yes
Mobile phone usage	0: No
	1: Yes
Social media usage	0: No
	1: Yes
Digital banking	0: No
	1: Yes
Laptop usage	0: No
	1: Yes
Education	0: University
	1: Vocational
	2: High school
Poverty status	0: Affluent
	1: Poor

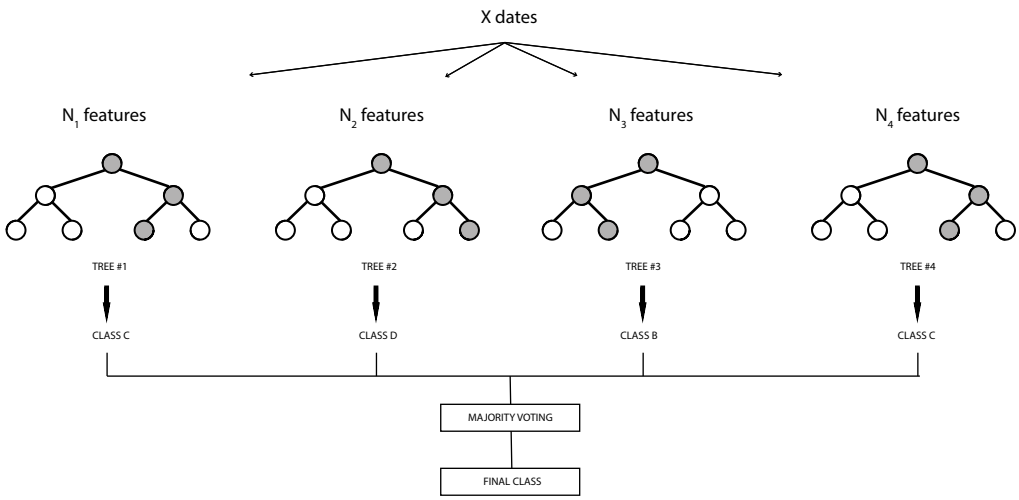
Source: Own construction

2.2 Random forest for machine learning agorithm

The machine learning technique used in this study is Random Forest (RF) analysis. RF is a type of machine learning algorithm that uses decision trees to make predictions. This method is particularly useful for this study because it can handle complex interactions between variables and handle large amounts of data. RF is one of the classification methods introduced by Breiman in 2001 as a development of the Classification and Regression Tree (CART) method. It is used to improve classification accuracy. The RF algorithm has demonstrated significant success as a classification and regression method for various general purposes. As a machine learning algorithm, it can handle large-scale problems involving big data and can be readily customized for various ad-hoc learning tasks (Mirzaei et al., 2021; Yangyudongnanxin, 2021).

The RF classification method consists of a combination of independent CART classification trees through a randomization process to form a tree on the sample and factor data. Therefore, this process will produce different classification trees. From a set of decision trees, it is expected to obtain a small correlation between trees to reduce prediction errors (Breiman, 2001). The underlying approach involves constructing multiple decision trees by randomly selecting training samples, and the output category is determined through the principle of majority voting, as demonstrated in Figure 2. In this context, the predictions for unknown samples are established based on the majority voting principle. In this study, random forest analysis will be used to examine the importance of technology to high-educated poverty modeling.

Figure 2 Poverty status on education



Source: Boateng et al. (2020)

The RF algorithm also provides a measure of feature importance. Feature importance is a measure that helps you understand which features (variables or attributes) are most influential in making predictions. Feature importance can provide insights into the relative significance of different input variables in a predictive model. It is important to note that feature importance does not establish causation, but it does help to identify which features have the most impact on the model's performance. Feature importance serves as a cornerstone for optimizing predictive models, aiding in the selection of critical features, and providing a deeper understanding of the interplay between variables and the target outcome. This information can be used to guide further research and policy interventions. This can be used to understand the importance of technology (social media, e-commerce, and digital banking) on poverty, as well as the role of education in alleviating poverty.

The use of the Susenas survey and RF analysis allows for a robust examination of the importance of technology on poverty in East Java. The survey data provides a rich and diverse data source, while the machine learning technique provides a powerful tool for analyzing complex relationships between variables. By combining these two methods, this study aims to provide a comprehensive and in-depth analysis of the impact of technology on poverty in East Java, with a focus on high-educated poverty.

The Susenas (national survey) data is collected and prepared for analysis. The data should be cleaned and checked for missing or incorrect values. The relevant variables that are believed to impact poverty in East Java, including education levels, use of technology (social media, e-commerce, and digital banking), and economic status, are selected. The data is then split into two parts, one for training and one for testing. The training data is used to build the RF model, while the testing data is used to validate the model. The random forest algorithm is applied to the training data. This involves creating a large number of decision trees, each of which is trained on a different subset of the data. The model then uses these decision trees to make predictions about the target variable (poverty status) based on the input variables.

The RF model is evaluated using the testing data. This involves comparing the model's predictions to the actual values and calculating various performance metrics, such as accuracy, precision, recall, and F1 score (Jiang et al., 2021; Matloob et al., 2021). For model evaluation and in Formulas (1), (2), (3), and (4), we expressed these evaluation metrics as follows:

$$Accuracy\left(Acc\right)=\frac{TP+TN}{TP+TN+FP+FN}, \tag{1}$$

$$Recall\left(Re\right)=\frac{TP}{TP+TN}, \tag{2}$$

$$Precision\left(Pre\right)=\frac{TP}{TP+FP}, \tag{3}$$

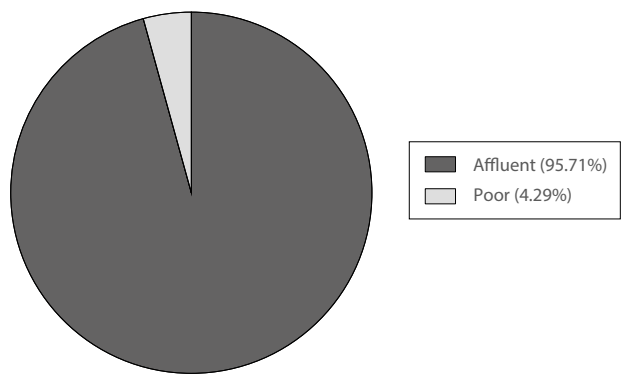
$$F1-Score=2\frac{\left(Pre\right)\left(Re\right)}{Pre+Re}. \tag{4}$$

In calculating accuracy for various types of models or performance evaluation methods, four terms are commonly used: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP occurs when the model or system correctly identifies a positive event as positive. TN occurs when a negative event is correctly identified as negative. On the other hand, FP happens when the model or system wrongly identifies a negative event as positive, while FN occurs when a positive event is wrongly identified as negative (Shah et al., 2020). In the context of classification, TP and TN refer to the number of true positive and true negative predictions, respectively, while FP and FN refer to the number of false positive and false negative predictions. Although accuracy can be calculated using TP, TN, FP, and FN, other performance metrics such as precision, recall, or F1-score may also be necessary to provide a more comprehensive evaluation of the model's performance If the performance of the model is not satisfactory, various parameters of the RF algorithm can be adjusted to improve its performance. This may involve changing the number of trees, the size of the subsets used for training, or the type of impurity measure used.

3 RESULTS AND DISCUSSION

The examination aimed to provide a preliminary understanding of the distribution of the data, which would aid in further analysis and interpretation of the results. The graphical representation provided valuable insights into the nature of the data, allowing for the identification of any potential issues or trends in the data. Through this examination, the study aimed to lay the foundation for a more in-depth and comprehensive analysis of the data in the future. The study aims to investigate the importance of ICTs on poverty in educated individuals in East Java. One of the key findings from the study is that the majority of the educated population in East Java is classified as affluent, with poverty rates at 4.29% as shown in Figure 3.

Figure 3 Percentage of poverty status in East Java



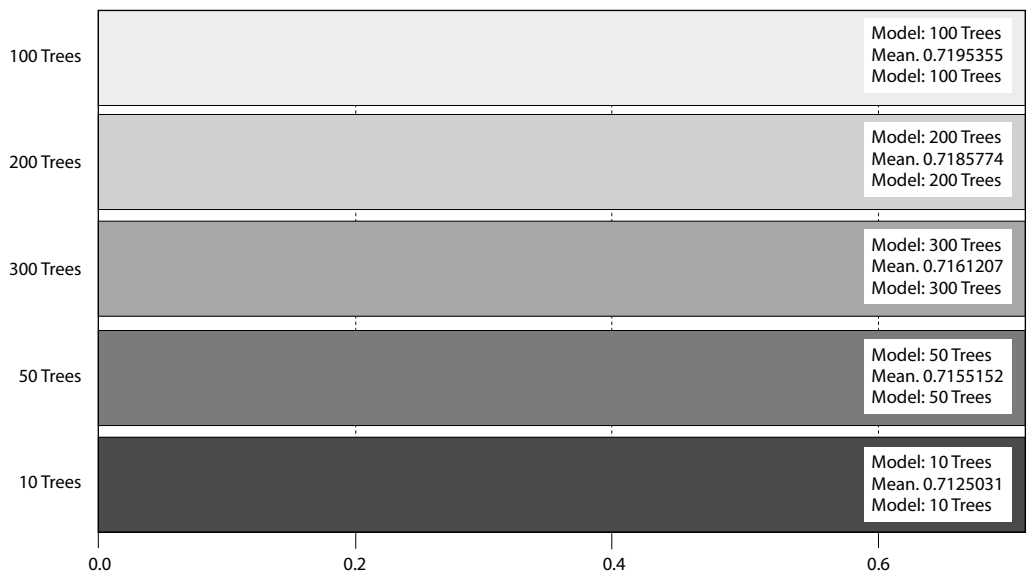
Source: Own construction

This result highlights the challenge of data imbalance in machine learning problems, as the majority class is more frequently represented in the data. Imbalance data can negatively impact the quality of the results obtained from machine learning models. The models tend to predict the majority class more often, as they have more examples to learn from. This can lead to biased results and poor performance in predicting minority or rare classes in the data. To address this issue, the study utilizes Synthetic Minority Over-Sampling Technique (SMOTE), a popular balanced data sampling technique. SMOTE creates synthetic data by modifying the unbalanced dataset and generating a balanced dataset from the unbalanced dataset. This helps to overcome the imbalance and increase the representation of the minority class in the data (Karthik and Krishnan, 2021).

The study highlights the importance of considering data imbalance and utilizing appropriate techniques such as SMOTE to overcome the challenge in machine learning problems. The results from the RF model provide a valuable insight into the importance of ICTs on poverty in educated individuals in East Java and demonstrate the effectiveness of the balanced data sampling technique in overcoming data imbalance. These findings can be used to inform future research and guide the development of more effective strategies to address poverty through the use of ICTs.

The RF model was optimized using a 10-fold cross-validation test mode with 1 000 iterations. As a result, the model is constructed using 10 trees on the grounds that a simple model can have the same accuracy as 300 trees (Figure 4). Based on the law of parsimony, we used the simplest model with 10 trees.

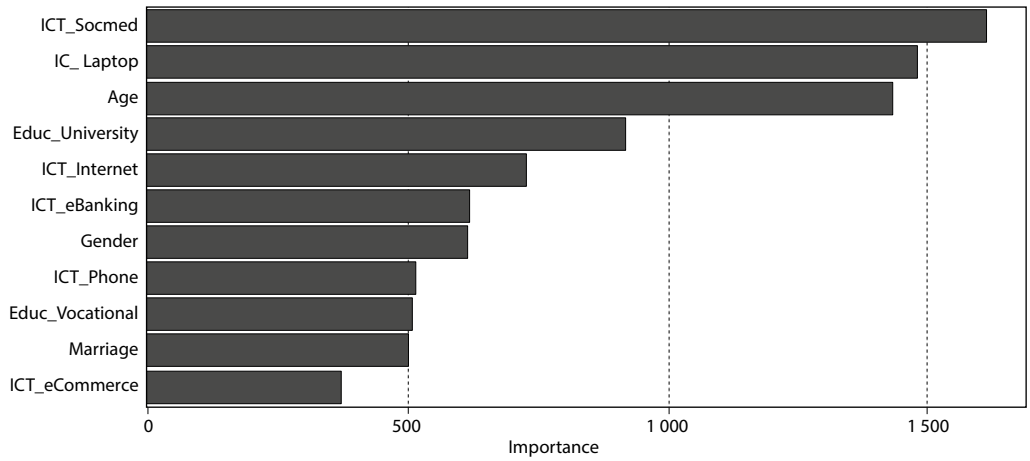
Figure 4 Selecting number of trees



Source: Own construction

Identifying the important variables in this experiment is crucial for determining if there are other factors that affect poverty classification aside from income. This is important in guiding us towards building multidimensional poverty indicators for poverty classification. Conversely, the RF algorithm is capable of generating a list of important variables by using the average merit calculation. Figure 5 presents the ranking of important variables from this experiment. It is apparent that out of the 11 variables, social media literacy, laptop ownership, age, higher education, and internet access are the five most important variables for high-educated poverty classification in East Java.

Figure 5 The variable importance



Source: Own construction

The accuracy of the RF model is evaluated using the testing data. This involves comparing the model's predictions to the actual values and calculating various performance metrics, such as accuracy, precision, recall, and F1 score are shown in Table 2.

Table 2 Evaluation model

Metric	Value
Accuracy	0.713
Precision	0.972
Recall	0.719
F1 score	0.826

Source: Own construction

Accuracy with a value of around 71.3% gives an overall picture that the model is able to correctly predict around 71.3% of individuals in the high-educated poverty group. However, a deeper focus on the precision metric indicates that the model is very efficient in identifying individuals who are actually in high-educated poverty. With a precision of about 97.2%, about 97.2% of the model's positive predictions are true positives, reducing the risk of misclassifying individuals who are not actually experiencing poverty as being in high-educated poverty. Similarly, the recall metric (Recall or Sensitivity) which has a value of around 71.9% highlights the model's ability to identify individuals who are actually in high-educated poverty. This reinforces the belief that the model is able to capture about 71.9% of all high-educated poverty cases in the data, creating a strong basis for more appropriate countermeasures. Furthermore, the F1 Score value of around 0.826 reflects a good balance between the model's ability to classify outcomes with a high degree of precision and its ability to identify actual cases of high-educated poverty. In this context, the F1 Score illustrates that the model is successful in providing overall good results in modeling high-educated poverty, considering both precision and recall.

Based on Figure 4, the results from the RF model suggest that using social media, having laptop, using internet and digital banking are among the top variables ICTs that have importance to classification high-educated poverty in East Java. The direct applications of ICTs, such as providing information about

markets by Nwafor et al. (2020), opportunities, and employment, as well as improving skills and education, health care by Tortorella et al. (2022), and delivery of government services, are critical in addressing the welfare of the poor. These technologies also have the potential to empower the poor by improving communication between them and the government, giving them new avenues to voice their grievances. It should be noted that while access to ICTs is important, it is not the only a variable that matters in high educated poverty. Age, gender, education level, marital status, and mobile phone ownership also play a role in classification of poverty. Age, gender, education level, marital status, and posession of mobile phone also play a role in this model.

The presence of social media in an individual‘s life is also seen as a critical factor of high-educated poverty classification. Social media provides people with access to information and communication that can improve their economic and social well-being. For instance, individuals can use social media to connect with potential employers, access job opportunities, market their goods and services, and access education and training. Oh and Syn (2015) indicated the advantages of social media. Social media can also serve as a platform for building networks, seeking support, and exchanging ideas and information. All these benefits can contribute to an improvement in the economic status and overall well-being of individuals. However, excessive and uncontrolled use of social media can disrupt productivity and lead to unhealthy dependence (Achmad et al., 2022). Therefore, it is important for individuals to manage their social media usage and find the right balance between work and social life.

The importance of having a laptop in high-educated poverty is shown in the RF results. This is because someone who is in the poverty category will usually find it difficult to buy a laptop. Of course, this will hamper one‘s work skills and creativity considering that this is the digital era. Laptops provide access to information and opportunities that can improve economic and social well-being. This includes access to online markets, education and employment opportunities, as well as communication platforms and social media. With laptops, individuals can engage in digital commerce and grow their businesses, access financial services, and receive and transmit information. This enables them to improve their financial literacy and overall standard of living.

Age is also a critical factor in high-educated poverty classification. Generally, as individuals grow older and gain experience, they are able to acquire more skills, knowledge, and assets. These resources can help individuals to participate more effectively in the economy, thus improving their financial stability and reducing poverty. This is consistent with research by Baker et al. (2018) that the older individuals may have a greater sense of financial literacy, which can enable them to access financial services and make better financial decisions.

The use of the Internet is also the importance variable in the model of high-educated poverty in East Java. The Internet provides individuals with access to a wealth of information, as well as new and innovative ways of participating in the economy. For instance, individuals can use the Internet to access online markets, education, and job opportunities. The Internet can also be used to connect with others and build networks, which can be critical in securing support, resources, and information. All of these benefits can contribute to an improvement in the financial stability and overall well-being of individuals.

Having a mobile phone is also seen as an important factor in high-educated poverty modeling. Mobile phones provide individuals with access to information and communication that can improve their economic and social well-being. For instance, individuals can use mobile phones to access online markets, education, and job opportunities. Mobile phones can also be used for financial services, such as mobile banking, which can greatly enhance financial inclusion and improve financial stability. Furthermore, mobile phones can provide individuals with a sense of security and connect them with others, which can be critical in times of need and uncertainty.

CONCLUSION

The results of the RF model revealed that the top five most important variables in high-educated poverty in East Java are owning a laptop, having a social media account, age, using the internet, and having a mobile phone. The findings show that ICT use is the most important variable for classifying high-educated poverty in East Java. The ownership of a laptop and having a social media account, both ranking high in importance, signal the growing relevance of digital tools and online presence in today's interconnected world. These findings may hint at the potential opportunities and challenges faced by highly educated individuals, as these two variables could relate to job search, remote work, or online networking for economic advancement.

Technology, in the form of social media presence, laptop ownership and Internet access, is identified in this study as a factor that is important to modeling the high-educated poverty. However, we understand that poverty cannot be simplified into a direct result of not having technology. Poverty is a complex and often multidimensional phenomenon, influenced by various factors such as income, education, employment, and access to resources. Therefore, this study found that technology was one of the factors that contribute to overcoming poverty, but not the only cause. In-depth understanding needs to be done on how technology can act as a tool of reducing high-educated poverty, while still understanding the complexity of the poverty phenomenon and its impact on technology access.

In order to reduce the level of high-educated poverty in East Java, a number of concrete recommendations can be made to the government. First, it is necessary to launch an ICTs training program that focuses on educated people who still experience poverty. This includes training in computer use, applications, and internet access. In addition, initiatives should provide free or affordable internet access in areas where the majority of the population is out of reach thereof. This program should cover both rural and less developed urban areas in East Java.

Second, the government needs to allocate resources for digital infrastructure development in remote and isolated areas. This includes the development of strong internet networks and stable electricity. Collaboration with internet and electricity service providers, as well as Village-owned enterprises, can ensure reliable access to digital technology. Furthermore, digital-based economic empowerment programs should be a priority. This involves training in app creation, e-commerce, digital marketing and online business management. Financial support and mentoring for local digital entrepreneurs should be strengthened. Fourth, the government should pass strict regulations related to data security and user privacy in the use of digital technology. An independent oversight body should be established to deal with data and privacy breaches. Finally, regular monitoring and evaluation of these programs should be conducted. Collaboration with the private sector, financial institutions, and non-governmental organizations that have competencies in digital technology and poverty alleviation can support the implementation of these recommendations. With these measures, the government is expected to make a significant contribution in addressing high-educated poverty in East Java through better and effective utilization of the potential of digitalization.

There are several limitations that need to be acknowledged when interpreting the results of this study. Firstly, the study was conducted solely in East Java, Indonesia, which may limit its generalizability to other regions or countries. Secondly, the study was based on cross-sectional data, meaning that causality between the independent and dependent variables cannot be established. Finally, self-reported data, such as the information collected in this study, may be prone to measurement errors such as recall bias or social desirability bias. As a result, these limitations need to be considered when interpreting the findings of this study.

To further advance the understanding of the relationship between ICT and poverty among educated individuals in East Java, future research could be enhanced to address the limitations of this study. One potential approach is to include other socio-economic variables that may have an impact

on poverty, such as household income and employment status, which could provide a more comprehensive understanding of the relationship between ICT and poverty. In addition, analyzing the impact of ICTs on poverty at the community level can reveal potential benefits and challenges in implementing ICT interventions in poor areas. Finally, conducting longitudinal studies can help to better understand changes in poverty levels over time, and the long-term impact of ICT interventions on poverty alleviation. By exploring these recommendations, future research can provide a more in-depth and comprehensive understanding of the role of ICT in alleviating poverty among educated individuals in East Java.

References

ABDUVALIEV, M., BUSTILLO, R. (2019). Impact of remittances on economic growth and poverty reduction amongst CIS countries [online]. *Post-Communist Economies*, 1–22. <https://doi.org/10.1080/14631377.2019.1678094>.

ACHMAD, W., SUDRAJAT, A., FAIZA, S., OLLIANTI, R. N. (2022). The Influence of Social Media on Teenagers' Lifestyles: Behavioral Analysis Among Adolescents in Bandung. *Journal on Education*, 5(3): 10356–10363.

ALSHARKAWI, A., AL-FETYANI, M., DAWAS, M., SAADEH, H., ALYAMAN, M. (2021). Poverty Classification Using Machine Learning: The Case of Jordan [online]. *Sustainability* (Switzerland), 13(3): 1–16. <https://doi.org/10.3390/su13031412>.

AO, Y., LI, H., ZHU, L., ALI, S., YANG, Z. (2019). The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling [online]. *Journal of Petroleum Science and Engineering*, 174: 776–789. <https://doi.org/10.1016/j.petrol.2018.11.067>.

BAKER, H. K., KUMAR, S., GOYAL, N., GAUR, V. (2018). How financial literacy and demographic variables relate to behavioral biases [online]. *Emmerald Insight*, 45(1): 124–146. <https://doi.org/10.1108/MF-01-2018-0003.

BHAVNANI, A., CHIU, R. W.-W., JANAKIRA, S., SILARZSKY, P. (2008). Ict Policy Division Global Information and Communications Department (Gict) the Role of Mobile Phones in Sustainable Rural Poverty Reduction. In: *Global Information and Communications Department*.

BOATENG, E. Y., OTOO, J., ABAYE, D. A. (2020). Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review [online]. *Journal of Data Analysis and Information Processing*, 8(4): 341–357. <https://doi.org/10.4236/jdaip.2020.84020>.

BREIMAN, L. (2001). Random Forests [online]. *Machine Learning*, 45: 5–32. <https://doi.org/10.1023/A:1010933404324>.

CHATTERJEE, A. (2020). Financial inclusion, information and communication technology diffusion, and economic growth: A panel data analysis [online]. *Information Technology for Development*, 26(3): 607–635. <https://doi.org/10.1080/02681102.2020.1734770>.

CHERNOVA, V. Y., ZOBOV, A. M., STAROSTIN, V. S. (2019). Dynamics of digital inequality in Russia [online]. *Media Watch*, 10(2): 225–234. <https://doi.org/10.15655/mw/2019/v10i2/49622>.

CHIB, A., MAY, J., BARRANTES, R. (2015). Impact of information society research in the global south [online]. In: *Impact of Information Society Research in the Global South*. <https://doi.org/10.1007/978-981-287-381-1>.

DIOP, S., ASONGU, S. A. (2021). The Covid-19 Pandemic and the New Poor in Africa: The Straw That Broke the Camel's Back [online]. In: *Forum for Social Economics*, Routledge. <https://doi.org/10.1080/07360932.2021.1884583>.

ERLANDO, A., RIYANTO, F. D., MASAKAZU, S. (2020). Financial inclusion, economic growth, and poverty alleviation: Evidence from eastern Indonesia [online]. *Heliyon*, 6(10): e05235. <https://doi.org/10.1016/j.heliyon.2020.e05235>.

FAHARUDDIN, F., ENDRAWATI, D. (2022). Determinants of working poverty in Indonesia [online]. *Journal of Economics and Development*, 24(3): 230–246. <https://doi.org/10.1108/jed-09-2021-0151>.

FINCH, W. H., FINCH, M. E. H. (2020). Poverty and Covid-19: Rates of Incidence and Deaths in the United States During the First 10 Weeks of the Pandemic [online]. *Frontiers in Sociology*, 5(June): 1–10. <https://doi.org/10.3389/fsoc.2020.00047>.

GUVEN, H. (2020). Industry 4.0 and Marketing 4.0 [online]. In: *Perspective of Digitalization and E-Commerce, Agile Business Leadership Methods for Industry 4.0*, 25–46. <https://doi.org/10.1108/978-1-80043-380-920201003>.

HOFMARCHER, T. (2021). The effect of education on poverty: A European perspective [online]. *Economics of Education Review*, 83(April): 102124. <https://doi.org/10.1016/j.econedurev.2021.102124>.

ISUKUL, A., TANTUA, B. (2021). Financial Inclusion in Developing Countries: Applying Financial Technology as a Panacea [online]. *South Asian Journal of Social Studies and Economics*, January: 42–60. <https://doi.org/10.9734/sajsse/2021/v9i230237>.

JACOBUS, E. H., KINDANGEN, P., WALEWANGKO, E. N. (2019). Analisis Faktor-Faktor Yang Mempengaruhi Kemiskinan Rumah Tangga Di Sulawesi Utara [online]. *Jurnal Pembangunan Ekonomi Dan Keuangan Daerah*, 19(7): 86–103. <https://doi.org/10.35794/jpekd.19900.19.7.2018>.

JIANG, T., LI, J. P., HAQ, A. U., SABOOR, A., ALI, A. (2021). A Novel Stacking Approach for Accurate Detection of Fake News [online]. *IEEE Access*, 9: 22626–22639. <https://doi.org/10.1109/ACCESS.2021.3056079>.

KARTHIK, M. G., KRISHNAN, M. B. M. (2021). Hybrid random forest and synthetic minority over sampling technique for detecting internet of things attacks [online]. *Journal of Ambient Intelligence and Humanized Computing*, 0123456789. <https://doi.org/10.1007/s12652-021-03082-3>.

KAUR, I., KAUR, J. (2020). Customer churn analysis and prediction in banking industry using machine learning [online]. *PDGC 2020–2020 6th International Conference on Parallel, Distributed and Grid Computing*, 434–437. <https://doi.org/10.1109/PDGC50313.2020.9315761>.

KENDALL, J., MYLENKO, N., PONCE, A. (2010). Measuring Financial Access around the World. *Policy Research Working Paper 5253 Abstract*, March: 1–21.

LOVARI, A., VALENTINI, C. (2020). Public Sector Communication and Social Media [online]. In: *The Handbook of Public Sector Communication*, 315–328. <https://doi.org/10.1002/9781119263203.ch21>.

MATLOOB, F., GHAZAL, T. M., TALEB, N., AFTAB, S., AHMAD, M., KHAN, M. A., ABBAS, S., SOOMRO, T. R. (2021). Software defect prediction using ensemble learning: A systematic literature review [online]. *IEEE Access*, 9: 98754–98771. <https://doi.org/10.1109/ACCESS.2021.3095559>.

MCQUIRE, S. (2021). Urban Digital Infrastructure, Smart Cityism, and Communication: Research Challenges for Urban E-Planning [online]. *International Journal of E-Planning Research*, 10(3): 1–10. <https://doi.org/10.4018/IJEPR.20210701.oa1>.

MEO, M. S., KUMAR, B., CHUGHTAI, S., KHAN, V. J., DOST, M. K. BIN, NISAR, Q. A. (2020). Impact of Unemployment and Governance on Poverty in Pakistan: A Fresh Insight from Non-linear ARDL Co-integration Approach [online]. *Global Business Review*. <https://doi.org/10.1177/0972150920920440>.

MIRZAEI, S., VAFAKHAH, M., PRADHAN, B., ALAVI, S. J. (2021). Flood susceptibility assessment using extreme gradient boosting (EGB), Iran [online]. *Earth Science Informatics*, 14(1): 51–67. <https://doi.org/10.1007/s12145-020-00530-0>.

MISHRA, V., SINGH BISHT, S. (2013). Mobile banking in a developing economy: A customer-centric model for policy formulation [online]. *Telecommunications Policy*, 37(6–7): 503–514. <https://doi.org/10.1016/j.telpol.2012.10.004>.

NILI, S. (2019). Global poverty, global sacrifices, and natural resource reforms [online]. *International Theory*, 11(1): 48–80. <https://doi.org/10.1017/S1752971918000209>.

NWAFOR, C. U., OGUNDEJI, A. A., VAN DER WESTHUIZEN, C. (2020). Adoption of ICT-based information sources and market participation among smallholder livestock farmers in South Africa [online]. *Agriculture* (Switzerland), 10(2). <https://doi.org/10.3390/agriculture10020044>.

OH, S., SYN, S. Y. (2015). Motivations for Sharing Information and Social Support in Social Media: A Comparative Analysis of Facebook, Twitter, Delicious, YouTube, and Flickr [online]. *Journal of the American Society for Information Science and Technology*, 66: 2045–2060. <https://doi.org/10.1002/asi>.

PRADHAN, R. P., ARVIN, M. B., NAIR, M. S., HALL, J. H., BENNETT, S. E. (2021). Sustainable economic development in India: The dynamics between financial inclusion, ICT development, and economic growth [online]. *Technological Forecasting and Social Change*, 169(May 2020): 120758. <https://doi.org/10.1016/j.techfore.2021.120758>.

SEBASTIANI, A. (2021). Digital artifacts and landscapes. Experimenting with placemaking at the impero project [online]. *Heritage*, 4(1): 281–303. <https://doi.org/10.3390/heritage4010018>.

SENGAJI, M., SASONGKO, S., SAKTI, R. (2019). Determinant of Regional Economic Growth in East Java, Indonesia [online]. *Jurnal Ekonomi Dan Studi Pembangunan*, 11(2): 104–115. <https://doi.org/10.17977/um002v11i22019p104>.

SHAH, K., PATEL, H., SANGHVI, D., SHAH, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification [online]. *Augmented Human Research*, 5(1). <https://doi.org/10.1007/s41133-020-00032-0>.

SIDDIQA, A. (2021). Determinants of Unemployment in Selected Developing Countries: A Panel Data Analysis [online]. *Journal of Economic Impact*, 3(1): 19–26. <https://doi.org/10.52223/jei3012103>.

TORTORELLA, G. L., FOGLIATTO, F. S., SAURIN, T. A., TONETTO, L. M., MCFARLANE, D. (2022). Contributions of Healthcare 4.0 digital applications to the resilience of healthcare organizations during the Covid-19 outbreak [online]. *Technovation*, 111(September 2021): 102379. <https://doi.org/10.1016/j.technovation.2021.102379>.

VAKULENKO, Y., SHAMS, P., HELLSTRÖM, D., HJORT, K. (2019). Service innovation in e-commerce last mile delivery: Mapping the e-customer journey [online]. *Journal of Business Research*, 101(June 2018): 461–468. <https://doi.org/10.1016/j.jbusres.2019.01.016>.

WEWEGE, L., LEE, J., THOMSETT, M. C. (2020). Disruptions and Digital Banking Trends. *Journal of Applied Finance & Banking*, 10(6): 15–56.

YANGYUDONGNANXIN, G. (2021). Financial Credit Risk Control Strategy Based on Weighted Random Forest Algorithm [online]. *Scientific Programming*. <https://doi.org/10.1155/2021/6276155>.

YAO, Y., ZHOU, J., SUN, Z., GUAN, Q., GUO, Z., XU, Y., ZHANG, J., HONG, Y., CAI, Y., WANG, R. (2023). Estimating China's poverty reduction efficiency by integrating multi-source geospatial data and deep learning techniques [online]. *Geo-Spatial Information Science*, 1–17. <https://doi.org/10.1080/10095020.2023.2165975>.

Trade Integration of India with South and Central Asia: Evidence from Gravity Model

Suadat Hussain Wani[1] | *Indian Council of Social Science Research, New Delhi, India*
Effat Yasmin[2] | *University of Kashmir, Srinagar, India*

Received 8.7.2023 (revision received 7.9.2023), **Accepted** (reviewed) 16.9. 2023, **Published** 15.3.2024

Abstract

The present study investigates exports of India with selected South and Central Asian countries and their top trade partners. The panel data from 1996–2020 is used and PPML estimation technique which takes care of heterogeneity, zero trade and endogeneity is applied. In addition to basic variables, institutional setup in selected countries is included in the model to get better results. To check robustness of results, in addition to PPML, Heckman selection model is also estimated. The findings indicate that institutions play an important role in export promotion of India to these countries. Among various institutional factors, effectiveness of government in different aspects has considerable influence on exports from the country. The study concludes that institutional quality needs to be improved to enhance trade cooperation of India with these countries.

Keywords	***DOI***	***JEL code***
International trade, gravity model, institutional factors, regional trade agreement, panel data	*https://doi.org/10.54694/stat.2023.35*	*F13, F14, F15*

INTRODUCTION

From the beginning of 21st century, due to changing economic situation around the world, India took a keen interest in improving cooperation with Central Asian states. On the other hand, given the land-locked nature and lack of connectivity, these countries expressed keen interest to increase cooperation with India (Sharma, 2012). To enhance cooperation, Connect Central Asia policy was announced by India in addition to India-Central Asia dialogue at track II to be held annually in one of these republics. It is under the mantra of the four C's (commerce, connectivity, community and consular) that define the renewed interest of India in the region (Roy, 2013). In addition, the country took an active part in International "North-South Transport Corridor (INSTC)" and along with Pakistan joined the Shanghai Cooperation Organization (SCO) in 2016. The country wants to develop bilateral and multilateral

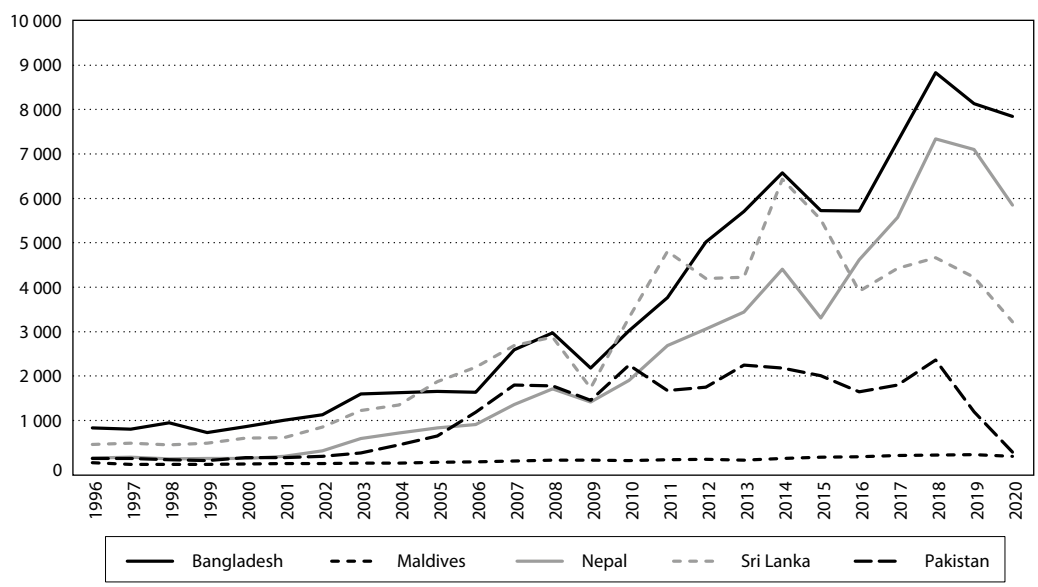
[1] Indian Council of Social Science Research (ICSSR), New Delhi 110067, India. Also Department of Economics, University of Kashmir, Hazratbal, Srinagar 190006, Jammu and Kashmir, India. E-mail: suadat.scholar@kashmiruniversity.net.
[2] Department of Economics, University of Kashmir, Hazratbal, Srinagar 190006, Jammu and Kashmir, India. E-mail: fgulwani@gmail.com.

partnership with the region (Chatterjee, 2018). Balooch (2009) suggests that Iran can act as gateway for trade enhancement between India and the region by providing access through Chabahar port. To avail this India acceded to "Customs Convention on International Transport of Goods" under cover of TIR Carnets in 2017 and joined "Ashgabat Agreement" in 2018.

In case of South Asia, in line with various trade agreements signed worldwide, SAARC member countries[3] also signed "South Asia Free Trade Agreement (SAFTA)" on January 6, 2006. In addition, various bilateral and multilateral trade agreements have been signed to enhance trade cooperation. However, despite 21 percent population of world and common culture and history, only 5 percent intra-regional trade takes place as compared to 35 percent in case of East Asia (Kathuria and Shahid, 2017). According to the World Bank Report (2014), South Asian countries have recorded average growth rate of 6 percent during the last 20 years. However, the contribution of trade among SAARC member countries continues to be insignificant as compared to trade with the outside world. Similarly, Iqbal and Nawaz (2017) concluded that SAFTA has been unable to create regional trade, which raises the question of what makes SAFTA ineffective. However, despite many measures, trade relations of India with South and Central Asian countries have not increased up to the expected magnitude, as is clear from Figures 1 and 2.

There are various obstacles to international trade ranging from geographical distance to formal trade barriers (Trefler, 1995; Anderson and Van Wincoop, 2004) which add to trade costs and reduce trade volume between countries. Anderson and Van Wincoop (2004) concluded that they account for ad valorem tax equivalent of around 170 percent for a representative developed country. However, Deardoff (1998) is of the view that in addition to observed costs, trade also includes unobserved trade costs, particularly difference in institutional quality between nations. It is due to these differences that firms have to bear

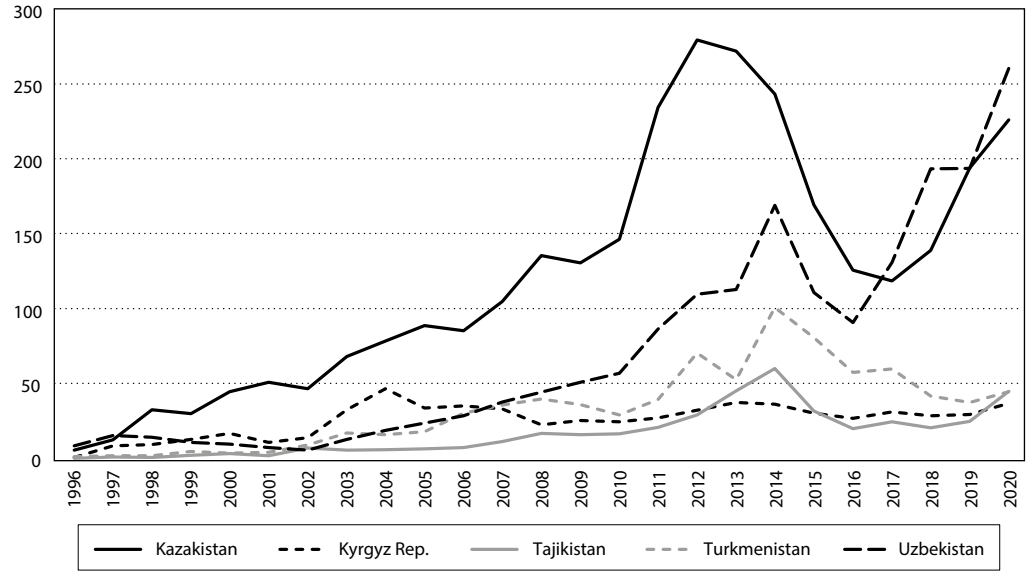
Figure 1 Exports from India to South Asia (million USD)



Source: Author's calculation based on data from DOTS, IMF

[3] Afghanistan, Bangladesh, Bhutan, Maldives, Nepal, India, Pakistan, Sri Lanka.

Fiure 2 Exports from India to Central Asia (million USD)



Source: Author's calculation based on data from DOTS, IMF

additional costs related to foreign markets which leads to increase in trade costs (Berthelon and Freund, 2004; Hofstede, 2001). Institutional factors affect international trade in different ways. First, it is the quality of institutions mainly related to political and legal system that implements policies ensuring security of trade between countries. The effectiveness of formal regulations in a particular country impacts the way of doing business and interpersonal trust. Zhang et al. (2003) concluded that it is the implementation of property rights and adherence to trade contracts that vary significantly between countries.

Similarly, Achrol et al. (1983), Skarmeas et al. (2002) argue that environmental uncertainty leads to conflicts between partner countries. In addition, Rodrik (2000) has highlighted difficulties to enforce international contracts as compared to domestic ones. At the same time, Wei (2000) has concluded that low quality of institutions led to negative externality for private transactions, which raises costs and reduce bilateral trade. Highlighting the positive role of institutions, Anderson and Marcouiller (2002), or Jansen and Kyvik Nordas (2004) are of the view that business supported by effective rule of law and government regulations leads to enhanced trade cooperation between countries.

Various studies (Bown, 2017; Handley and Limao, 2017) conclude that well-functioning institutions diminish uncertainty related to trade and facilitate transaction between countries. However, threats to trade agreements which arise from delay could be determinantal to economic growth of a nation. Kamal and Zaki (2018) argue that countries sign trade agreements to reduce uncertainty related to bilateral trade. However, it is the structure of institutions in these countries that regulates implementation of these agreement. Thus bilateral trade is directly related to the quality of institutions. Improvement in institutional quality enhances bilateral trade, whereas low institutional quality impedes bilateral trade cooperation. Various factors like corruption, inadequate information and inadequate enforcement of contracts leads to increase in costs related to trade which adversely impacts export competitiveness of a country (Bakhsh et al., 2022).

The present study contributes to the current literature in many significant ways. First, to our knowledge, this is the first attempt to investigate regional integration of India with South and Central Asia. Second,

although plethora of studies (Kumar and Ahmed, 2015; Jan and Shah, 2019; Sharma and Kumar, 2021; Wani and Mir, 2023) highlight many factors which deter the progression of regional integration in South Asia, role of institutions has been neglected in available literature. The present study tests the hypothesis that institutional quality of partner countries matters for bilateral trade. To check the influence of institutional factors on exports, average of institutional factors is included in the model. Then, in line with objectives of the study, individual institutional factors are incorporated. Third, previous studies were criticized for failing to consider econometric issues related to trade costs. To address various econometric issues related to trade data, PPML is applied, which considers both time and country-pair fixed effects. In addition to PPML, Heckman selection model is estimated to check the coherence of results.

1 REVIEW OF LITERATURE

In available literature, quite a few studies have focused on the relationship between trade and institutional quality of different countries. Various studies (Anderson and Marcouiller, 2002; Francois and Manchin, 2013; De Mendonca, 2014) concluded that well-defined institutional setup is positively related to trade cooperation between countries. Whereas Iqbal and Nawaz (2017), and Nawaz (2020) argue that institutional arrangements support regional integration and impact cooperation between countries both directly and indirectly. Kea et al. (2019) is of the view that Cambodia needs to reform its institutions and monetary system to reduce the negative impact of external shocks and economic recession on rice exports. Aluko et al. (2021) concluded that FDI can boost economic growth in countries that enjoy certain degree of economic freedom. However, in case of those countries which enjoy less economic freedom, FDI can hardly improve their economic performance.

Chi and Kilduff (2010) argue that regional trade agreements, language and infrastructure have played a crucial role in increasing apparel imports of the USA from partner countries. Dincer (2014) has highlighted the role of R&D (research and development) in imports of Turkey from BRICS countries. The study concludes that R&D in Turkey negatively impacts imports from BRICS countries. However, increase in R&D in BRICS has enhanced imports of Turkey from these countries. Given the significance of agri-food exports in case of Nigeria, Abdullahi et al. (2021) argue that the national income of partner countries, membership of trade agreements and common borders enhance agri-food exports. On the other hand, geographical distance, exchange rate and landlocked geography negatively impact agri-food exports. The country enjoys agri-food export potential with major countries of the world which include Brazil, China, EU, Russia, USA and border countries. Guan et al. (2021) argue that growth rate of national income, exchange rate and FDI significantly affect trade between Mauritius and China. The study concludes that the two countries need to formulate policies to remove remaining restrictions to attract more mutual investment, which can further enhance bilateral trade. In case of Bangladesh, Shahriar et al. (2021) examined trade performance of leather products from 1989 to 2015. The study highlights the problem of high trade and transport costs associated with trade performance of the country.

In case of Central Asia, Suvankulov and Guc (2012) argue that though many countries, including China, India, Iran, Russia and Turkey have attempted to enhance trade cooperation with this region, China and Turkey have been most successful. Similarly, Agrawal and Sangita (2013), and Malik and Mir (2014) suggest that there is vast trade potential between India and Central Asia. According to Roy (2013), India still struggles to boost collaboration with Central Asia on trade and economic front.

In case of South Asian, various studies (Shaikh and Rahpoto, 2009; Shaikh et al., 2012; Regmi, 2017) argue that SAFTA has been instrumental in enhancing regional trade. Besides, Derosa and Govindan (1996), and Pigato et al. (1997) suggest that regional cooperation can benefit participating nations. Similarly, Hassan (2001), and Rahman, Shadat and Das (2006) concluded that opportunities for trade enhancement existed among South Asian nations. However, many other studies argue that, for various reasons, SAFTA has been ineffective in enhancing regional integration. Baysan et al. (2006) argue that due to small market

size and high non-tariff restrictions among member countries, SAFTA has been ineffective. According to Akhter and Ghani (2010), in short run SAFTA has not been able to benefit the member countries, though in long run it may be beneficial. Similarly, Sultana and Asrat (2014) argue that a high level of protection and infrastructural and institutional factors have been the main reason for less economic integration in South Asian. In addition, Kathuria and Shahid (2017) also concluded that political differences and weak regulating framework negatively impact trade in the region. Finally, Abbas and Waheed (2019) argue that member countries must revisit the regional trade agreement as SAFTA has been ineffective to create opportunities for trade among them. Thus from the above discussed literature, it can be concluded that institutional quality plays a vital role in trade cooperation between countries.

2 METHODOLOGY

2.1 Gravity model

The gravity model of trade propounded by Tinbergen (1962) and Poyhonen (1963) specifies that trade flow from one country to another country can be explained by economic forces in these countries and those economic and other forces resisting or aiding this flow. The model explains the flow of trade between the set of countries as being proportional to their national income (GDP) and inversely proportional to the geographical distance between them. The linear form of classical gravity model can be expressed as:

$$\ln(T_{ij}) = \alpha_0 + \alpha_1 \ln(GDP_i) + \alpha_2 \ln(GDP_j) + \alpha_3 \ln(Dis_{ij}), \quad (1)$$

where T_{ij} represents bilateral trade between countries i and j, α_0 is a constant, GDP_i and GDP_j denote the national income of two countries, Dis_{ij} shows geographical distance, and α_1, α_2, α_3 are parameters. From Formula (1), it is clear that countries with high income tend to have high volume of bilateral trade than countries with low levels of income. Head (2003) argues that the economic size of partner countries specifies demand and supply forces. Similarly, geographically close countries tend to trade more as compared to geographically distant countries. Thus, distance reflects trade costs as it represents economic and non-economic factors.

In addition to national income and distance, various other explanatory factors like common border, common language, regional trade agreements, difference in per capita gross domestic product (PCGDPD) and institutional factors[4] and average of institutional factors are included in the model. After adding these variables, augmented gravity model is given as follows:

$$\begin{aligned}\ln(T_{ijt}) = {} & \alpha_0 + \alpha_1 \ln(GDP_{it}) + \alpha_2 \ln(GDP_{jt}) + \alpha_3 \ln(PCGDPD_{ijt}) + \alpha_4 \ln(Dis_{ij}) + \alpha_5 (Contig_{ij}) + \\ & \alpha_6 (Comlang_{ij}) + \beta_1 \ln(Inst_{jt-1}) + \beta_2 (LGEP_{jt}) + \beta_3 (LPSP_{jt}) + \beta_4 (LRQP_{jt}) + \beta_5 (LRolP_{jt}) + \\ & \beta_6 (LVAP_{jt}) + \beta_7 (LCCP_{jt}) + \varepsilon_{ijt}, \end{aligned} \quad (2)$$

where basic variables are the same as in equation (1), $PCGDPD_{ijt}$ measures the difference in per capita gross domestic product of partner countries to test the presence of intra-industry or inter-industry trade. The gravity model forecasts that nations with different per capita GDP, trade less than those with similar level of per capita GDP. Similarly, $Contig_{ij}$ and $Comlang_{ij}$ represent common border and common language between partner countries. In addition, $Inst_{jt-1}$ measures partner countries institutional quality index at the level. Besides individual institutional factors at level have been incorporated to get a better idea

[4] Which include control of corruption (CC), government effectiveness (GE), political stability and absence of violence/terrorism (PA), rule of law (ROL), regulatory quality (RQ), voice and accountability (VA).

of their role in trade cooperation. Given the objectives of study, institutional factors of partner countries are included. In the available literature, gravity model has been used to answer various questions ranging from role of border, effect of various trade agreements (Soloaga and Wintersb, 2011; Baier et al., 2019) to role of various cultural and institutional factors in trade cooperation.

2.2 Econometric approach and data source

In traditional gravity model, use of linear logarithmic form was widespread. However, the recent literature argues that linear approach leads to biased results, which are inconsistent in presence of heteroscedasticity (Silva and Tenreyro, 2006). Besides, logarithm operator applied to trade flow leads to the problem of zero trade flow[5] as log of zero is undetermined. In the present study, estimation is performed using Poisson Pseudo Maximum Likelihood (PPML) proposed by Silva and Tenreyro (2006, 2010, 2015), taking into account the fact that present sample includes large number of zeros. This estimator has been widely used in recent studies due to its consistent results (Egger and Nigai 2015). Silva and Teneyro (2006) strongly recommends use of PPML rather than OLS because the former includes difference in size of coefficients which are smaller and more suitable. Head and Mayer (2014) argue that in presence of dummies, PPML gives several advantages as compared to other estimators. One obvious challenge in gravity model estimation is that the multilateral resistance terms highlighted by Anderson and van Wincoop (2003) are not directly observable. Following Olivero and Yotov (2012), exporter-time and importer-time fixed effect are used to account for multilateral resistance terms. Another challenge in gravity model estimation is issue of endogeneity related to trade policy variable. Following Baier and Bergstrand (2007) country-pair fixed effect are used to control for potential endogeneity concerns. In addition, year-fixed effect is used to control for macroeconomic shocks. It is important to note here that set of country-pair fixed effect absorbs all bilateral time invariant variables that are used in gravity regression. Taking into account all the above mentioned issues and following (Westerlund and Wilhelmsson 2006; Alvarez et al. 2018; Nawaz, 2020), gravity model is estimated as:

$$\begin{aligned} T_{ij} =& \alpha_0+\alpha_1 \ln(GDP_i) +\alpha_2 \ln(GDP_j) +\alpha_3 \ln(PCGDPD_{ij}) +\alpha_4 \ln(Dis_{ij}) +\alpha_5 (Contig_{ij}) + \\ & \alpha_6 (Comlang_{ij}) +\beta_1(SAFTA_{ij}) + \beta_2(ECO_{ij}) + \beta_3(Inst_{jt-1}) + \beta_4(LGEP_{jt}) + \beta_5(LPSP_{jt}) + \\ & \beta_6(LRQP_{jt}) + \beta_7(LRolP_{jt}) + \beta_8(LVAP_{jt}) + \beta_9(LCCP_{jt}) + E_{ij} + I_{IJ} + C_{ij} + \varepsilon_{ijt}. \end{aligned} \quad (3)$$

In addition to the variables discussed above, Formula (3) includes SAFTA and ECO,[6] two regional trade agreements of South Asian and Central Asian states. Moreover, along the lines of suggestions of Matyas (1997, 1998), exporter (E_{ij}), importer (I_{ij}), and country dummies (C_{ij}) are added to take care of unobserved factors.

The dataset used is a balanced panel that includes 11 countries of South and Central Asia and top 10 trade partners[7] of these two regions from 1996 to 2020.

Finally, to check the robustness of results, Heckman Sample Selection model is applied, consisting of sample selection (Formulas 4 and 5) and an outcome model (Formula 6) to confirm the results of previous models. It is important to mention here that different econometric techniques supplement each other

$$E_{ijt} = nZ_{ijt} + \mu_{ijt}, \quad (4)$$

[5] For further detail – Golovko and Sahin (2021).

[6] Afghanistan, Azerbaijan, Iran, Kazakhstan, Kyrgyz Republic, Pakistan, Tajikistan, Turkey, Turkmenistan, Uzbekistan.

[7] Bangladesh, Bhutan, Maldives, Nepal, Pakistan, Sri Lanka, Kazakhstan, Kyrgyz, Tajikistan, Turkmenistan, Uzbekistan, China, Myanmar, Russia, Iran, Singapore, Switzerland, USA, Saudi Arabia, UAE, Hong Kong.

Table 1 Sources of data

Variable	Source of data
Bilateral exports	Direction of trade statistics, IMF
GDP	WDI, World Bank
Per capita GDP	WDI, World Bank
Distance	CEPII
Language	CEPII
Common border	CEPII
Regional trade agreement	World Trade Organization
Institutional factors	WGI, World Bank

Source: Author's calculation

where E_{ijt} is the latent variable and is not observed; however, in the present study, it is observed whether countries have trade, such as $E_{ijt} = 1$ if $E_{ijt} > 0$ and $E_{ijt} = 1$if $E_{ijt} = 0$ and is a vector of variables that effect E_{ijt}.

Selection model (institutions at level):

$$T_{ij}=\alpha_0 +\beta_1 \ln GDP_i + \beta_2 \ln GDP_j + \beta_3 \ln Dis_{ij} + \beta_4 \ln PCGDPD_{ij} + \beta_5 contig_{ij} + \beta_6 comlang_{ij} + \beta_7 SAFTA_{ij} + \beta_8 ECO_{ij} + \beta_9 INS_{jt-1} + \beta_{10}(LGEP_{jt}) + \beta_{11}(LPSP_{jt}) + \beta_{12}(LRQP_{jt}) + \beta_{13}(LRolP_{jt}) + \beta_{14}(LVAP_{jt}) + \beta_{15}(LCCP_{jt}) + \varepsilon_{ij}. \quad (5)$$

Outcome model:

$$T_{ij}=\alpha_0 +\beta_1 \ln GDP_i + \beta_2 \ln GDP_j + \beta_3 \ln Dis_{ij} + \beta_4 \ln PCGDPD_{ij} + \beta_5 contig_{ij} + \beta_6 comlang_{ij} + \beta_7 SAFTA_{ij} + \beta_8 ECO_{ij} + \beta_9 INS_{jt-1} + \beta_{10}(LGEP_{jt}) + \beta_{11}(LPSP_{jt}) + \beta_{12}(LRQP_{jt}) + \beta_{13}(LRolP_{jt}) + \beta_{14}(LVAP_{jt}) + \beta_{15}(LCCP_{jt}) + \varepsilon_{ij}. \quad (6)$$

3 RESULTS AND DISCUSSION

The gravity model estimation requires substantial preparation for data collection and organization, given that a large data set is needed for analysis. The advantage of this data set is that the estimations made are more precise and stable. However, collection of data from different sources requires much time and effort.

The extended gravity model expressed in Formula (4) is estimated in six different specifications to examine the impact of different variables on the bilateral export cooperation of India with South Asian and the Central Asian states, as presented in Table 2.

The results show that the sign and size of coefficients align with available literature. Model (1) included the average quality of institutional factors in partner countries. In the rest of the model, the impact of quality of individual institutional factors is examined. It is important to mention here that institutional factors at the level are included, given that changes in any of these factors impact trade after a time gap.

From the available literature, it is clear that the GDP of reporting country indicates supply capacity, and GDP of partner country shows demand capacity. The coefficient of GDP in both reporting and partner country is statistically significant and have a positive sign. The results indicate that 1 percent increase in GDP of reporting country increases trade by 0.81 percent. Similarly, 1 percent increase in the GDP of partner country enhances trade by 0.83 percent. The higher coefficient of reporting country compared

to the partner country confirms the findings of Feenstra et al. (2001). The coefficient of distance is significant and has expected negative sign along the lines of available literature. The distance variable confirms classical gravity model (Anderson and Van Wincoop, 2003). These findings contradict the thesis that "distance is dead" due to increased integration worldwide. Thus it can be concluded that despite the rapid pace of globalization, distance has maintained its significance as an explanatory variable of the gravity model.

Table 2 PPML with time-fixed effect

Variable	–1	–2	–3	–4	–5	–6	–7
LnGDPi	0.818***	0.757***	0.784***	0.752***	0.760***	0.806***	0.759***
	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)
LnGDPj	0.833***	0.810***	0.875***	0.845***	0.806***	0.818***	0.848***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
PCGDPD	–0.020	0.210***	0.167***	0.234***	0.159***	0.004	0.207***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
LnDISij	–1.042***	–1.070***	–1.033***	–1.122***	–1.066***	–1.096***	–1.053***
	(0.06)	(0.06)	(0.06)	(0.05)	(0.06)	(0.06)	(0.06)
SAFTAij	–2.456***	–1.852***	–1.492***	–1.918***	–2.216***	–2.674***	–1.967***
	(0.20)	(0.20)	(0.21)	(0.19)	(0.20)	(0.21)	(0.19)
ECOij	–1.111***	–0.529***	–0.719***	–0.583***	–0.761***	–1.148***	–0.576***
	(0.13)	(0.13)	(0.14)	(0.14)	(0.13)	(0.13)	(0.13)
Contig	0.289***	0.675***	0.552***	0.616***	0.665***	0.379***	0.722***
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
Comlang	1.292***	0.873***	1.055***	0.810***	0.946***	1.227***	0.903***
	(0.07)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)
INSTjt-1	0.647*						
	(0.35)						
LGEP		0.836***					
		(0.04)					
LPSP			0.743***				
			(0.04)				
LRQP				0.788***			
				(0.04)			
LRolP					0.628***		
					(0.04)		
LVAP						0.169***	
						(0.03)	
LCCP							0.687***
							(0.03)
Constant	–28.98***	–25.99***	–28.36***	–26.25***	–25.79***	–26.78***	–27.06***
	(1.13)	(1.14)	(1.11)	(1.05)	(1.04)	(1.03)	(1.09)
Observations	10 499	10 079	10 079	10 079	10 079	10 079	10 079
R-squared	0.574	0.685	0.665	0.702	0.672	0.593	0.683

Note: Standard errors in parentheses * p<.10, ** p<.05, *** p<.01.
Source: Author's calculation

The "GDP per capita income difference" (PCGDPD) is used to study the comparative existence of Heckscher-Ohlin (H-O) theory with reference to Linder hypothesis. According to H-O theory, countries trade with each other is based on factor endowment, which leads to comparative advantage in respective countries. The difference in factor endowment in different countries leads to more intensive inter-industry trade. However, at present, even countries with same factor endowment trade with each other, which leads to intra-industry trade. The estimated coefficient of PCGDPD is significant and has a positive sign which indicates higher volume of bilateral trade between partner countries. Thus based on the sign of coefficient, H-O hypothesis predominates the Linder hypothesis, which suggests that countries with different factor endowments have a higher inter-industry trade flow than intra-industry trade (Krugman, 1981).

Following Silva and Tenreyro (2006), effect of change in variable X on variable Y is calculated by $\{\left(e^{\propto}-1\right)\times 100\}$, where α is the coefficient of dummy variable. The common border enhances trade by $\left(e^{0.289}-1\right)\times 100=0.49$ percent compared to countries without a direct link. Similarly, the common language between countries enhances trade by $\left(e^{1.292}-1\right)\times 100=133.91$ percent compared to countries which face language barriers in bilateral trade cooperation. In addition to these variables, trade policy liberalization is expected to enhance trade between countries. For this reason, regional trade agreements are incorporated, which play a significant role in trade cooperation. The coefficient of SAFTA and ECO have negative sign and are statistically significant. This results indicates that South and Central Asian states have failed to enhance intra-regional trade cooperation.

The primary variable of interest in present study is the role of institutional setup in selected countries. Institutions play significant role in the international trade cooperation between countries. There are several dimensions that capture institutional quality. In the present study, in addition to the institutional quality index, six dimensions are included separately that capture institutional quality to avoid overlapping impacts. It is the lag of institutional variables which are included because bilateral trade slowly adjusts to institutional changes. Improvement in institutional quality is expected to impact bilateral trade cooperation positively. Regarding bilateral exports of India to these countries, the results in column (2–7) and Table (2) show that institutions play an important role. The positive sign of institutional factors indicates that improvement in quality of institutions is expected to enhance bilateral trade cooperation. Improvement in government effectiveness by 1 percent enhances exports by 0.83 percent. Similarly, the coefficient of regulatory quality indicates that 1 percent improvement in government ability to implement policies leads to enhancement of exports by 0.78 percent to partner countries. Among other variables, PS, VA, RoL and CC indicates an enhancement of bilateral exports between these countries with improvement in quality of these indicators. These variables are related to the participation of general public in government formation, freedom of expression, control of corruption, policy formulation and government credibility regarding policy implementation.

The results reveal that the impact of these variables on bilateral exports of these countries is outstanding. The results in the present study align with the findings of Bakhsh et al. (2022), who examined exports from China to Belt and Road initiative countries. Moreover, the findings indicate that trade agreements in South and Central Asia have not generated enough trade opportunities for member countries. The signs of coefficients in the Heckman model (Table 3) align with the results presented in Table 2. This model (Table 3) was run to check the robustness of the results as different econometric techniques complement each other. The signs of the variables are same both in PPML and Heckman models. Thus there is no need to discuss the results again.

Table 3 Heckman selection model results

Variable	–1	–2	–3	–4	–5	–6	–7
LnGDPi	0.947***	0.938***	0.943***	0.930***	0.952***	0.949***	0.978***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
LnGDPj	0.739***	0.685***	0.716***	0.815***	0.736***	0.696***	0.723***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
PCGDPD	0.064***	0.309***	0.298***	0.274***	0.274***	0.261***	0.117***
	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
Contig	1.217***	1.444***	1.468***	1.252***	1.382***	1.452***	1.291***
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)	(0.09)
Comlang	2.065***	1.515***	1.580***	1.838***	1.470***	1.620***	1.847***
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
LnDIS	–0.862***	–1.262***	–1.240***	–1.177***	–1.268***	–1.218***	–1.076***
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
ECOij	–0.612***	–0.454***	–0.364***	–0.519***	–0.515***	–0.411***	–0.552***
	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)
SAFTAij	–0.787***	–0.658***	–0.669***	–0.421***	–0.801***	–0.846***	–1.088***
	(0.13)	(0.12)	(0.12)	(0.13)	(0.12)	(0.12)	(0.12)
INSTjt-1	0.471***						
	(0.14)						
LGEP		0.939***					
		(0.04)					
LCCP			0.811***				
			(0.03)				
LPSP				0.703***			
				(0.04)			
LRQP					0.744***		
					(0.03)		
LRoIP						0.809***	
						(0.04)	
LVAP							0.460***
							(0.05)
Constant	–32.54***	–27.49***	–28.44***	–31.04***	–28.96***	–28.26***	–30.68***
	(0.54)	(0.57)	(0.56)	(0.54)	(0.55)	(0.57)	(0.56)
Select							
LnGDPi	0.376***	0.383***	0.384***	0.384***	0.384***	0.384***	0.384***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
LnGDPj	0.299***	0.308***	0.309***	0.308***	0.308***	0.308***	0.309***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
PCGDPD	–0.028***	–0.027***	–0.027***	–0.027***	–0.027***	–0.027***	–0.027***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Contig	0.486***	0.486***	0.482***	0.482***	0.483***	0.482***	0.477***
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
Comlang	7.305	6.925	7.264	7.097	6.991	7.074	7.261
	(333 109.48)	(66 549.81)	(253 675.07)	(123 947.02)	(84 124.11)	(118 679.46)	(237 557.10)
LnDIS	–0.330***	–0.338***	–0.335***	–0.336***	–0.337***	–0.336***	–0.335***
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Constant	–12.98***	–13.33***	–13.39***	–13.36***	–13.34***	–13.36***	–13.39***
	(0.40)	(0.41)	(0.41)	(0.41)	(0.41)	(0.41)	(0.41)
Observations	10 499	10 163	10 163	10 163	10 163	10 163	10 163

Notes: Standard errors in parentheses * p<.10, ** p<.05, *** p<.01.
Source: Author's calculation

CONCLUSION

The present study examined bilateral export determinants of India with selected South and Central Asian countries and top trade partners to determine the impact of quality of institutions on trade cooperation between these countries. The gravity model with advanced estimation techniques was applied following the available literature. In addition to basic variables, role of institutional factors at level, regional trade agreements, cultural factors and related trade costs were also incorporated. The result confirms that institutions play positive and significant role in trade integration between countries, in line with the findings of Gani and Prasad (2008).

The results discussed above indicate that various factors of institutional quality play an important role in export promotion from India to these countries. Among various factors, GE, RQ and PS have the most substantial impact, followed by CC, RoL and VA. This highlights the need to improve the quality of public and civil services and their independence from political pressure. Similarly, the ability of the government to formulate and implement policies to promote private sector development needs to be paid adequate attention to promote exports. Further, the focus should be diverted towards control of corruption, misuse of public offices for private gains, and increased participation of common masses in government formation, which is conducive to economic development and international trade. Besides, the results indicate that regional trade agreements have been ineffective to enhance trade cooperation among member countries, which points out the ineffective implementation of these agreements.

The findings of the present study highlight various areas which needs attention from the competent authority of these countries. First, an effective government role in formulation and implementation of policies is important to promote trade cooperation between India and these countries. Similarly, the participation of people in government formation and freedom of expression, association and free media needs to be promoted in these countries. Similarly, there is a need to check corruption and abuse of public office for private gains needs to be curbed in addition to the improvement in law and order situation and control of terrorist activities.

In general, the results indicate that institutional quality promotes trade between countries. The present study confirms the hypothesis regarding the role of institutions in trade cooperation between countries. Thus institution of democracy need to be strengthened to enhance economic development and trade cooperation among these countries.

ACKNOWLEDGEMENT

This paper is largely an outcome of the Post-Doctoral Fellowship sponsored by the Indian Council of Social Science Research (ICSSR). However, the responsibility for the facts stated, opinions expressed, and the conclusions drawn are entirely of the authors.

References

ABBAS, S., WAHEED, A. (2019). Pakistan's global trade potential: A gravity model approach [online]. *Global Business Review*, 20(6): 1361–1371. <https://doi.org/10.1177/0972150919848936>.

ABDULLAHI, N. M., SHAHRIAR, S., KEA, S., ABDULLAHI, A. M., ZHANG, Q., HUO, X. (2021). Nigeria's cocoa exports: a gravity model approach [online]. *Ciência Rural*, 51. <https://doi.org/10.1590/0103-8478cr20201043>.

ACHROL, R., REVE, T., STERN, W. (1983). The Environment of Marketing Channel Dyads: a framework for comparative analysis [online]. *Journal of Marketing*, 47(4): 55–67. <https://doi.org/10.1177/002224298304700407>.

AGRAWAL, P., SANGITA, S. (2013). *India and Central Asia: Trade Routes and Trade Potential*. IEG working paper No. 334.

AKHTER, N., GHANI, E. (2010). Regional integration in South Asia: An analysis of trade flows using the gravity model [online]. *The Pakistan Development Review*, 105–118. <https://www.jstor.org/stable/41263356>.

ALUKO, O. A., IBRAHIM, M., VO, X. V. (2021). On the foreign direct investment–economic growth relationship in Africa: does economic freedom mediate this relationship? [online]. *International Journal of Emerging Markets*. <https://doi.org/10.1108/IJOEM-04-2021-0568>.

ANDERSON, J. E., MARCOUILLER, D. (2002). Insecurity and the pattern of trade: An empirical investigation [online]. *Review of Economics and statistics*, 84(2): 342–352. <https://doi.org/10.1162/003465302317411587>.

ANDERSON, J. E., VAN WINCOOP, E. (2003). Gravity with gravitas: A solution to the border puzzle [online]. *American economic review*, 93(1): 170–192. <https://doi.org/10.1257/000282803321455214>.

BAIER, S. L., YOTOV, Y. V., ZYLKIN, T. (2019). On the widely differing effects of free trade agreements: Lessons from twenty years of trade integration [online]. *Journal of International Economics*, 116: 206–226. <https://doi.org/10.1016/j.jinteco.2018.11.002>.

BAKHSH, S., YIN, H., SHABIR, M., ALI, K. (2022). China trade with belt and road countries: the role and impact of institutions [online]. *China Economic Journal*, 15(1): 29–48. <https://doi.org/10.1080/17538963.2021.1994709>.

BALOOCH, M. (2009). Iran and India's cooperation in Central Asia. *China and Eurasia Forum Quarterly*, 7(3): 25–29.

BAYSAN, T., PANAGARIYA, A., PITIGALA, N. (2006). Preferential Trading in South Asia. *World Bank Policy Research Working Paper,* 3813.

BERTHELON, M., FREUND, C. (2004). On the Conservation of Distance in International Trade. *World Bank Policy Research Department Working Paper*, 3293.

BOWN, C. (2017). *The Slow Unraveling of US Trade Policy in the Age of Trump*. International Agricultural Trade Research Consortium, No. 266842.

CHATTERJEE, S. (2018). International North South Transport Corridor (INSTC) connecting India Central–Asia. *International Journal of Research in Social Sciences*, 8(4): 601–616.

CHI, T., KILDUFF, P. P. (2010). An empirical investigation of the determinants and shifting patterns of US apparel imports using a gravity model framework [online]. *Journal of Fashion Marketing and Management: An International Journal*, 14(3): 501–520. <https://doi.org/10.1108/13612021011061915>.

DEARDOFF, A. (1998). Determinants of Bilateral Trade: Does Gravity Model Work in a Classical World? In: FRANKEL, J. (eds.) *The Regionalization of the World Economy,* Chicago University Press.

DE JONG, E., BOGMANS, C. (2011). Does corruption discourage international trade? [online]. *European Journal of Political Economy*, 27(2): 385–398. <https://doi.org/10.1016/j.ejpoleco.2010.11.005>.

DE MENDONÇA, T. G., LIRIO, V. S., BRAGA, M. J., DA SILVA, O. M. (2014). Institutions and bilateral agricultural trade [online]. *Procedia Economics and Finance*, 14: 164–172. <https://doi.org/10.1016/S2212-5671(14)00699-6>.

DEROSA, D. A., GOVINDAN, K. (1996). Agriculture, trade, and regionalism in South Asia [online]. *Journal of Asian Economics*, 7(2): 293–315. <https://doi.org/10.1016/S1049-0078(96)90009-3>.

DINCER, G. (2014). *Turkey's rising imports from BRICS: A Gravity model approach* [online]. <https://mpra.ub.uni-muenchen.de/61979/ MPRA Paper No. 61979>.

FEENSTRA, R. C., MARKUSEN, J. R., ROSE, A. K. (2001). Using the gravity equation to differentiate among alternative theories of trade [online]. *Canadian Journal of Economics/Revue canadienne d'économique*, 34(2): 430–447. <https://doi.org/10.1111/0008-4085.00082>.

FRANCOIS, J., MANCHIN, M. (2013). Institutions, infrastructure, and trade [online]. *World development*, 46: 165–175. <https://doi.org/10.1016/j.worlddev.2013.02.009>.

GOLOVKO, A., SAHIN, H. (2021). Analysis of international trade integration of EurAsian countries: Gravity model approach [online]. *EurAsian Economic Review*, 11(3): 519–548. <https://doi.org/10.1007/s40822-021-00168-3>.

GUAN, Z., ZHANG, Y., KWEE FAT, I. P. S. J. (2021). Trade Relations Between Mauritius and China: A Gravity Model Approach [online]. *SAGE Open*, 11(4). <https://doi.org/10.1177/21582440211058184>.

HANDLEY, K., LIMAO, N. (2017). Policy Uncertainty, Trade, and Welfare: Theory and Evidence for China and the United States [online]. *American Economic Review,* 107(9): 2731–2783. <https://doi.org/10.1257/aer.20141419>.

HASSAN, M. K. (2001). Is SAARC a viable economic block? Evidence from gravity model. *Journal of Asian Economics*, 12(2): 263–290.

HEAD, K. (2003). *Gravity for beginners.* University of British Columbia, 2053.

HOFSTEDE, G. (2016). Culture's consequences: Comparing values, behaviors, institutions, and organisations across nations. *Collegiate Aviation Review*, 34(2): 108.

IQBAL, N., NAWAZ, S. (2017). Pakistan's bilateral trade under MFN and SAFTA: Do institutional and non-institutional arrangements matter? [online]. *The Pakistan Development Review*, 59–78. <https://www.jstor.org/stable/44986516>.

JAN, W. U., SHAH, M. (2019). A gravity model approach towards Pakistan's bilateral trade with SAARC countries. *Comparative Economic Research. Central and Eastern Europe*, 22(4): 23–38.

JANSEN, M., KYVIK NORDAS, H. (2004). Institutions, trade policy and trade flows. *Staff Working Paper ERSD*, Geneva: World Trade Organization, 2004-02.

KAMAL, Y., ZAKI, C. (2018). How Do Technical Barriers to Trade Affect Exports? Evidence from Egyptian Firm-Level Data [online]. *Journal of Economic Integration,* 33(4): 659–721. <https://doi.org/ doi:10.11130/jei.2018.33.4.659>.

KATHURIA, S., SHAHID, S. (2017). Boosting trade and prosperity in South Asia. In: DE, P., REHMAN, M. (eds.) *Regional Integration in South Asia: Essays in Honour of Dr. M. Rahmatullah*, 1st Ed., New Delhi: KW Publishers Pvt. Limited, 7–34.

KEA, S., LI, H., SHAHRIAR, S., ABDULLAHI, N. M., PHOAK, S., TOUCH, T. (2019). Factors influencing Cambodian rice exports: An application of the dynamic panel gravity model [online]. *Emerging Markets Finance and Trade*, 55(15): 3631–3652. <https://doi.org/10.1080/1540496X.2019.1673724>.

KRUGMAN, P. R. (1981). Intraindustry specialisation and the gains from trade [online]. *Journal of political Economy*, 89(5): 959–973. <https://www.journals.uchicago.edu/doi/abs/10.1086/261015>.

KUMAR, S., AHMED, S. (2015). Gravity model by panel data approach: An empirical application with implications for South Asian countries [online]. *Foreign Trade Review*, 50(4): 233–249. <https://doi.org/10.1177/0015732515598587>.

MALIK, I. A., MIR, M. A. (2014). India's trade potential with Central Asia: an application of gravity model analysis. *International Journal on World Peace*, 31(3): 53.

MATYAS, L. (1997). Proper econometric specification of the gravity model [online]. *World Economy*, 20(3): 363–368. <https://doi.org/10.1111/1467-9701.00074>.

MATYAS, L. (1998). The gravity model: Some econometric considerations [online]. *World Economy*, 21(3): 397–401. <https://doi.org/10.1111/1467-9701.00136>.

NAWAZ, S. (2020). Institutions, Regional Integration and Bilateral Trade in South Asia: PPML Based Evidence. *The Pakistan Development Review*, 59(2): 221–242.

PARAKH, Y., ADITYA, A. (2021). Analysing the Determinants of Trade Agreements: A Cross Country Socio-Economic-Political Analysis [online]. *The International Trade Journal*, 1–27. <https://doi.org/10.1080/08853908.2021.2007180>.

PIGATO, M. (1997). *South Asia's integration into the world economy*. World Bank Publications.

RAHMAN, M., SHADAT, W. B., DAS, N. C. (2006). Trade potential in SAFTA: An application of augmented gravity model. *CPD occasional paper series*, 61: 512.

RAYCHAUDHURI, A., HALDAR, S. K. (2009). An investigation into the inter-district disparity in West Bengal, 1991–2005. *Economic and political weekly*, 258–263.

REGMI, R. K., DEVKOTA, S. C., UPADHYAY, M. P. (2017). Impact of SAFTA on South Asian trade [online]. *Asian Economic and Financial Review*, 7(3): 232. <https://doi.org/10.18488/journal.aefr/2017.7.3/102.3.232.247>.

RODRIK, D. (2000). Institutions for high-quality growth: what they are and how to acquire them [online]. *Studies in comparative international development*, 35: 3–31. <https://doi.org/10.1007/BF02699764>.

ROY, M. S. (2013). India's' Connect Central Asia' Policy: Building Cooperative Partnership. *Indian Foreign Affairs Journal*, 8(3): 301.

SHAHRIAR, S., KEA, S., ABDULLAHI, N. M., RAHMAN, R., ISLAM, R. M. (2021). Determinants of Bangladesh's Leather Exports to Its Major Trade Partners: A Panel Gravity Model Approach. *Global Business Review*, 09721509211036288.

SHAIKH, F. M., RAHPOTO, M. S. (2009). Impact of trade liberalisation and SAFTA on Pakistan's Economy by Using CGE Model. *International Journal of Business and Management*, 4(4): 192.

SHAIKH, F. M., SYED, A. A. S. G., SHAH, H., SHAH, A. A. (2012). Observing Impact of SAFTA on Pakistan's Economy by Using CGE Model. *Pakistan Journal of Commerce & Social Sciences*, 6(1).

SHARMA, V., KUMAR, V. (2021). Trade potential under the SAFTA between India and other SAARC countries: The augmented gravity model approach. *Statistics in Transition New Series*, 22(3).

SILVA, J. S., TENREYRO, S. (2006). The log of gravity. *The Review of Economics and statistics*, 88(4): 641–658.

SKARMEAS, D., KATSIKEAS, C. K., SCHLEGELMILCH, B. (2002). Drivers of Commitment and its Impact on Performance in Cross-Cultural Buyer-Seller Relationships: The Importer's Perspective [online]. *Journal of International Business Studies*, 33(4): 757–783.

SOLOAGA, I., WINTERSB, L. A. (2001). Regionalism in the nineties: What effect on Trade? *The North American Journal of Economics and Finance*, 12(1): 1–29.

SULTANA, S., ASRAT, J. (2014). South Asian Countries in Regional Integration Perspective: A Critical Review. *Journal of Business and Technology* (Dhaka), 9(2): 43–59.

SUVANKULOV, F., GUC, Y. (2012). Who is trading well in Central Asia? A gravity analysis of exports from the regional powers to the region. *EurAsian Journal of Business and Economics*, 5(9): 21–43.

TREFLER, D. (1995). The case of missing trade and other mysteries. *American Economic Review*, 85: 1029–1046.

WANI, S. H., MIR, M. A. (2023). Import Determinants and Potential Markets of Pakistan: An Application of PPML Gravity Model [online]. *Studia Universitatis Vasile Goldiș Arad, Seria Științe Economice*, 33(1): 57–73. <https://doi.org/10.2478/sues-2023-0003>.

WEI, S. J. (2000). Natural openness and good government. *NBER Working Paper*, 7765.

YUSUF, H. A., AFOLABI, L. O., SHITTU, W. O., GOLD, K. L., MUHAMMAD, M. (2021). Institutional quality and trade flow: Empirical evidence from Malaysia and other OIC member Countries in Africa [online]. *Insight on Africa*, 13(2): 177–191. <https://doi.org/10.1177/0975087820987174>.

ZHANG, C. et al. (2003). Manufacturing Governance of Foreign Distributor Relationships: Do relational norms enhance competitiveness in the export market? *Journal of International Business Studies*, 34: 550–566.

Impact of Tobacco Taxation on Rural-Urban Cigarette Consumption in Indonesia

Imade Yoga Prasada[1] | *Universitas Putra Bangsa, Kebumen, Indonesia*
Aura Dhamira[2] | *Universitas Pembangunan Nasional Veteran Yogyakarta, Yogyakarta, Indonesia*
Dinda Dewi Aisyah[3] | *Universitas Pembangunan Nasional Veteran Yogyakarta, Yogyakarta, Indonesia*
Ayu Putri Merry Anisya[4] | *Universitas Putra Bangsa, Kebumen, Indonesia*
Rizqika Puspajanati[5] | *Universitas Putra Bangsa, Kebumen, Indonesia*

Received 1.7.2023, **Accepted** (reviewed) 15.9.2023, **Published** 15.3.2024

Abstract

Tobacco taxes in Indonesia have always been increasing over the past period. However, little was known about the impact of increasing tobacco taxes on cigarette consumption in urban and rural communities. Therefore, we examine the impact of implementing tobacco taxation policies in Indonesia on cigarette consumption levels in urban and rural areas. This study uses panel data for the 2007–2022 period from 33 provinces in Indonesia. Data were sourced from the Central Bureau of Statistics of the Republic of Indonesia and the Ministry of Finance of the Republic of Indonesia and were analyzed using the two-stage least squares (2SLS) model and the propensity score matching (PSM) model. Our finding shows that increased tobacco taxes negatively impact cigarette consumption in urban and rural areas. Even so, the decline in cigarette consumption in urban areas is still much lower than the reduction in rural areas.

Keywords	***DOI***	***JEL code***
Tobacco taxation, cigarette consumption, rural, urban, Indonesia	*https://doi.org/10.54694/stat.2023.29*	*C33, E20, E62*

[1] Faculty of Science and Technology, Universitas Putra Bangsa, Kebumen 54361, Ronggowarsito Street 18, Sudagaran, Kedawung, Pejagoan District, Kebumen Regency, Central Java, Indonesia. Corresponding author: e-mail: imade.yogap@gmail.com. ORCID: <https://orcid.org/0000-0002-4878-5965>.
[2] Agribusiness Department, Faculty of Agriculture, Universitas Pembangunan Nasional Veteran Yogyakarta, Yogyakarta 55283, Ring Road Utara 104, Ngropoh, Condongcatur, Kec. Depok, Sleman Regency, Yogyakarta Special Region, Indonesia. E-mail: dhamiraaura@gmail.com.
[3] Agribusiness Department, Faculty of Agriculture, Universitas Pembangunan Nasional Veteran Yogyakarta, Yogyakarta 55283, Ring Road Utara 104, Ngropoh, Condongcatur, Kec. Depok, Sleman Regency, Yogyakarta Special Region, Indonesia. E-mail: dinda.dewiaisyah@upnyk.ac.id.
[4] Faculty of Science and Technology, Universitas Putra Bangsa, Kebumen 54361, Ronggowarsito Street 18, Sudagaran, Kedawung, Pejagoan District, Kebumen Regency, Central Java, Indonesia. E-mail: merryanisya@gmail.com.
[5] Faculty of Science and Technology, Universitas Putra Bangsa, Kebumen 54361, Ronggowarsito Street 18, Sudagaran, Kedawung, Pejagoan District, Kebumen Regency, Central Java, Indonesia. E-mail: rizqikapuspa04@gmail.com.

INTRODUCTION

Tobacco is a commodity that has a strategic role in the economy of Indonesia. This can be seen from the value of state revenues originating from tobacco taxation which tends to grow rapidly. The Indonesian Central Bureau of Statistics shows that state revenues from tobacco taxes in 2022 will touch IDR 224 200 billion or the equivalent of 11.65% of Indonesia's total tax revenues in 2022 (Central Bureau of Statistics, 2023c). The tobacco taxes also contributed around 1.14% of Indonesia's total GDP in 2022. This figure is higher than state revenue from tobacco taxes in 2021 and 2020, which only touched IDR 195 518 billion and IDR 176 309 billion, respectively. In addition, tobacco production in Indonesia is quite high, with the average annual production reaching 256.03 thousand tons. Tobacco produced in Indonesia is mostly used by the tobacco industry and eventually processed into several types of cigarettes, namely clove, and white.

The high production of tobacco and cigarettes in Indonesia has pushed Indonesia to become one of the most cigarette-exporting countries in the world. Indonesia's cigarette commodity export market share (HS code: 2402) in the international market reached 3.60%, slightly lower than the market share of the Dominican Republic, ranked the fourth largest cigarette exporting country globally, which touched 4.40%. (International Trade Centre, 2023). Nonetheless, many studies show that consumption of tobacco in the form of cigarettes can harm public health, which can trigger lung cancer (Singh and Kathiresan, 2015; Zhang et al., 2022), chronic obstructive pulmonary disease (Dai et al., 2022), ischemic heart disease (Holipah, Sulistomo and Maharani, 2020; Lim et al., 2017), and can even increase the risk of mental health disorders (Milic et al., 2020; Taylor and Treur, 2023). Therefore, the marketing of tobacco products in Indonesia is limited and regulated through the Regulation of the Ministry of Finance of the Republic of Indonesia concerning Taxation for Tobacco Products.

It was recorded that from 2007 to 2022, the Ministry of Finance of the Republic of Indonesia issued five modifications to the tobacco taxation policy. Both of the policies are implementing the increase in tobacco taxation. The increase in tobacco taxes alone has had a mixed impact on various sectors of the economy in Indonesia. Based on data from the Central Bureau of Statistics of the Republic of Indonesia shows that the increase in tobacco taxes has led to a decrease in the number of tobacco product industries in Indonesia. The number of tobacco product industries in Indonesia has decreased by an average of 5.23% per year during the 2007–2022 period (Central Bureau of Statistics, 2023b). Hence, the increase in tobacco taxes has reduced labor absorption in the tobacco production and processing sector (Nguyen, Giang and Pham, 2020).

Another study has provided finding about the increasing tobacco tax impacts that reduce tobacco prices at the farm level due to the low demand for tobacco by the tobacco industry (Suprihanti, Harianto, Sinaga and Kustiari, 2018). On the other hand, the price of processed tobacco products such as cigarettes at the consumer level tends to increase due to increased tobacco taxes (Ho, Schafferer, Lee, Yeh and Hsieh, 2018; Zhao, 2022). Several studies have shown that increasing tobacco taxes encourages sustainable and equitable economic growth (Bardach et al., 2022; Bella et al., 2023).

Implementing fiscal policies, especially tobacco taxation policies, may impact rural and urban areas differently. Van den Boogaard and Beach (2023) concluded that implementing tax policies in rural areas tended to be inefficient compared to urban areas due to low tax revenues and high tax collection costs by the government. In particular, people in rural and urban areas have different tax preferences. Rural communities focus more on taxation policies related to their assets, while urban communities pay more attention to taxation policies related to consumption activities (Andersson, 2018). Indonesia itself has both the regional characteristics, rural and urban areas. It is recorded that Indonesia has 98 urban areas and 416 rural areas to date (Central Bureau of Statistics, 2023a). Rural areas are defined as areas dominated by community structures working in the agricultural sector with a relatively low population density (Putri, Russell, O'Sullivan, Meliala and Kippen, 2022), while urban areas are dominated by people working in industrial sectors with high population density (Wang, Ma, Sun and Zhang, 2021).

Previous studies have demonstrated the effects of increasing tobacco taxes in various case studies. However, there were needs for the research linking the impact of the increase in tobacco taxes on cigarette consumption in rural and urban communities. Furthermore, rural and urban areas are known to have many differences, including differences in topography, population density, income level, poverty level, and so on, where these differences have the potential to produce different responses to cigarette consumption when changes in tobacco taxation are implemented. Hence, this research examined the impact of tobacco tax policies on cigarette consumption in rural-urban communities in Indonesia.

1 LITERATURE REVIEW

Many previous studies have proven that the policy of increasing tobacco taxation provided various direct effects on the tobacco product industry and cigarette consumers, including reducing cigarette consumption in society. The results of a recent study conducted by Boachie et al. (2022) show that increasing tobacco taxation and health education can reduce cigarette consumption in the community. In other previous studies, Cheng and Estrada (2020) specifically calculated changes in the elasticity of demand for cigarettes that are affected by increases in cigarette prices. The results can be confirmed that with every increase in the price of cigarettes by 10.00%, the overall demand decreases by 5.60%. In the long term, Friedson, Li, Meckel, Rees and Sacks (2023) proved that the impact of increasing tobacco tax policies would reduce the death rate caused by diseases suffered by smokers. In adolescent smokers, it was found that a $1.00 increase in tobacco taxes was associated with an 8.00% decrease in cigarette consumption as adolescents grew older.

Our analysis begins by exploring previous studies that concluded cultural differences and social interactions in rural and urban communities as the determining factors for cigarette use. Roberts, Teferra, Keller-Hamilton, Patterson and Ferketich (2020) looked into the profile of rural cigarette consumers associated with the presence of male family members who smoked in the household. This indicates that masculinity and smoking culture can spread between generations. Disparities in cigarette consumption that occur in different sociodemographics based on race, ethnicity, poverty status and sexual orientation in the research conducted by Golden, Kong, Lee and Ribisl (2018) are implicated in a recommendation for imposing taxes by adjusting groups that are vulnerable to smoking consumption. These results are supported by the findings of Vallarta-Robledo et al. (2022), who stated that spatially, the environment was associated with cigarette consumption, and behavioral change was associated with smoking independently. Therefore, the first hypothesis to be developed is as follows:

Hypothesis 1: Tobacco taxation in Indonesia can reduce the level of consumption of cigarettes in rural and urban areas.

Despite the fact that several previous studies have shown that the level of cigarette consumption is not only influenced by tobacco taxation but also by other interrelated factors. Aadahl et al. (2021) concluded that a higher unemployment rate could encourage a person to consume more cigarettes. This increase was attributed to the growing volume of leisure time which could be used for smoking (Verkooijen, Nielsenand and Kremers, 2009). In addition, Crespo Cuaresma, Kubala and Petrikova (2018) stated that the increase in cigarette consumption by the public was also triggered by growing income disparities in an area. Another factor that can affect the level of cigarette consumption is the real wage of workers, where an increase in workers' real wages can cause the amount of cigarette consumption to increase (Huang, Liu and You, 2021). Furthermore, the number of incidents of violence that occurred among adolescents and children also encouraged increased cigarette consumption (Kleppang and Skille, 2022; Ouyang et al., 2020). Therefore, we develop the second hypothesis as follows:

Hypothesis 2: Cigarette consumption in rural and urban areas is influenced by unemployment rates, workers' wages, income disparities, and the amount of physical and mental violence that occurs in rural and urban areas.

At the same time, the unemployment rate is simultaneously influenced by various macroeconomic variables. Siregar (2020) shows that an increase in the minimum wage can lead to an increase in the number of the unemployed due to decreased labour demand. In addition, the number of the unemployed is also influenced by economic growth. Economic growth can encourage more opportunities to create new jobs to reduce the number of unemployed (Mushtaq, Ahmed, Fahlevi, Aljuaid and Saniuk, 2022). Furthermore, a person's education level also influences the unemployment rate. The higher the level of education, the greater the attractiveness of becoming a worker compared to someone with a low education level (Tamvada, Shrivastava and Mishra, 2022). Hailu Demeke (2022) also emphasizes that someone uneducated tends to have limitations in obtaining opportunities to get a good job. Based on these studies, we develop a third hypothesis, namely:

Hypothesis 3: Unemployment rates in urban and rural areas are influenced by the provincial minimum wage, economic growth rate, education level, and the percentage of illiterate population.

2 RESEARCH METHOD

2.1 Data source and variables

This study aims to see the impact of tobacco taxation on cigarette consumption in rural and urban communities, and several stages of study have to be completed before evaluating the impact of this policy. We chose our variables based on previous research. Then, we collect annual time series data from the Central Bureau of Statistics of the Republic of Indonesia and the Ministry of Finance of the Republic of Indonesia covering 2007–2021. We also use cross-sectional data from 33 provinces in Indonesia, then the data is organized into panel data.The data used such as real consumption of cigarettes in urban

Table 1 Variable's information

Variables	Description	Measurement	Source
URB	Real cigarette consumption in urban communities	IDR/month/capita	Central Bureau of Statistics of the Republic of Indonesia
RUR	Real cigarette consumption in rural communities	IDR/month/capita	Central Bureau of Statistics of the Republic of Indonesia
UNE	Open unemployment rate	Percent	Central Bureau of Statistics of the Republic of Indonesia
MIN	Provincial minimum wage	IDR	Central Bureau of Statistics of the Republic of Indonesia
GRO	Economic growth	Percent	Central Bureau of Statistics of the Republic of Indonesia
ENR	Pure enrolment rate for high school education	Percent	Central Bureau of Statistics of the Republic of Indonesia
ILL	Percentage of illiterate population	Percent	Central Bureau of Statistics of the Republic of Indonesia
WAG	The average real wage of workers	IDR/hour	Central Bureau of Statistics of the Republic of Indonesia
INE	Gini index	Index	Central Bureau of Statistics of the Republic of Indonesia
VIO	Percentage of population who are victims of violence	Percent	Central Bureau of Statistics of the Republic of Indonesia
R08	Dummy tobacco taxation policy in 2008	1=period of increase in tobacco taxation rates in 2008, 0=others	Ministry of Finance of the Republic of Indonesia
R09	Dummy tobacco taxation policy in 2009	1=period of increase in tobacco taxation rates in 2009, 0=others	Ministry of Finance of the Republic of Indonesia
R12	Dummy tobacco taxation policy in 2012	1=period of increase in tobacco taxation rates in 2012, 0=others	Ministry of Finance of the Republic of Indonesia
R17	Dummy tobacco taxation policy in 2017	1=period of increase in tobacco taxation rates in 2017, 0=others	Ministry of Finance of the Republic of Indonesia

Source: Authors

and rural communities, provincial minimum wage, economic growth, percentage of illiterate population, Gini index, and dummy variables of tobacco taxation policy in 2008, 2009, 2012 and 2017 to see whether there are changes in cigarette consumption after the implementation of the tariff (Table 1).

2.2 Test for Unit Root

The data was then analyzed using Levin Lin and Chu (LLC) test to determine the stationarity level (Levin, Lin and Chu, 2002). Thereafter, using a two-stage least squares (2SLS) model, the stationary variables are analyzed. A unit root test was conducted to determine the level of stationarity of the variables used in this study. Stationarity indicates no change in the data's statistic properties over time. The stationarity test shows that URB, UNE, MIN, GRO, ENR, ILL, INE, and VIO are stationary at level (Table 2). Meanwhile, RUR and WAG are stationary at the first-difference level.

Table 2 Stationarity test

Variable	Urban			Rural		
	Stage	LLC statistic	Prob.	Stage	LLC statistic	Prob.
URB	Level	–2.04	0.02	–	–	–
RUR	–	–	–	1st difference	–12.09	0.00
UNE	Level	–9.11	0.00	Level	–9.17	0.00
MIN	Level	–2.21	0.01	Level	–2.31	0.01
GRO	Level	–6.52	0.00	Level	–6.38	0.00
ENR	Level	–6.61	0.00	Level	–6.55	0.00
ILL	Level	–2.60	0.00	Level	–2.51	0.01
WAG	1st difference	–11.92	0.00	1st difference	–12.22	0.00
INE	Level	–7.43	0.00	Level	–7.67	0.00
VIO	Level	–3.80	0.00	Level	–3.75	0.00

Note: H0: Panels contain unit roots. Ha: Panels are stationary.
Source: Authors analysis, 2023

2.3 Data analysis

2.3.1 Determinants of tobacco consumption and unemployment in urban and rural area

The two-stage least squares (2SLS) model was conducted after the data became stationary. Endogenous variables are often correlated with random variables, and it is possible to solve this problem in each equation using 2SLS (López-Espín, Vidal and Giménez, 2012). This model has two stages, the first stage is to see the factors impacting the unemployment rate, while the second and final stages see the factors affecting tobacco consumption in urban or rural communities. The (1) and (2) Formulas are estimated simultaneously using the 2SLS model. The equation is as follows.

First stage:

$$UNE_t^i = \gamma_0 + \gamma_1 MIN_t^i + \gamma_2 GRO_t^i + \gamma_3 ENR_t^i + \gamma_4 ILL_t^i + \mu\,. \tag{1}$$

Second stage:

$$URB_t^i \; or \; RUR_t^i = \beta_0 + \beta_1 UNE_t^i + \beta_2 WAG_t^i + \beta_3 INE_t^i + \beta_4 VIO_t^i + \beta_5 R08_t^i + \beta_6 R09_t^i + \beta_7 R12_t^i + \beta_8 R17_t^i + \mu\,. \tag{2}$$

Expected estimation mark $\beta_1, \beta_2, \gamma_1, \gamma_4 > 0; \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \gamma_2, \gamma_3 < 0,$

where: *i* represents the cross-section data used in this study that refers to the data from rural and urban areas, while *t* shows the year of observation. The time lag variable was not used in this model. In this research, the time lag variable is not used in the model because it's only used when the policy being implemented is deemed to require a certain amount of time before it can be actively implemented properly. On the other hand, tobacco taxation has been implemented in Indonesia for a long time, so changes in the value of tobacco taxes do not require time lag to be implemented properly (Bella et al., 2023).

At the first stage, the independent variables are provincial minimum wage, economic growth, enrolment rate for high school education, and the percentage of the illiterate population. Meanwhile, the variables to determine impacting factors on tobacco consumption in urban and rural areas are unemployment, the real wage average, Gini index, the percentage of violence victims, and the dummy variables consisting of the implementation of tobacco taxations in 2008, 2009, 2012, and 2017.

2.3.2 Impact evaluation of tobacco taxations on cigarette

The 2SLS model can only determine the difference before and after the policy was implemented in 2008, 2009, 2012, and 2017, but it cannot show how big the difference is. The PSM method is one of the several methods used to determine the impact of Tobacco taxation on cigarette consumption. The propensity score matching (PSM) method is a commonly used method to deal with the problem of sample selection bias (Li, Han and Zhu, 2023). The idea is to balance and make the two situations, before and after the implementation of the tariff policy, comparable. There are several stages, first is to identify the two groups, control and treatment, followed by identifying which outcome would be measured in this study. The last stage was matching the two groups to see the impact of tariffs implementation (Kuss, Blettner, & Börgermann, 2016). The propensity score matching can be calculated through this formula:

$$ATT = E(R_1 \mid I = 1) - E(R_0 \mid I = 0),$$

$$ATT = E\left\{R_1 | I = 1, p\left(Z\right)\right\} - E\{R_0 \mid I = 0, p\left(Z\right)\},$$

here: ATT is the Average Treatment effect of the treated group, which is the impact of implementing the policy, the I symbol shows the indicators of tobacco taxations implementation (I = 0 control, I = 1 treatment), R_0 and R_1 indicate the outcome value of the treatment and control data. Lastly, p(Z) is the propensity score obtained from the dummy variables from each year of the tariff implementation (2008, 2009, 2012 and 2017). The propensity score matching was conducted twice, the first for the urban area and the second for the rural area. After that, we compare the results from the two areas. A balance test is needed to evaluate the conditional independence assumptions. On the other hand, we must also fulfill the overlapping assumptions through the PSM chart to interpret the PSM results properly. Most of the total bias reduction almost reached 100% which means that the impact evaluation result is not biased (Table 3).

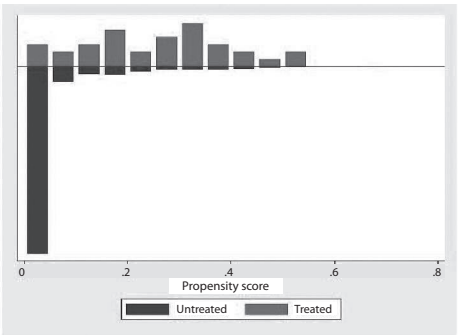
The propensity score distribution shows that most of the data can be matched perfectly except for the distribution of tobacco taxation propensity scores in 2012 (R12) in urban and rural areas that was off support even though it was not dominant, so we can conclude that the matching quality was well maintained in each period of the tobacco taxation implementation (Figure 1).

Table 3 Balancing test for matching based on the propensity score

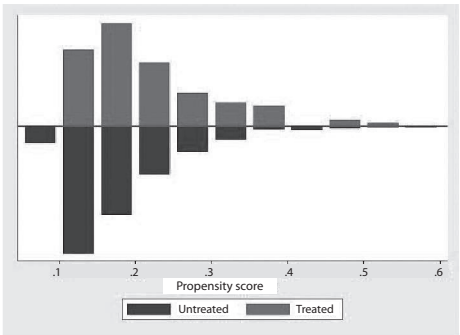
Tobacco excise tariffs period	Pseudo R2 before matching	Pseudo R2 after matching	LR chi2 before matching	LR chi2 after matching	Mean standardized Bias before matching	Mean standardized Bias after matching	Total % |Bias| reduction
Urban							
R08	0.34	0.01	84.64	1.23	87.70	2.10	97.70
R09	0.04	0.00	19.15	0.00	49.50	0.10	99.80
R12	0.11	0.00	69.14	0.20	37.30	4.20	88.90
R17	0.11	0.02	54.45	5.75	66.50	2.00	97.10
Rural							
R08	0.09	0.00	21.63	0.00	89.00	1.50	98.30
R09	0.04	0.00	17.79	0.00	48.60	0.20	99.70
R12	0.04	0.00	22.96	0.16	39.70	2.20	94.50
R17	0.11	0.01	51.96	2.22	68.80	5.10	92.60

Source: Authors analysis, 2023

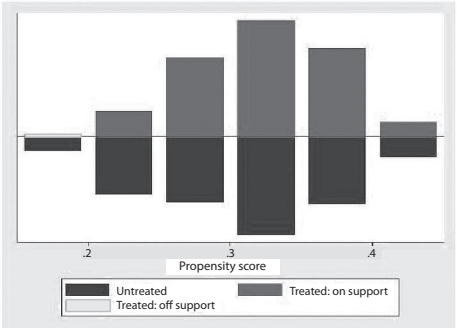
Figure 1 Distribution of tobacco taxation propensity scores



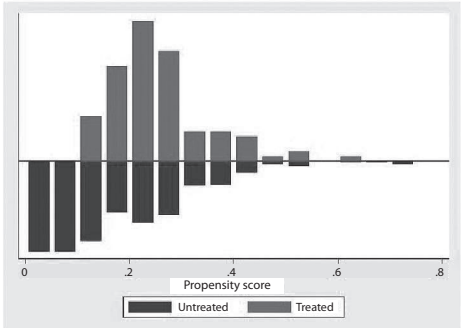
(a) Distribution of tobacco taxation propensity scores in 2008 (R08) in urban areas



(b) Distribution of tobacco taxation propensity scores in 2009 (R09) in urban areas

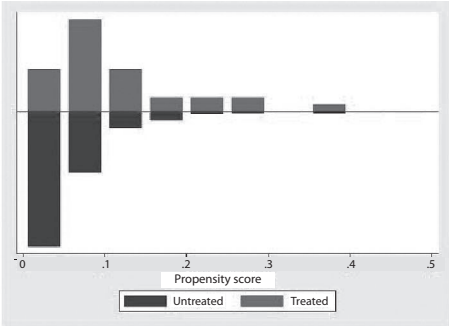


(c) Distribution of tobacco taxation propensity scores in 2012 (R12) in urban areas

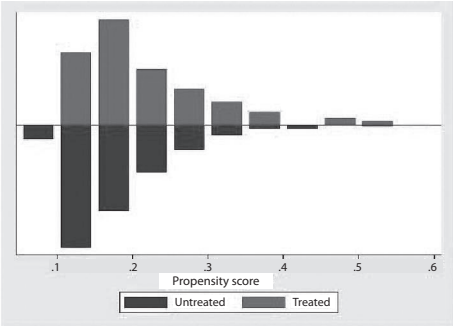


(d) Distribution of tobacco taxation propensity scores in 2017 (R17) in urban areas

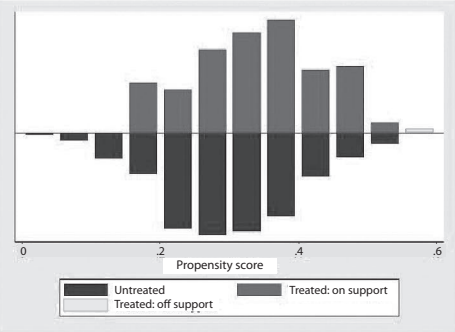
Figure 1 Distribution of tobacco taxation propensity scores (continuation)



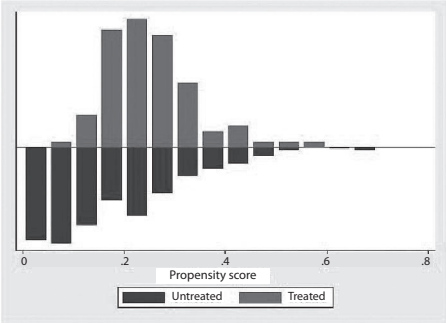
(e) Distribution of tobacco taxation propensity scores in 2008 (R08) in rural areas



(f) Distribution of tobacco taxation propensity scores in 2009 (R09) in rural areas



(g) Distribution of tobacco taxation propensity scores in 2012 (R12) in rural areas



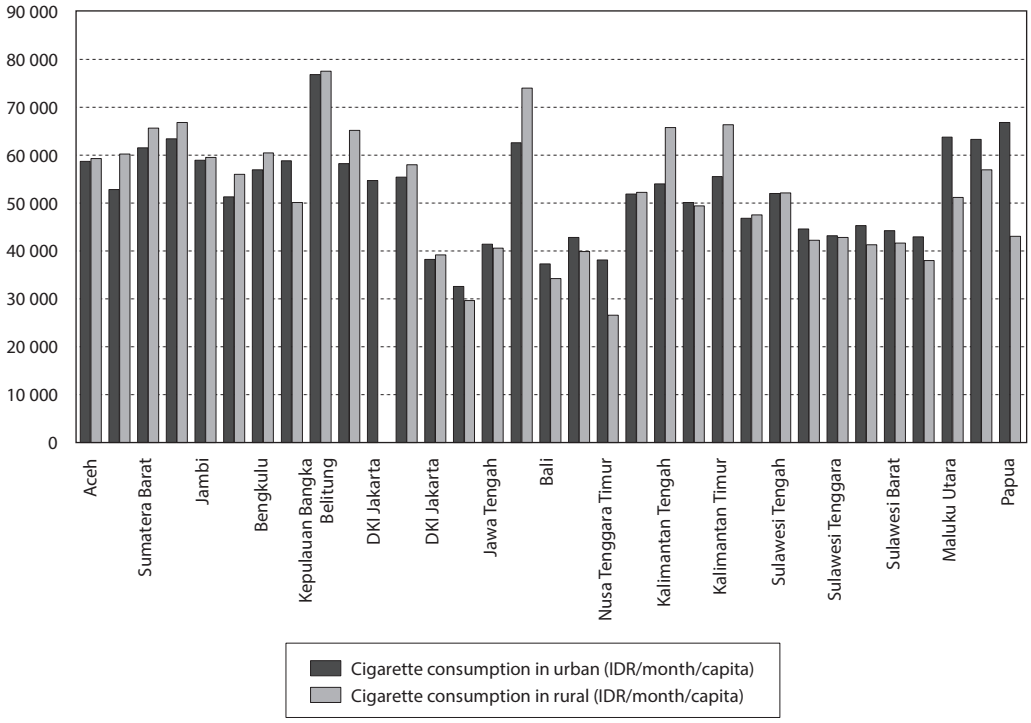
(h) Distribution of tobacco taxation propensity scores in 2017 (R17) in rural areas

Source: Authors analysis, 2023

3 RESULTS

The level of cigarette consumption in rural and urban areas seems to have differences where cigarette consumption in urban areas is higher than in rural areas. The average cigarette consumption in urban areas from 2007 to 2022 reaches IDR 54 647.56/month/capita, while in rural areas the average cigarette consumption is only IDR 51 655.23/month/capita. Furthermore, cigarette consumption in both rural and urban areas shows quite diverse differences at the provincial level (Figure 2). Bangka Belitung Province is recorded to have the highest level of cigarette consumption in rural and urban areas. The level of cigarette consumption in rural areas reached IDR 77 536.75/month/capita followed by IDR 76 821.06/month/capita in the urban areas. The high level of cigarette consumption in Bangka Belitung Province compared to other provinces is caused by the tendency of people to consume cigarettes in every social activity. Social relations are closely related to the level of one's cigarette consumption (Denney, Sharp and Kimbro, 2022; Fithria, Adlim, Jannah and Tahlil, 2021). Someone will tend to consume more cigarettes in a smoker's environment than a non-smoker's (Thomeer, Hernandez, Umberson and Thomas, 2019). Furthermore, the province with the lowest level of cigarette consumption in rural areas is East Nusa Tenggara Province, while in urban areas is the Special Region of Yogyakarta Province where cigarette consumption in each of these provinces is only IDR 26 516.31/month/capita and IDR 32 555.13/month/capita.

Figure 2 Cigarette consumption in urban-rural (IDR/month/capita)



Source: Central Bureau of Statistics of the Republic of Indonesia, 2007–2021

The 2SLS model shows that the equations compiled have met the post-estimation test where this test is a requirement that must be met to use the model properly. Using the 2SLS model requires endogeneity problems in the built model (Prasada and Dhamira, 2022; Prasada, Nugroho and Lakner, 2022). Endogeneity equations can occur when the independent variable is thought to affect a dependent variable, but, at the same time, the independent variable is affected by other independent variables. The post-estimation test shows that the Hausman statistic values for the urban and rural models are 10.07 (prob.=0.00) and 48.66 (prob.=0.00), respectively (Table 4). The probability value of the Hausman statistic is significant at the 1% alpha level, so it can be interpreted that this model has an endogeneity problem, so it fulfills the requirements for using the 2SLS model.

Furthermore, this model is indicated to have strong instrument variables through the Stock & Yogo statistic value of 15.95 (prob.=0.00) for the urban model and 9.08 (prob.=0.00) for the rural model. In addition, the Sargan statistical values for the urban and rural models each show a value of 70.52 (prob.=0.00) and 33.39 (prob.=0.00), meaning that the equation used is included in the overidentified equation category. Therefore, solving this equation must be done using the 2SLS model. Another statistical indicator, namely the Adj.R^2 value in urban and rural models, shows high values, 0.76 and 0.42, respectively. The value of the F statistic also looks significant at the 1% alpha level. Various statistical indicators indicate that selecting the 2SLS model in this study is the right decision because all the requirements for using the model can be fulfilled properly.

Table 4 Determinants of cigarette consumption in Indonesia

Variable	Urban				Rural			
	Coefficient	Std. error	t-statistic	Prob.	Coefficient	Std. error	t-statistic	Prob.
Dependent variable: UNE								
MIN	0.30	0.07	4.10	0.00 ***	0.24	0.08	3.07	0.00 ***
GRO	–0.23	0.04	–5.37	0.00 ***	–0.18	0.05	–3.97	0.00 ***
ENR	0.07	0.06	1.19	0.24 ns	0.07	0.06	1.27	0.21 ns
ILL	–0.05	0.04	–1.24	0.22 ns	–0.04	0.04	–1.01	0.31 ns
Cons.	–0.80	0.13	–6.38	0.00 ***	–0.75	0.13	–5.63	0.00 ***
Adj. R2				0.26				0.25
F statistic				16.58				15.58
F prob.				0.00				0.00
Dependent variable: URB or RUR								
UNE	0.30	0.07	4.04	0.00 ***	0.82	0.15	5.33	0.00 ***
WAG	0.15	0.09	1.73	0.08 *	0.16	0.14	1.14	0.25 ns
INE	–0.23	0.02	–9.98	0.00 ***	–0.03	0.04	–0.74	0.46 ns
VIO	–0.02	0.02	–0.92	0.36 ns	0.08	0.04	1.93	0.05 *
R08	–3.20	0.13	–24.43	0.00 ***	–3.55	0.23	–15.12	0.00 ***
R09	–2.46	0.09	–28.12	0.00 ***	–2.76	0.16	–17.07	0.00 ***
R12	–1.25	0.06	–20.04	0.00 ***	–1.45	0.10	–14.61	0.00 ***
R17	–0.85	0.07	–12.30	0.00 ***	–0.81	0.11	–7.32	0.00 ***
Cons.	1.40	0.06	24.01	0.00 ***	1.56	0.10	15.56	0.00 ***
Adj. R2				0.76				0.42
F statistic				1 630.61				627.76
F prob.				0.00				0.00
Overidentification test				70.52				33.39
Weak instruments test				15.95				9.08
Endogeneity test				10.07				48.66

Note: *** significant at 1% alpha; * significant at 10% alpha; ns not significant.
Source: Authors analysis, 2023

The results of the 2SLS analysis on urban and rural models show that the variable unemployment rate (UNE) is positively influenced by the provincial minimum wage (MIN), meaning that the higher the provincial minimum wage, the higher the unemployment rate will occur, and vice versa. The UNE variable in the urban and rural models is also negatively affected by the economic growth variable (GRO), meaning that the lower the economic growth rate, the higher the unemployment rate will occur.

Furthermore, the UNE variable has a positive effect on the level of cigarette consumption in urban (URB) and rural (RUR) communities. This shows that the higher the unemployment level in urban

and rural areas, the higher the level of public consumption of cigarettes. In the urban model, the average real wage of workers (WAG) also positively affects the cigarette consumption variable, meaning that the higher the real wage of workers, the higher the consumption of cigarettes by people in urban areas, and vice versa. In addition, the Gini index variable (INE), which shows income disparities in urban areas, has a negative effect on urban cigarette consumption, so the higher the income disparity, the lower the cigarette consumption will be. In contrast to the urban model, in the rural model, the VIO variable, which shows the level of population who are victims of violence, has a positive effect on cigarette consumption, meaning that the higher the number of people who are victims of violence, the higher the level of cigarette consumption in rural areas.

Another variable, namely the dummy variable for the implementation of the tobacco tax increase policy in 2008 (R08), 2009 (R09), 2012 (R12), and 2017 (R17) affects the level of cigarette consumption in Indonesia, both in urban and rural areas with negative regression coefficient. The negative regression coefficient indicates that the level of cigarette consumption after implementing the policy is lower than before the policy was implemented. In addition, the results of the impact evaluation analysis using the PSM method show similarities in the decline in cigarette consumption between urban and rural areas (although with different impact sizes). During the implementation of the 2008 increase in tobacco tax, cigarette consumption in urban and rural areas decreased by IDR 5 783.67/month/capita and IDR 9 393.71/month/capita, respectively (Table 5). In the increase in tobacco tax in 2009, consumption of cigarettes in urban and rural areas again decreased by IDR 21 438.58/month/capita and IDR 26 959.85/month/capita, respectively. The decline in this period was the largest decrease in cigarette consumption due to the increased tobacco taxation compared to other periods. Cigarette consumption in urban and rural areas decreased by IDR 4 416.51/month/capita and IDR 12 838.25/month/capita, respectively, when the 2012 tobacco tax increase policy was officially implemented. This decline continued when the increase in tobacco tax was carried out again in 2017, where the increase in tobacco tax reduced cigarette consumption in urban and rural areas by IDR 3 355.02/month/capita and IDR 4 123.43/month/capita, respectively.

Table 5 Impact evaluation results of tobacco taxation policy

Tobacco excise tariffs period	Impact on RCA index (difference after matching)	t-statistics
Urban		
R08	–5 783.76	–8.43 ***
R09	–21 438.58	–7.85 ***
R12	–4 416.51	–2.88 ***
R17	–3 355.02	–4.67 ***
Rural		
R08	–9 393.71	–2.83 ***
R09	–26 959.85	–7.97 ***
R12	–12 838.25	–4.25 ***
R17	–4 123.43	–4.78 ***

Note: *** significant at 1% alpha (t-table = 2.33).
Source: Authors analysis, 2023

4 DISCUSSION

4.1 Factors affecting the unemployment rate

In the urban model, it can be seen that the provincial minimum wage (MIN) has a positive and significant effect on the unemployment rate with a regression coefficient of 0.30. The regression coefficient value indicates that an increase in the provincial minimum wage of 1.00% can increase urban unemployment by 0.30%. The provincial minimum wage is closely related to the ability of a company to provide wages to all its workers. Increased minimum wages (exceeding the market clearing rate) encourage companies to increase the efficiency of their production activities, thereby triggering a reduction in the number of workers (Clemens and Wither, 2019; Dreepaul-Dabee and Tandrayen-Ragoobur, 2023; Kawaguchi and Mori, 2021). The influence of the provincial minimum wage in rural areas is lower than in urban areas. This can be seen from the coefficient of the MIN variable in the rural model which is 0.24 (lower than the UNE regression coefficient in the urban model), meaning that an increase in the provincial minimum wage of 1.00% can increase the numbers of unemployment in rural areas by 0.24%. These results can occur because the informal sector dominates the composition of employment in rural areas, so implementing the provincial minimum wage policy in rural areas has a lower elasticity of the unemployment rate than urban areas (Pérez Pérez, 2020; Siregar, 2020).

Economic growth has a negative and significant effect on unemployment rates in urban and rural areas with regression coefficient values of 0.23 and 0.18, respectively. Higher economic growth encourages an increase in the number of jobs so that the unemployment rate can be reduced (Hjazeen, Seraj and Ozdeser, 2021). The effect of economic growth on the unemployment rate in urban areas is greater than in rural areas. This is due to better connectivity in urban areas in response to rapid economic growth (Tiwasing, Clark and Gkartzios, 2022). Higher economic growth will require increased interconnection between one business unit and another. This encourages business units to develop more quickly and ultimately can encourage increased employment opportunities that can be accessed by the community.

4.2 Factors affecting cigarette consumption

The 2SLS analysis shows that the unemployment rate has a positive and significant effect on the level of cigarette consumption both in urban and rural areas. In the urban model, the UNE coefficient is 0.30, which means that a 1,00% increase in the unemployment rate will increase cigarette consumption by 0.30%. The effect of the unemployment rate in rural areas is bigger than in urban areas. The UNE coefficient in rural areas is 0.82. It can be interpreted that a 1.00% increase in the unemployment rate will increase the cigarette consumption by 0.82% in the rural area. The unemployment rate can trigger an increase or decrease in consumption where higher unemployment forces consumers to lower their expectations of future income levels, so that consumption levels will be lower (Campos and Reggio, 2015). Furthermore, the effect of unemployment on cigarette consumption is bigger in the rural area. This is due to the difficulty in obtaining other alternative jobs in rural areas compared to urban areas, so that when unemployment increases, the decline in future income expectations in rural areas becomes higher than in urban areas (Chen, Huang, Cheng, Tang and Huang, 2023; Nieto Masot, Cárdenas Alonso and Engelmo Moriche, 2020). In addition, an increase in unemployment causes an increase in leisure time which can encourage an increase in cigarette consumption for residents in both urban and rural areas (Aadahl et al., 2021; Verkooijen et al., 2009).

Real wages of workers in urban areas positively and significantly affect cigarette consumption in urban areas. The regression coefficient of this variable is 0.15, meaning that an increase in workers' real wages of 1.00% can increase cigarette consumption in urban areas by 0.15%. Growing real wages increase disposable income (Avram, Brewer, Fisher and Fumagalli, 2022; Ku, Lee, Lee and Han, 2018). Therefore, the higher the real wage of workers, the greater the amount of cigarette consumption. An increase must follow an increase in workers' real wages in public awareness about the dangers of smoking,

so it is expected that an increase in workers' real wages will not impact increasing cigarette consumption. Raising awareness regarding the dangers of smoking can be done by implementing a socialization program on the dangers of smoking, especially for teenagers, mass advertising campaigns in various electronic and print media, mobilizing communities who are activists for the smoking cessation movement, health warnings on tobacco products, and prohibiting smoking in public places (Golechha, 2016).

The income disparity variable (INE) in the urban model has a negative and significant effect on cigarette consumption with a regression coefficient of 0.23, meaning that an increase in income disparity of 1.00% can reduce cigarette consumption in urban areas by 0.23%. Increasing income inequality causes a decrease in people's purchasing power (Crespo Cuaresma et al., 2018). In addition, higher inequality encourages people to be more selective in choosing the products they buy (Shen, Fan and Hu, 2022; Velandia-Morales, Rodríguez-Bailón and Martínez, 2022). Inequality influences consumer behavior, where consumers with high levels of inequality tend to consume products that can improve their social status in society (Velandia-Morales et al., 2022). On the other hand, smoking is considered as an activities carried out by lower class people (Golden et al., 2018). Hence, people with high incomes tend to reduce cigarette consumption. At the same time, people with low income levels cannot afford to buy cigarettes due to increasing purchase prices. Therefore, when income inequality gets bigger, people will prioritize buying their basic needs of goods or services, so that consumption of cigarettes will be lower.

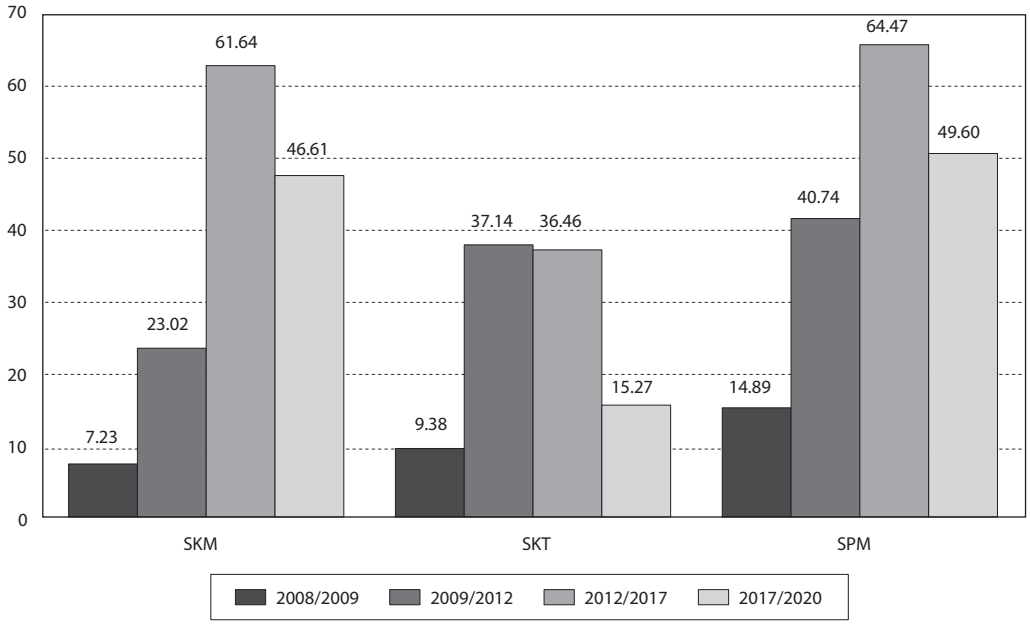
The percentage of population who are victims of violence (VIO) has a positive and significant effect on the level of cigarette consumption in rural areas with a regression coefficient of 0.08. This coefficient means that an increase in the number of victims of violence by 1.00% can increase the amount of cigarette consumption in rural areas by 0.08%. Someone who has experienced physical or sexual violence tends to seek an escape from the problems they are experiencing, so the fastest and easiest alternative to do this is smoking (Lewis, Oberleitner, Morgan, Picciotto and McKee, 2016). Furthermore, negative emotions due to acts of violence experienced by someone can trigger an increase in the frequency of smoking, so the consumption of cigarettes will increase (Spaducci et al., 2020; Y. Wang, Chen, Gong and Yan, 2016). Therefore, to reduce cigarette consumption, appropriate policies are needed to reduce violence in society. Several policies that can be implemented include encouraging the formation of communication that focuses on levels of violence, especially on teenagers and children, media campaigns that highlight preventing acts of violence, and counseling and therapy for mental health support (Araten-Bergman and Bigby, 2023; Pundir, Saran, White, Adona and Subrahmanian, 2019).

4.3 Impact of increasing tobacco taxation on cigarette consumption

Indonesia has a very complex tobacco tax policy where tobacco taxation is differentiated based on several categories at once, namely production volume, production technique (i.e., machine-rolled vs hand-rolled), aroma (i.e., white cigarettes vs clove cigarettes or kretek cigarettes), and retail price (Bella et al., 2023). Tobacco taxation in Indonesia has always been increasing from one period to another. In 2009 there was an increase in tobacco taxation compared to 2008 for the categories of machine-rolled kretek cigarettes (SKM), hand-rolled kretek cigarettes (SKT), and machine-made white cigarettes (SPM) respectively by 7.23%, 9.38% and 14. 89% (Figure 3). Tobacco tax increases continued in 2012, 2017 and 2020 with different percentages of tobacco tax increases.

The gradual increase in tobacco taxation has a direct effect on the level of cigarette consumption in both urban and rural areas. The results of the 2SLS analysis show that the dummy variable of the application of tobacco tax policy in each period (R08, R09, R12, and R17) has a negative sign regression coefficient, meaning that the level of cigarette consumption in the period before the increase in tobacco tax was higher than the period after. This result is also in line with the impact analysis results using the PSM model, where the impact value in each period of increased tobacco tax has a negative value.

Figure 3 The development of tobacco taxation in Indonesia



Source: Ministry of Finance of the Republic of Indonesia, 2007–2021

Impact evaluation analysis using the PSM method shows that the reduction in cigarette consumption rates in rural areas is higher than in urban areas. The increase in tobacco taxation in 2008 led to a decrease in cigarette consumption in rural areas by IDR 9 393.71/month/capita, while tobacco consumption in urban areas decreased by IDR 5 783.76/month/capita (Table 5.). Furthermore, in 2009 the increase in tobacco taxation caused cigarette consumption in urban areas to fall by IDR 21 438.58/month/capita, lower than the decrease in cigarette consumption in rural areas, reaching IDR 26 959.85/month/capita. In 2012, the increase in tobacco taxation led to a decrease in the level of consumption of cigarettes in urban and rural areas by IDR 4 416.51/month/capita and IDR 12 838.25/month/capita, respectively. The decline in cigarette consumption occurred again after the tobacco tax increase policy was implemented in 2017, where the policy pushed cigarette consumption in urban and rural areas to fall by IDR 3 355.02/month/capita and IDR 4 123.43/month/capita, respectively. The impact of increasing tobacco taxation is greater in rural areas due to the high-income gap between urban and rural communities, so the response to reduced cigarette consumption to increased tobacco taxation in rural areas is higher than in urban areas (Liu and Long, 2021; Zhong, Wang, Zhu, Chen and Huang, 2022). These results also indicate that implementing tobacco taxation in urban areas is ineffective compared to rural areas. Differences in the effectiveness of implementing tobacco taxation in urban areas and rural areas are driven by differences in population density, income level, poverty level, and level of public education in each region (Darden, 2021).

The tobacco tax increase policy will be more effective if it is carried out with other interventions such as the anti-smoking campaign (Kalousova et al., 2020; Parks et al., 2021). The anti-smoking campaign acts as a complement that accommodates the behaviour of smokers with various characteristics (Colombo and Galmarini, 2023). Another effort to significantly reduce people's consumption of cigarettes is to provide images of the dangers of consuming cigarettes on packaging (Kim and Khang, 2020). Therefore, a proportional combination of policies between tobacco taxationes and accurate public education

is needed regarding the relative risks that can result from cigarette use (Cummings, Ballin and Sweanor, 2020). Sæbø and Lund (2022) emphasizes that perceptions of the risks of smoking use can be formed by conveying accurate information to the public. The implementation of the tobacco tax increase policy in 2009 appears to have had the greatest impact compared to other tobacco tax increase periods. This was driven by low economic growth due to the global crisis in 2008–2009, causing an increase in tobacco taxation in that period to have a greater impact than in other periods (Nützenadel, 2020; Resosudarmo, Abdurohman, Yusuf and Hartono, 2021; Tambunan, 2019).

Furthermore, it is important for the Indonesian government to reform tobacco taxation policy. The tobacco taxation currently implemented in Indonesia follows a very complex taxation system, where tobacco tax is determined based on various categories, starting from production volume, production technique (i.e., machine-rolled vs. hand-rolled), aroma (i.e., white cigarettes vs. clove cigarettes or kretek cigarettes), and retail price (Bella et al., 2023). This could potentially lead to inefficiencies in the implementation of tobacco taxation. In addition, higher tobacco taxes increase the potential for the phenomenon of untaxed cigarette smuggling to emerge, so supervision regarding the implementation of tobacco taxes needs to be tightened.

CONCLUSION

Tobacco policy improvements have negatively impacted cigarette consumption levels in both urban and rural areas. This indicates that the government's tobacco tax increase policy has successfully reduced cigarette consumption in both areas. The results also show that the decline in cigarette consumption due to the implementation of tobacco taxation in rural areas is greater than in urban areas. Rural and urban areas have different characteristics, both in terms of income level, population density, poverty level, and education level, so they have different responses to changes in tobacco taxes. These results indicate an ineffective implementation of the tobacco tax in urban areas. Therefore, it is necessary to increase the effectiveness of tobacco tax in urban areas by simultaneously implementing other intervention policies complementary to tobacco tax policies. In addition, cigarette consumption in urban and rural areas can be reduced by taking into account the unemployment rate variable, where a lower unemployment rate can reduce the level of public consumption of cigarettes. Furthermore, the unemployment rate can be reduced by carefully considering the increase in the provincial minimum wage every year and implementing policies that can encourage economic growth in both urban and rural areas.

Workers‘ real wages and income disparity are other variables that need to be considered to reduce cigarette consumption in urban areas. An increase in workers' real wages can encourage an increase in cigarette consumption. Hence, an increase in real wages needs to be balanced by implementing policies that can increase public awareness about the dangers of smoking. In rural areas, reducing cigarette consumption can be reduced by considering the violence in society. Higher levels of violence can lead to increased consumption of cigarettes in rural communities.

References

AADAHL, M., ANDREASEN, A. H., PETERSEN, C. B., GUPTA, N., HOLTERMANN, A., LAU, C. J. (2021). Should leisure-time sedentary behavior be replaced with sleep or physical activity for prevention of diabetes? [online]. *Scandinavian Journal of Medicine & Science in Sports*, 31(5): 1105–1114. <https://doi.org/10.1111/sms.13924>.

ANDERSSON, P. F. (2018). Democracy, Urbanization, and Tax Revenue [online]. *Studies in Comparative International Development*, 53(1): 111–150. <https://doi.org/10.1007/s12116-017-9235-0>.

ARATEN-BERGMAN, T., BIGBY, C. (2023). Violence Prevention Strategies for People with Intellectual Disabilities: A Scoping Review [online]. *Australian Social Work*, 76(1): 72–87. <https://doi.org/10.1080/0312407X.2020.1777315>.

AVRAM, S., BREWER, M., FISHER, P., FUMAGALLI, L. (2022). Household Earnings and Income Volatility in the UK, 2009–2017 [online]. *The Journal of Economic Inequality*, 20(2): 345–369. <https://doi.org/10.1007/s10888-021-09517-3>.

BARDACH, A., CASARINI, A., RODRIGUEZ CAIROLI, F., ADENIRAN, A., CASTRADORI, M., AKANONU, P., PALACIOS, A. et al. (2022). The estimated benefits of increasing cigarette prices through taxation on the burden of disease and economic burden of smoking in Nigeria: A modeling study [online]. *PLOS ONE*, 17(3): e0264757. <https://doi.org/10.1371/journal.pone.0264757>.

BELLA, A., SWARNATA, A., VULOVIC, V., NUGROHO, D., MEILISSA, Y., USMAN, U., DARTANTO, T. (2023). Macroeconomic impact of tobacco taxation in Indonesia [online]. *Tobacco Control*, tc-2022-057735. <https://doi.org/10.1136/tc-2022-057735>.

BOACHIE, M. K., AYIFAH, R. N. Y., IMMURANA, M., AGYEMANG, J. K., SINGH, A., ROSS, H. (2022). Effect of cigarette prices on cigarette consumption in Ghana [online]. *Drug and Alcohol Dependence Reports*, 5: 100102. <https://doi.org/10.1016/j.dadr.2022.100102>.

CAMPOS, R. G., REGGIO, I. (2015). Consumption in the shadow of unemployment [online]. *European Economic Review*, 78: 39–54. <https://doi.org/10.1016/j.euroecorev.2015.04.006>.

CENTRAL BUREAU OF STATISTICS. (2023a). *Number of Districts 2014–2016* [online]. [cit. 19.9.2023]. <https://www.bps.go.id/indicator/101/154/1/jumlah-kabupaten.html>.

CENTRAL BUREAU OF STATISTICS. (2023b). Number of Large and Medium Industrial Companies [online]. [cit. 19.5.2023]. <https://www.bps.go.id/indicator/9/200/5/jumlah-perusahaan-ibs-kbli-2020-.html>.

CENTRAL BUREAU OF STATISTICS. (2023c). *State Revenue Realization (Billion Rupiah)* [online]. [cit. 19.5.2023]. <https://www.bps.go.id/indicator/13/1070/1/realisasi-pendapatan-negara.html>.

CHEN, M., HUANG, X., CHENG, J., TANG, Z., HUANG, G. (2023). Urbanization and vulnerable employment: Empirical evidence from 163 countries in 1991–2019 [online]. *Cities*, 135: 104208. <https://doi.org/https://doi.org/10.1016/j.cities.2023.104208>.

CHENG, K. J. G., ESTRADA, M. A. G. (2020). Price Elasticity of cigarette smoking demand in the Philippines after the 2012 Sin Tax Reform Act [online]. *Preventive Medicine*, 134: 106042. <https://doi.org/10.1016/j.ypmed.2020.106042>.

CLEMENS, J., WITHER, M. (2019). The minimum wage and the Great Recession: Evidence of effects on the employment and income trajectories of low-skilled workers [online]. *Journal of Public Economics*, 170: 53–67. <https://doi.org/10.1016/j.jpubeco.2019.01.004>.

COLOMBO, L., GALMARINI, U. (2023). Taxation and anti-smoking campaigns: Complementary policies in tobacco control [online]. *Journal of Policy Modeling*, 45(1): 31–57. <https://doi.org/10.1016/j.jpolmod.2022.11.006>.

CRESPO CUARESMA, J., KUBALA, J., PETRIKOVA, K. (2018). Does income inequality affect aggregate consumption? Revisiting the evidence [online]. *Empirical Economics*, 55(2): 905–912. <https://doi.org/10.1007/s00181-017-1302-x>.

CUMMINGS, K. M., BALLIN, S., SWEANOR, D. (2020). The past is not the future in tobacco control [online]. *Preventive Medicine*, 140: 106183. <https://doi.org/10.1016/j.ypmed.2020.106183>.

DAI, X., GIL, G. F., REITSMA, M. B., AHMAD, N. S., ANDERSON, J. A., BISIGNANO, C., GAKIDOU, E. et al. (2022). Health effects associated with smoking: a Burden of Proof study [online]. *Nature Medicine*, 28(10): 2045–2055. <https://doi.org/10.1038/s41591-022-01978-x>.

DARDEN, M. E. (2021). Cities and Smoking [online]. *Journal of Urban Economics*, 122: 103319. <https://doi.org/10.1016/j.jue.2021.103319>.

DENNEY, J. T., SHARP, G., KIMBRO, R. T. (2022). Community social environments and cigarette smoking [online]. *SSM – Population Health*, 19: 101167. <https://doi.org/10.1016/j.ssmph.2022.101167>.

DREEPAUL-DABEE, V., TANDRAYEN-RAGOOBUR, V. (2023). Minimum wage and employment: a gender perspective for Mauritius [online]. *Journal of Business and Socio-Economic Development*, 3(1): 1–16. <https://doi.org/10.1108/JBSED-10-2021-0134>.

FITHRIA, F., ADLIM, M., JANNAH, S. R., TAHLIL, T. (2021). Indonesian adolescents' perspectives on smoking habits: a qualitative study [online]. *BMC Public Health*, 21(1): 82. <https://doi.org/10.1186/s12889-020-10090-z>.

FRIEDSON, A., LI, M., MECKEL, K., REES, D. I., SACKS, D. W. (2023). Cigarette taxes, smoking, and health in the long run [online]. *Journal of Public Economics*, 222: 104877. <https://doi.org/10.1016/j.jpubeco.2023.104877>.

GOLDEN, S. D., KONG, A. Y., LEE, J. G. L., RIBISL, K. M. (2018). Disparities in cigarette tax exposure by race, ethnicity, poverty status and sexual orientation, 2006–2014, USA [online]. *Preventive Medicine*, 108: 137–144. <https://doi.org/10.1016/j.ypmed.2017.12.017>.

GOLECHHA, M. (2016). Health promotion methods for smoking prevention and cessation: A comprehensive review of effectiveness and the way forward [online]. *International Journal of Preventive Medicine*, 7(7): 1–7. <https://doi.org/10.4103/2008-7802.173797>.

HAILU DEMEKE, Y. (2022). Youth unemployment and political instability: Evidence from IGAD member countries [online]. *Cogent Economics & Finance*, 10(1): 2079211. <https://doi.org/10.1080/23322039.2022.2079211>.

HJAZEEN, H., SERAJ, M., OZDESER, H. (2021). The nexus between the economic growth and unemployment in Jordan [online]. *Future Business Journal*, 7(1): 42. <https://doi.org/10.1186/s43093-021-00088-3>.

HO, L.-M., SCHAFFERER, C., LEE, J.-M., YEH, C.-Y., HSIEH, C.-J. (2018). Raising cigarette excise tax to reduce consumption in low-and middle-income countries of the Asia-Pacific region: A simulation of the anticipated health and taxation revenues impacts [online]. *BMC Public Health*, 18(1): 1187. <https://doi.org/10.1186/s12889-018-6096-z>.

HOLIPAH, H., SULISTOMO, H. W., MAHARANI, A. (2020). Tobacco smoking and risk of all-cause mortality in Indonesia [online]. *PLOS ONE*, 15(12): e0242558. <https://doi.org/10.1371/journal.pone.0242558>.

HUANG, C., LIU, F., YOU, S. (2021). The impact of minimum wage increases on cigarette smoking [online]. *Health Economics*, 30(9): 2063–2091.<https://doi.org/10.1002/hec.4362>.

INTERNATIONAL TRADE CENTRE. (2023). List of exporters for the selected product in 2021 Product : 2402 Cigars, cheroots, cigarillos and cigarettes of tobacco or of tobacco substitutes [online]. [cit. 19.5.2023]. <https://www.trademap.org/Country_SelProduct.aspx?nvpm=1%7C%7C%7C%7C%7C2402%7C%7C%7C4%7C1%7C1%7C2%7C1%7C1%7C2%7C1%7C1%7C1>.

KALOUSOVA, L., LEVY, D., TITUS, A. R., MEZA, R., THRASHER, J. F., ELLIOTT, M. R., FLEISCHER, N. L. (2020). Cigarette taxes, prices, and disparities in current smoking in the United States [online]. *SSM – Population Health*, 12: 100686. <https://doi.org/10.1016/j.ssmph.2020.100686>.

KAWAGUCHI, D., MORI, Y. (2021). Estimating the effects of the minimum wage using the introduction of indexation [online]. *Journal of Economic Behavior & Organization*, 184: 388–408. <https://doi.org/10.1016/j.jebo.2021.01.032>.

KIM, I., KHANG, Y.-H. (2020). Differential changes in quitting smoking by daily cigarette consumption and intention to quit after the introduction of a tobacco tax increase and pictorial cigarette pack warnings in Korea, 2013–2017 [online]. *Drug and Alcohol Dependence*, 213: 108085. <https://doi.org/10.1016/j.drugalcdep.2020.108085>.

KLEPPANG, A. L., SKILLE, E. Å. (2022). Leisure-Time Activities in Different Contexts and Depressive Symptoms in Norwegian Adolescents: A Cross-Sectional Study [online]. *International Journal of Environmental Research and Public Health*. <https://doi.org/10.3390/ijerph191710769>.

KU, I., LEE, W., LEE, S., HAN, K. (2018). The Role of Family Behaviors in Determining Income Distribution: The Case of South Korea [online]. *Demography*, 55(3): 877–899. <https://doi.org/10.1007/s13524-018-0670-y>.

KUSS, O., BLETTNER, M., BÖRGERMANN, J. (2016). Propensity Score: an Alternative Method of Analyzing Treatment Effects [online]. *Deutsches Arzteblatt International*, 113(35–36): 597–603. <https://doi.org/10.3238/arztebl.2016.0597>.

LEVIN, A., LIN, C. F., CHU, C. S. J. (2002). Unit root tests in panel data: Asymptotic and finite-sample properties [online]. *Journal of Econometrics*, 108(1): 1–24. <https://doi.org/10.1016/S0304-4076(01)00098-7>.

LEWIS, A. S., OBERLEITNER, L. M. S., MORGAN, P. T., PICCIOTTO, M. R., MCKEE, S. A. (2016). Association of Cigarette Smoking With Interpersonal and Self-Directed Violence in a Large Community-Based Sample [online]. *Nicotine & Tobacco Research*, 18(6): 1456–1462. <https://doi.org/10.1093/ntr/ntv287>.

LI, L., HAN, J., ZHU, Y. (2023). Does farmland inflow improve the green total factor productivity of farmers in China? An empirical analysis based on a propensity score matching method [online]. *Heliyon*, 9(3): e13750. <https://doi.org/10.1016/j.heliyon.2023.e13750>.

LIM, Y. K., SHIN, D. W., KIM, H. S., YUN, J. M., SHIN, J.-H., LEE, H., CHO, M. H. et al. (2017). Persistent smoking after a cardiovascular event: A nationwide retrospective study in Korea [online]. *PLOS ONE*, 12(10): e0186872. <https://doi.org/10.1371/journal.pone.0186872>.

LIU, Y., LONG, C. (2021). Urban and Rural Income Gap: Does Urban Spatial Form Matter in China? [online]. *SAGE Open*, 11(1): 21582440211002216. <https://doi.org/10.1177/21582440211002215>.

LÓPEZ-ESPÍN, J. J., VIDAL, A. M., GIMÉNEZ, D. (2012). Two-stage least squares and indirect least squares algorithms for simultaneous equations models [online]. *Journal of Computational and Applied Mathematics*, 236(15): 3676–3684. <https://doi.org/10.1016/j.cam.2011.07.005>.

MILIC, M., GAZIBARA, T., PEKMEZOVIC, T., KISIC TEPAVCEVIC, D., MARIC, G., POPOVIC, A., LEVINE, H. et al. (2020). Tobacco smoking and health-related quality of life among university students: Mediating effect of depression [online]. *PLOS ONE*, 15(1): e0227042. <https://doi.org/10.1371/journal.pone.0227042>.

MUSHTAQ, M., AHMED, S., FAHLEVI, M., ALJUAID, M., SANIUK, S. (2022). Globalization and employment nexus: Moderating role of human capital [online]. *PLOS ONE*, 17(10): e0276431. <https://doi.org/10.1371/journal.pone.0276431>.

NGUYEN, H. T. T., GIANG, L. T., PHAM, T. N. (2020). Impacts of higher tobacco tax on output and employment in Vietnam [online]. *Journal of Economics and Development*, 22(1): 167–182. <https://doi.org/10.1108/JED-11-2019-0058>.

NIETO MASOT, A., CÁRDENAS ALONSO, G., ENGELMO MORICHE, Á. (2020). Spatial Analysis of the Rural-Urban Structure of the Spanish Municipalities [online]. *ISPRS International Journal of Geo-Information*, 9(4). <https://doi.org/10.3390/ijgi9040213>.

NÜTZENADEL, A. (2020). The financial crisis of 2008 – Experience, memory, history [online]. *Journal of Modern European History*, 19(1): 3–7. <https://doi.org/10.1177/1611894420973590>.

OUYANG, C., LI, D., LI, X., XIAO, J., SUN, W., WANG, Y. (2020). Cyber victimization and tobacco and alcohol use among adolescents: A moderated mediation model [online]. *Children and Youth Services Review*, 114: 105041. <https://doi.org/10.1016/j.childyouth.2020.105041>.

PARKS, M. J., PATRICK, M. E., LEVY, D. T., THRASHER, J. F., ELLIOTT, M. R., FLEISCHER, N. L. (2021). Tobacco Taxation and Its Prospective Impact on Disparities in Smoking Initiation and Progression Among Young Adults [online]. *Journal of Adolescent Health*, 68(4): 765–772. <https://doi.org/10.1016/j.jadohealth.2020.08.031>.

PÉREZ PÉREZ, J. (2020). The minimum wage in formal and informal sectors: Evidence from an inflation shock [online]. *World Development*, 133: 104999. <https://doi.org/10.1016/j.worlddev.2020.104999>.

PRASADA, I. Y., DHAMIRA, A. (2022). Non-Tariff Measures and Competitiveness of Indonesia's Natural Rubber Export in Destination Countries [online]. *AGRARIS: Journal of Agribusiness and Rural Development Research*, 8(2): 181–197. <https://doi.org/10.18196/agraris.v8i2.11392>.

PRASADA, I. Y., NUGROHO, A. D., LAKNER, Z. (2022). Impact of the FLEGT licence on Indonesian plywood competitiveness in the European Union [online]. *Forest Policy and Economics*, 144(November): 102848. <https://doi.org/10.2139/ssrn.4150386>.

PUNDIR, P., SARAN, A., WHITE, H., ADONA, J., SUBRAHMANIAN, R. (2019). PROTOCOL: The effectiveness of interventions for reducing violence against children: An evidence and gap map in low- and middle-income countries [online]. *Campbell Systematic Reviews*, 15(3): e1040. <https://doi.org/10.1002/cl2.1040>.

PUTRI, L. P., RUSSELL, D. J., O'SULLIVAN, B. G., MELIALA, A., KIPPEN, R. (2022). A critical review of definitions of rural areas in Indonesia and implications for health workforce policy and research [online]. *Health Research Policy and Systems*, 20(1): 46. <https://doi.org/10.1186/s12961-022-00847-w>.

RESOSUDARMO, B. P., ABDUROHMAN, YUSUF, A. A., HARTONO, D. (2021). Spatial impacts of fiscal stimulus policies during the 2009 global financial crisis in Indonesia [online]. *Asia-Pacific Journal of Regional Science*, 5(1): 305–326. <https://doi.org/10.1007/s41685-020-00176-2>.

ROBERTS, M. E., TEFERRA, A. A., KELLER-HAMILTON, B., PATTERSON, J. G., FERKETICH, A. K. (2020). Shared and unique risk factors for tobacco use among rural versus urban adolescents [online]. *Preventive Medicine*, 140: 106239. <https://doi.org/https://doi.org/10.1016/j.ypmed.2020.106239>.

SÆBØ, G., LUND, I. (2022). Public support for further regulating smoking, snus and e-cigarettes in Norway, and its associations with risk perceptions and tobacco use [online]. *International Journal of Drug Policy*, 101: 103559. <https://doi.org/10.1016/j.drugpo.2021.103559>.

SHEN, Z., FAN, W., HU, J. (2022). Income inequality, consumption, and the debt ratio of Chinese households [online]. *PLOS ONE*, 17(5): e0265851. <https://doi.org/10.1371/journal.pone.0265851>.

SINGH, C. R., KATHIRESAN, K. (2015). Effect of cigarette smoking on human health and promising remedy by mangroves [online]. *Asian Pacific Journal of Tropical Biomedicine*, 5(2): 162–167. <https://doi.org/10.1016/S2221-1691(15)30337-3>.

SIREGAR, T. H. (2020). Impacts of minimum wages on employment and unemployment in Indonesia [online]. *Journal of the Asia Pacific Economy*, 25(1): 62–78. <https://doi.org/10.1080/13547860.2019.1625585>.

SPADUCCI, G., MCNEILL, A., HUBBARD, K., STEWART, D., YATES, M., ROBSON, D. (2020). Smoking-related violence in a mental health setting following the implementation of a comprehensive smoke-free policy: A content analysis of incident reports [online]. *International Journal of Mental Health Nursing*, 29(2): 202–211. <https://doi.org/10.1111/inm.12659>.

SUPRIHANTI, A., HARIANTO, H., SINAGA, B. M., KUSTIARI, R. (2018). The Impact of Cigarette Excise Tax Policy on Tobacco Market and Clove Market in Indonesia [online]. *International Journal of Economics and Financial Issues*, 8(6SE-articles): 54–60. <https://www.econjournals.com/index.php/ijefi/article/view/7228>.

TAMBUNAN, T. T. H. (2019). The impact of the economic crisis on micro, small, and medium enterprises and their crisis mitigation measures in Southeast Asia with reference to Indonesia [online]. *Asia & the Pacific Policy Studies*, 6(1): 19–39. <https://doi.org/https://doi.org/10.1002/app5.264>.

TAMVADA, J. P., SHRIVASTAVA, M., MISHRA, T. K. (2022). Education, social identity and self-employment over time: Evidence from a developing country [online]. *Small Business Economics*, 59(4): 1449–1468. <https://doi.org/10.1007/s11187-021-00583-5>.

TAYLOR, G. M. J., TREUR, J. L. (2023). An application of the stress-diathesis model: A review about the association between smoking tobacco, smoking cessation, and mental health [online]. *International Journal of Clinical and Health Psychology*, 23(1): 100335. <https://doi.org/10.1016/j.ijchp.2022.100335>.

THOMEER, M. B., HERNANDEZ, E., UMBERSON, D., THOMAS, P. A. (2019). Influence of social connections on smoking behavior across the life course [online]. *Advances in Life Course Research*, 42: 100294. <https://doi.org/10.1016/j.alcr.2019.100294>.

TIWASING, P., CLARK, B., GKARTZIOS, M. (2022). How can rural businesses thrive in the digital economy? A UK perspective [online]. *Heliyon*, 8(10): e10745. <https://doi.org/10.1016/j.heliyon.2022.e10745>.

VALLARTA-ROBLEDO, J. R., MARQUES-VIDAL, P., SANDOVAL, J. L., DE RIDDER, D., SCHAFFNER, E., HUMAIR, J.-P., GUESSOUS, I. et al. (2022). The neighborhood environment and its association with the spatio-temporal footprint of tobacco consumption and changes in smoking-related behaviors in a Swiss urban area [online]. *Health & Place*, 76: 102845. <https://doi.org/10.1016/j.healthplace.2022.102845>.

VAN DEN BOOGAARD, V., BEACH, R. (2023). Tax and governance in rural areas: The implications of inefficient tax collection [online]. *Journal of International Development* (n/a). <https://doi.org/https://doi.org/10.1002/jid.3756>.

VELANDIA-MORALES, A., RODRÍGUEZ-BAILÓN, R., MARTÍNEZ, R. (2022). Economic Inequality Increases the Preference for Status Consumption [online]. *Frontiers in Psychology*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.809101>.

VERKOOIJEN, K. T., NIELSEN, G. A., KREMERS, S. P. J. (2009). Leisure time physical activity motives and smoking in adolescence [online]. *Psychology of Sport and Exercise*, 10(5): 559–564. <https://doi.org/10.1016/j.psychsport.2009.01.001>.

WANG, Y., CHEN, X., GONG, J., YAN, Y. (2016). Relationships Between Stress, Negative Emotions, Resilience, and Smoking: Testing a Moderated Mediation Model [online]. *Substance Use & Misuse*, 51(4): 427–438. <https://doi.org/10.3109/10826084.2015.1110176>.

WANG, Z., MA, D., SUN, D., ZHANG, J. (2021). Identification and analysis of urban functional area in Hangzhou based on OSM and POI data [online]. *PLOS ONE*, 16(5): e0251988. <https://doi.org/10.1371/journal.pone.0251988>.

ZHANG, P., CHEN, P.-L., LI, Z.-H., ZHANG, A., ZHANG, X.-R., ZHANG, Y.-J., MAO, C. et al. (2022). Association of smoking and polygenic risk with the incidence of lung cancer: A prospective cohort study [online]. *British Journal of Cancer*, 126(11): 1637–1646. <https://doi.org/10.1038/s41416-022-01736-3>.

ZHAO, S. (2022). A Study on China's Tobacco Taxation and Its Influencing Factor on Economic Growth [online]. *Frontiers in Psychology*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2022.832040>.

ZHONG, S., WANG, M., ZHU, Y., CHEN, Z., HUANG, X. (2022). Urban expansion and the urban–rural income gap: Empirical evidence from China [online]. *Cities*, 129: 103831. <https://doi.org/10.1016/j.cities.2022.103831>.

Self-Employment Dynamics amid High Unemployment: An Empirical Analysis of Transitional Labour Market in India

Sonu Madan[1] | *Chaudhary Bansi Lal University, Bhiwani, India*
Surender Mor[2] | *BPS Women University, Sonepat, India*

Received 24.5.2023 (revision received 26.8.2023), **Accepted** (reviewed) 3.10.2023, **Published** 15.3.2024

Abstract

There is a high unemployment rate in India, and self-employment is considered a way to create more jobs and stimulate the economy. However, personal and family characteristics and occupational structure play a role in determining who is likely to pursue self-employment. We conducted a study to assess the probability of Indian youth choosing self-employment over regular-wage employment. Our findings show that females living in rural areas and married individuals are more likely to choose self-employment than their male, urban, and unmarried counterparts. Education is also a factor, as those with higher education tend to seek regular-wage employment. Certain occupations, such as technicians and clerical support workers, are less likely to pursue self-employment. To promote self-employment among youth, there should be a focus on enhancing entrepreneurial skills and capabilities.

Keywords	***DOI***	***JEL code***
Regular wage employment, self-employment, labour market state, youth	*https://doi.org/10.54694/stat.2023.22*	*J13, J24, J61*

INTRODUCTION

Labour market dynamics are indicators of changes in jobs caused by structural changes in the labour market. They are affected by employment, departures and the establishment and shutting down of self-employment activities. Labour market dynamics experience enters and deviations from economic activities leading to structural changes therein. The current high level of unemployment is not merely a result

[1] Department of Economics, Chaudhary Bansi Lal University, 127021 Bhiwani, Haryana, India. Corresponding author: e-mail: sonu.economics@igu.ac.in.
[2] BPS Women University, 131305 Sonepat, Haryana, India.

of the pandemic, as the economy has already been struggling with a high unemployment rate of 6.1 per cent even before the pandemic. A Periodic Labour Force Survey of July 2017–June 2018 reveals the bleak truth of leaving the labour force as nearly 20 per cent of unemployed individuals leave it due to their inability to find any job (Xia, 2021). This is one of the main reasons that the self-employed labour force in India constitutes 52.1 per cent of the total workforce. The troops to move up into self-employment, over regular wage workers, as a job option defines the transition rates between labour market states (Alaniz et al., 2019; Fiaschi and Tealdi, 2021). In India, the youth population accounts for almost 27 to 28 per cent of the total population, comprising individuals aged between 15 to 29 years. Unfortunately, the unemployment rate among male youth in rural areas is 17.4 per cent, while in urban areas, it stands at 18.7 per cent. Similarly, female youth in urban areas face an unemployment rate of 13.6 per cent, while in rural areas, it goes up to 27.2 per cent. These figures are based on the Periodic Labour Force Survey conducted from July 2017 to June 2018.

Despite economic and social uncertainty and the lack of decent employment opportunities, young people are considered an asset to society worldwide. They continuously explore innovative means to create employment opportunities and seek professional growth. They are highly productive workers and entrepreneurs who contribute significantly as agents of change and innovation. In today's era of technological advancement, they have the potential to make significant contributions to economic activities as productive workers and entrepreneurs.

The question of whether young people should pursue economic activities as self-employed is an important one. Studies by Alaniz et al. (2019), and Oleneva (2020) suggest that pursuing self-employment can be an effective means for young people to earn a livelihood and grow gradually. It can also allow them to apply their creativity and innovation to build successful businesses.

The issue of youth unemployment is currently receiving a lot of attention from policymakers and social activists in India due to its significant and lasting impact on employment and earning opportunities. Research conducted globally has demonstrated that self-employment is not only an efficient way to tackle the problem of unemployment but also a viable means of enabling young people to explore their abilities and interests as productive economic agents. By engaging in self-employment, unemployed youth may overcome the demands of the labour market and bring dynamism to the job market.

The role of education has also been changing from producing merely educated job seekers to trained self-sufficient professionals to act as self-employed or entrepreneurs and has significantly affected wage levels in the Indian labour market (Robinson and Sexton, 1994; Dawson et al., 2014). Mor et al. (2020) indicate that attaining higher education decreases the chances of long-term survival in the micro-business world, as it opens more livelihood options, leading to exploration and experimentation of other areas of work engagement. While considering similar studies globally, Van Stel and Van der Zwan (2020) used macro-level data for 32 European countries to reveal that the number of solo self-employed workers, especially highly educated, has increased considerably for females.

So far as India is concerned, a significant chunk, i.e. 41.49 per cent of the Indian labour force, is in the agriculture sector contributing only 19.9 per cent to GDP (Government of India, 2021), which provides a sufficient reason to believe that the majority of workers in the agriculture sector are disguised unemployed resulting in an oversupply of unskilled labourers therein. A similar situation can be tracked in other countries; for instance, there is a strong evidence of a segmented self-employed sector in Argentina that co-exists with disguised unemployment and traditional entrepreneurial activities (Mandelman and Gabriel, 2009). In European countries, self-employment is usually involuntary (Hyytinen and Rouvinen, 2008) and is considered a job option to meet necessities and growth requirements (Desai, 2009). Expanding self-employment opportunities beyond the agricultural sector is crucial for bolstering productivity. Furthermore, the need for more skills that meet market demands and rigid employment protection laws negatively impacts job creation and overall revenue (Bertola et al., 1999; Blanchard and Portugal, 2001; Vindigni et al.,

2015). Woldetensay (2012) also found that formal education, family background, access to funding, risk-taking ability, and gender significantly influence the self-employment in Ethiopia's labour market.

This makes us investigate the motivation behind moving up of labour force into self-employed job option over wage employment for the analytical solution of problem of unemployment, under employment and disguised unemployment. Herein, personal, occupational, and family related factors need to be explored to investigate their strength for making the labour force for moving up into self-employed labour market state and to suggest a pathway ahead leading to solution.

1 REVIEW OF LITERATURE

1.1 Education and self-employment state of labour market

Investing in education can help individuals make informed choices when deciding between self-employment and regular-wage employment for their livelihood and well-being. Those with higher levels of schooling possess more skills, knowledge, and expertise, which can lead to better-paying job opportunities. Studies by Uusitalo (2001) and You (1995) have shown that individuals with higher levels of education typically earn more in regular-wage employment than in self-employment, which can be riskier due to the unstable nature of income. Kangasharju and Pekkala (2002) found that highly educated individuals are less likely to choose self-employment due to lower earning potential and less stable income stream. However, Robinson and Sexton (1994) argued that self-employment could be a form of entrepreneurship and that self-employed individuals have more formal education than those in regular wage employment. Recent studies by Mitze and Javakhishvili-Larsen (2020) and Stokke (2021) suggest that higher education can lead to better earning potential and job mobility.

A study conducted by Madan and Mor (2020) discovered that individuals who obtained an additional year of formal education experienced an average increase of 7 per cent in their wages. However, it is essential to note that this increase varies across different occupations. Specifically, managers can expect a 6 per cent increase, professionals can expect a 4.6 per cent increase, and technicians and associate professionals can expect a 4.3 per cent increase, compared to those in elementary occupations. While self-employment may offer higher pay for top earners, research conducted by Saridakis et al. (2019) and Pantea (2020) reveals that those who fall below the median of the earnings distribution in both Eastern and Western Europe tend to earn considerably less. This emphasizes the significance of exploring occupational diversity when considering the transition from wage employment to self-employment, particularly for young workers. In view of the above, we propose to test the undermentioned hypothesis:
H_0*: Moving up into self-employment labour market state is independent of formal education.*

1.2 Nativity of workforce and self-employment state of labour market

In today's fast-paced and globalized world, the opportunities available to young workers are vast and varied. The origin of these workers, whether from urban or rural areas, can significantly impact their career choices. It may shape their decision on whether to pursue self-employment or opt for wage employment. Research conducted by Faggio and Silva (2014) has uncovered that rural workers tend to choose self-employment over wage employment in areas where the labour market is limited. This research also found a strong correlation between self-employment, innovation, and business creation.

Conversely, Wang and Yang's (2013) study found no significant difference in the choice between self-employment and wage employment for non-agricultural activities in rural China among those who migrated back from urban areas. The nativity of workers plays a vital role in shaping their mindset and decision-making process and should be considered when evaluating their labour market preferences. To test this idea further, we propose the following hypothesis:
H_0: *The nativity of young workers does not significantly impact their decision to pursue self-employment as a labour market state.*

1.3 Marital status and self-employment state of labour market

When making advancements in one's career, there are many factors to consider. One of the most significant considerations is marital status. Marriage is deeply ingrained in society and can substantially impact various aspects of an individual's work life. This includes earning potential, skill development, immigration status, time horizon, investment opportunities, work motivation, and the willingness to accept well-paying but unpleasant job offers. Surprisingly, research has shown that the effect of marriage on wages can indirectly lead to an increased work commitment among male employees (Ahituv and Lerman, 2005). Meanwhile, for women, their marital status can significantly influence their behaviour and outcomes in the labour market. Studies in similar lines (Hamid, 1991; Muller and Posel, 2008; Ntuli, 2007; Yakubu, 2010) have revealed that married women are less likely to be employed than their unmarried counterparts, while married men are more likely to be used. To further explore this notion, we hypothesize that the *preference for self-employment in the labour market is unrelated to one's marital status.*

1.4 Occupational structure and self-employment state of labour market

As individuals contemplate advancing in the labour market, they are influenced by many factors that can impact their decision-making process. One significant consideration is the availability of opportunities in various occupational categories, including professional, technical, clerical, service and sales, agricultural, forestry and fishery work, craftsmanship, machine operating and assembling, and elementary work. Additionally, *gender, education, nativity*, and *marital status* of the labour force can all play a crucial role in deciding whether to pursue self-employment or wage employment (CEA et al., 2008).

It is worth noting that many self-employed workers operate outside fixed establishments, with a substantial proportion involved in transportation, wholesale, and retail trades, compared to other private sector workers (Reham and Salemi, 2019). Furthermore, Madan and Mor (2020) discovered that managerial, professional, technical, and associate professionals' workers who lack specific operational skills are more easily replaceable, resulting in lower wages.

Given these findings, it may be worthwhile to explore the possibility that the preference for self-employment across different occupational groups is uniform. As such, we propose testing the following hypothesis:

H_0: *The tendency for self-employment in various occupational categories is consistent and not influenced by factors such as gender, education, nativity,* or *marital status*.

1.5 Gender and self-employment state of labour market

There is a gradual shift in the labour market where women who have traditionally worked in low-productivity sectors with low pay are transitioning towards higher-productivity, modern jobs. However, despite this transition, the representation of women in such positions still needs improvement. The gender-based segregation of professions remains a prominent socio-structural feature of the labour market, where women participation is relatively lower than men. Research conducted by Correll in 2001 has shown that men and women tend to have different job preferences. This is reflected in the fact that women are less inclined to pursue careers in finance and consulting and are more interested in general management positions, as suggested by Barbulescu and Bidwell in 2013. Furthermore, there is a persistent wage gap between male and female workers, with male workers earning more than their female counterparts, as supported by studies conducted by Madan and Mor in 2021, Reshid in 2019, Ara in 2016, and Manning & Swaffield in 2008. Studies show that only 20–26% of the gender wage gap in early-career private-sector jobs can be attributed to gender, education, employer qualities, and mobility. This means that a significant portion of the gap remains unexplained (Napari, 2008). Additionally, data indicate that older men predominantly choose self-employment compared to other demographics (Reham and Salemi, 2019). Given these insights, we aim to test the undermentioned hypothesis:

H_0: *Gender does not significantly impact the preference for self-employment.*

2 CONCEPTUAL FRAMEWORK AND HYPOTHESIS FORMULATION

In today's job market, there are numerous avenues for individuals seeking employment, from traditional salary-based jobs to the increasingly popular option of self-employment. With such a diverse range of opportunities, individuals can explore and discover the best fit for their unique skills and interests. Unfortunately, for some, limited job options can lead to the challenge of unemployment, which can result in significant economic struggles (Blackburn and Mann, 1979).

In today's economy, self-employment has emerged as a significant contributor to job creation and the promotion of entrepreneurship. According to Mazzarol (1999), it allows individuals to establish their businesses and become self-reliant, thereby promoting economic stability and growth. Additionally, Woldetensay (2012) has noted that self-employment can be an asset for profitable entrepreneurial ventures. The ability to work for oneself and pursue one's passions and interests can lead to greater personal fulfilment and financial success. Overall, self-employment is a valuable avenue for those seeking to impact the economy and their own lives positively (Pardo and Jaime, 2011).

A critical goal of any society should be to prepare young people with the skills and knowledge they need to become successful self-stand workers rather than just job seekers. By creating and promoting opportunities for self-employment, we can help to achieve this goal and empower the next generation of entrepreneurs and business leaders (Mor, Madan and Chikhara, 2020).

This investigation aims to delve into the likelihood of young individuals opting for self-employment as opposed to traditional salaried jobs. Through this study, we aim to analyze the impact of various personal, family, and occupational diversity factors on the decision to pursue self-employment. Our primary objective is to understand better the key drivers behind this career choice among young people and identify any potential barriers or challenges that may impact their decision-making process. This research aligns with earlier studies conducted by Filipovich (1997), Patton and Creed (2001), Hartung et al. (2005), Ahituv and Lerman (2005), Barbulescu and Bidwell (2013), Wang and Yang (2013), Vindigni et al. (2015), and Punia (2020).

3 MATERIAL AND METHODS

3.1 Database

Principal data source of this study is the unit-level data obtained from Periodic Labour Force Survey, 2017–18 (PLFS) launched by National Statistical Office (NSO) in 2019. The survey covered 102 113 household and the total of 433 339 workers. Among all, 19 220 workers are in the age group of 15–29 are years. After data cleaning, a sample of 19 038 youth workers are considered to capture the predictive power of relevant factors to while opting for self-employment over regular wage employment. Herein, keeping in view relevant related studies, personal characteristics i.e. education and gender of workers (Ferrara et al., 2018; Dibeh et al., 2019), family related characteristics i.e. marital status and nativity of workers, (Ferrara et al., 2018; Msigwa and Kipesha, 2013), and occupational structure related characteristics (Todd, 2005) have been considered as important influential factors. Occupational structure has been considered in accordance with the International Standard Classification of Occupations-08 (ILO, 2012).

3.2 Description of variables

3.2.1 Dependent variable

State of labour market is considered as dependent variable with two broad work options, *self-employed* and *wage workers*, and is used as a proxy to estimate the probability of youth to opt for *self-employed* work option over *wage employment* in the labour market. A binomial variable is assigned value 1, if *self-employment work* option is opted over wage employment for which 0 value is assigned. Herein, wage employment is considered as a reference category to estimate the probability of opting *self-employment*, which is considered as a response/outcome state of labour market.

3.2.2 Independent variable(s)

Opting for any of the two broad states of labour market, is affected by several factors such as *education, structure of occupation, nativity, marital status,* and *gender* of a work aspirant in the labour market. To estimate the predictive power of these, education is considered as a continuous variable, and rest of the factors are considered as categorical variables with the assignment of value 1 to the category whose effect is to be estimated over the other category, which is assigned the 0 value (Table A1; Figure A1).

3.3 Estimation technique and model specification

The study employs binary logistic regression model to estimate the probability of opting self-employment in tune with research in the similar line (Msigwa and Kipesha, 2013; Ferrara et al., 2018; Dibeh et al., 2019). Purposefully, two broad states of labour market, i.e., *self-employment* and *regular wage employment* are considered keeping *self-employment* state of labour market as response/outcome category and *regular wage employment* as reference category. Influential factors such as demographic characteristics of work aspirants in labour market and occupational structure have been considered as predictors. The probability to opt for self-employment over paid worker is expressed as odds, which are transformed into log odds, logits, and the natural log of the odds. These transformations solve the problems that OLS regression faces when applied to data where the dependent variable is binary or categorical. Formally, the model follows as:

$$P(WS_{SE}) = \frac{1}{1+e^{-(\alpha+\beta 1X1+\beta 2X2+\ldots+\beta nXn)}} = \frac{1}{1+e^{-Zi}}, \quad (1)$$

here: $P(WS_{SE})$ = the probability of working as *self-employed*; and X_is = predictors used in the model; e is the natural logarithm. For ease of explanation:

$$P(WS_{SE}) = \frac{1}{1+e^{-Zi}} = \frac{e^{Zi}}{1+e^{Zi}}, \quad (2)$$

$P(WS_{SE})$ lies in between 0 to 1 and is non-linearly related to z_i which ranges from –α to α. From Formula (2), odds ratio of *self-employment* over *regular wage employment* can be estimated and defined as in Formula (3):

$$P(WS_{SE}) = \frac{1}{1+e^{Zi}}. \quad (3)$$

From Formulas (2) and (3), odds ratio of emerging as self-employed youth can be estimated and can be defined as in Formula (4). Specifically, odds of emerging as self-employed can be defined as the probability of opting *self-employment* over probability of opting for *wage employment*.

$$\frac{P(WSSE)}{1-P(WSSE)} = \text{Exp}(\beta) = \frac{1+e^{Zi}}{1+e^{-Zi}} = e^{Zi}, \quad (4)$$

$$\text{Logit (Log Odds)} = \text{Ln}\left(\frac{P(WSSE)}{1-P(WSSE)}\right) = a_m + \Sigma_{k-1}^{k}\beta_{m_k}x_{ik} = Z_{\text{mi}}. \quad (5)$$

Formula (5) explains the probability of moving up into *self- employment* labour market state over *regular wage employment*.

Model Assessment: The binary logistic regression employs a likelihood ratio test to provide a basis to find the optimal coefficient values and to assess the fit of the model. It is based on –2 log likelihood (–2LL) ratio. More precisely, it indicates the unexplained variation in the model and in this concern, larger value of log-likelihood indicates poorly fitted model.

Nagelkerke measure of R^2 (R^2_N) has been applied to ensure statistical significance of coefficient's (β's) associated with predictors to analyse significant contribution towards the prediction of the outcome. This measure can be interpreted in the same way as r^2 in the linear regression to analyse the significance of the model and is based on the log-likelihood of the new model with inclusion of predictors and the log-likelihood (LL) of the baseline/original model and the sample size (n).

$$R^2N = \frac{R^2_{cS}}{1 - e^{\left[\frac{2(LL(baseline))}{n}\right]}}, \tag{6}$$

$$\text{here: } R^2_{cS} = 1 - e^{\left[\frac{-2(LL(new) - (LL(baseline))}{n}\right]}. \tag{7}$$

Cox and Snell's measure of R^2 (R^2_{CS}), statistically never reaches its theoretical maximum value and Nagelkerke measure of R^2 amended it to overcome the limitation of R^2_{CS}. The parameters of logistic regression are estimated using the computer program SPSS (version 16.0).

4 RESULTS AND DISCUSSION

Opting to become *self-employed* rather than pursuing conventional employment can significantly impact the labour market. This decision can influence the number of entrepreneurial ventures, the rate of unemployment, and the advancement of novel and innovative business practices. In the upcoming sections, we will delve into the estimated parameters of a binary logistic regression model that scrutinizes how young people decide between self-employment and traditional employment by meticulously considering the pros and cons of both options.

4.1 Model fitting evaluation

Table 1 displays the log-likelihood value for the baseline model, which is denoted by 24 732.525. This model was able to accurately predict 64.7% of workers as regular wage employees by including only the constant. However, after the inclusion of all predictors and an intervention, the –2LL value decreased significantly to 17 763.752, which confirms that the model is an excellent fit. The Omnibus Tests of Model Coefficients and Chi-square statistics at 12 DOF ($\chi 212 = 6\ 968.773$; $p < .01$) indicate that there is a noteworthy difference between the model with predictors and the model with only the constant. The Nagelkerke R2 value is 42%, which shows that the predictive power of the model is substantial.

4.2 Education and self-employment state of labour market

To estimate the parameters in Table 1, the wage employment has been considered as the reference category while providing estimates for opting self-employment state of the labour market by youth workers in the labour market. The results uncovered that *educational achievements* play a significant role in the decision-making process of whether to opt for self-employment ($\beta = -.094$, Wald $\chi 2(1) = 301.184$, $p < .01$). The statistics presented in Table 1 indicate that for every additional year of formal education, the odds of selecting *wage employment* over *self-employment* increase by a factor of 1.09 (1/0.911).

This suggests that education is a critical factor in developing crucial cognitive, managerial, and technical skills necessary for skilled, growth-oriented, and innovative jobs. Additionally, higher levels of education not only increase the chances of finding wage employment and offer more significant promotional opportunities. Our results align with previous studies (Storey, 1994; You, 1995; Uusitalo, 2001; Kangasharju and Pekkala, 2002), which propose that higher education reduces the probability of exploring *self-employment* opportunities and encourages youth to seek *wage employment*. This seems

to correspond with current realities as wage employment for majority of work positions requires educational attainments which are defined in accordance with job profile of work aspirants. Contrarily, own account workers need enthusiasm and zeal to begin and continue with their work. As a result, we reject our initial null hypothesis and conclude that education is a significant motivating factor for transitioning into the self-employment state of the labour market.

4.3 Nativity and self-employment state of labour market

Nativity is also a significant predictor of a move of youth into *self-employment* job option rather than *wage employment*, $\beta = .670$, Wald $\chi2(1) = 312.029$, $p < .01$. It indicates that rural workers are more likely to become *self-employed* than urban workers. The odds of choosing *self-employment* compared to salaried employment is 1.95, showing rural youth are likelier to choose self-employment than wage employment. The reason is obvious to understand. The rural economy cannot provide enough paid work opportunities. At the same time, it has many work opportunities, especially in agriculture and allied activities, to opt for their livelihood. The indulgence in these activities make them to maintain their livelihoods at their native place. Put differently, sizable self-employment opportunities in divergent areas, such as farming, dairy, poultry, animal husbandry etc., are available in the rural economy of India. With this, self-employment is a convenient livelihood source for rural youth. In the light of the estimated impact of nativity on the choice of work status by youth, our maintained null hypothesis of the insignificant role of nativity in explaining the move up of youth into a self-employment labour market state is refuted.

4.4 Marital status and self-employment state of labour market

The *marital status* of young people affects their decision to work as self-employed individuals instead of regular-wage employees. This is supported by a beta value of –.543 and a Wald $\chi2(1)$ value of 208.899, with a significance level of $p < .01$. As a young person's marital status changes from being never married to currently married, their odds of choosing self-employment over salaried employment decreases by .540. This means that currently married individuals are 1.72 times more likely to choose self-employment than wage employment.

Thes finding suggests that when contemplating the prospect of self-employment, it is essential to note that unmarried individuals may encounter certain demographic and financial limitations that could pose a challenge. These limitations may include needing more necessary resources, limited access to business networks, and reduced social support systems. On the other hand, married individuals may view self-employment as a desirable option due to the unique demands of their family and child-care responsibilities, as well as the flexibility they can afford during the early stages of marriage. Previous research conducted by esteemed researchers such as Presser (1995), Casper and O'Connell (1998), and Bianchi (2000) has consistently demonstrated this trend.

Despite the potential advantages of self-employment for married individuals, the decision to pursue this path should be made carefully considering one's personal and professional goals and potential risks and challenges. Ultimately, everyone's circumstances and priorities will play a significant role in determining whether self-employment is viable. Hereby, we reject the hypothesis, which argues that marital status does not impact the preference for self-employment in the labour market.

4.5 Occupational structure and self-employment state of labour market

When young individuals embark on the journey of selecting a career path, they often lean towards a particular occupational category that aligns with their interests and skills. Upon exploring the various options available, self-employment emerges as a popular choice among youth in several occupational groups. These groups include managers, professionals, service, and sales workers, skilled agricultural, forestry

and fishery workers, craft and related trade workers, plant and machine operators, and assemblers. However, it is worth noting that elementary workers are less inclined to choose self-employment as a career path.

Present study has shed light on the fact that individuals who hold managerial or related positions and those who work in skilled agricultural, forestry, and fishery jobs tend to have a greater inclination towards self-employment rather than wage employment. The probability of choosing self-employment is significantly higher for those in management or related roles (with a β value of 2.658, Wald $\chi2(1) = 903.938$, $p < .01$). At the same time, it is even more pronounced for those in skilled agricultural, forestry, and fishery jobs (with a β value of 4.097, Wald $\chi2(1) = 1\ 010.884$, $p < .01$). This points towards a clear attraction towards self-employment opportunities in these fields among young people, as compared to those in elementary jobs.

These findings imply that self-employment within these occupational categories holds promising prospects, providing young people a distinct advantage over traditional wage employment. In addition, these occupations are often associated with greater autonomy and control over one's work, which may further increase the appeal of self-employment in these fields. These results have significant implications for policymakers and educators responsible for shaping the future workforce and highlight the importance of promoting self-employment opportunities in agriculture, forestry, and fisheries.

Similarly, the odds of choosing self-employment as professionals, service & sales workers, craft & related trade work and plant & machine operators and assemblers are 1.503 ($\beta = .408$, Wald $\chi2(1) = 19.422$, $p < .01$), 1.696 ($\beta = .529$, Wald $\chi2(1) = 58.086$, $p < .01$), 2.674 ($\beta = .983$, Wald $\chi2(1) = 204.537$, $p < .01$) & 1.526 ($\beta = .423$, Wald $\chi2(1) = 32.857$, $p < .01$), indicate that the youth is more likely to choose self-employment in these occupational categories in comparison to wage employment.

Some occupations, such as technicians, associate professionals, and clerical support workers, prefer regular wage work over self-employment. The odds of choosing self-employment for these occupations are .483 ($\beta = -.524$, Wald $\chi2(1) = 25.198$, $p < .01$) & 0.034 ($\beta = -2.788$, Wald $\chi2(1) = 80.829$, $p < .01$) respectively. This means that Indian youths are 2.07 times and 29.41 times more likely to choose regular wage employment over self-employment in these occupations than elementary workers (Table 1). Thus, our findings suggest that self-employment is not equally preferred across all occupational groups. This contradicts the 4th upheld hypothesis of uniformity of preference for self-employment labour market states across divergent occupational groups.

4.6 Gender of youth and self-employment state of labour market

The findings in Table 1 indicate that an individual's gender plays a significant role in predicting their preference for self-employment versus regular wage/salary employment. The regression coefficient (β) of –0.103 and Wald $\chi2(1)$ value of 3.885, with a significance level of $p < .001$, demonstrate that gender is a statistically significant predictor of self-employment preference.

Further analysis using odds ratios reveals that males are 0.902 times less likely than females to choose self-employment over salaried employment. It has been observed that women are more inclined towards self-employment as compared to men. The underlying reason for this preference difference could be attributed to various social barriers women face, such as household and childcare responsibilities. Research have shown that such duties often act as constraints that limit women's participation in the workforce or restrict their ability to seek full-time employment outside their homes (Presser, 1995; Casper and O'Connell, 1998; Bianchi, 2000; Yee, 2007; ILO, 2015).

Moreover, disparities in education and training and limited access to financial resources and networks can negatively impact female employment and earning potential, further pushing them towards considering self-employment as a viable alternative, as per Gangel and Ziefle's (2009) findings. It is worth noting that women who become self-employed often face unique challenges not encountered by their male

counterparts. These challenges include access to capital, market opportunities, and business networks, which can adversely impact their chances of success in entrepreneurship. Hence, addressing these challenges and providing women with the necessary support to achieve their entrepreneurial goals is essential. In view of this, *the fifth hypothesis that gender is unrelated to preferences for self-employment is not supported by the findings of this study.* This makes clear that social constraints and different levels of human capital endowments are major factors that lead women to choose self-employment over men. Policymakers and practitioners need to keep these findings in mind when creating strategies to support women's entrepreneurship and promote gender equality. By reducing these disparities, we can create a more equitable environment for entrepreneurs and encourage a more diverse and inclusive business community.

Table 1 Binary regression analysis of work status and demographic factors

Predictors	**β**	**S.E.**	**Wald**	**Exp(β)**
Gender (male)	–.103**	.052	3.88	.902
Nativity (rural)	.670*	.038	312.03	1.954
Marital status (never married)	–.543*	.038	208.90	.581
Educational attainments (in years)	–.094*	.005	301.18	.911
Occupation[a]			2 471.28	
Managers (i)	2.66*	.088	903.94	14.274
Professionals (ii)	.408*	.093	19.42	1.503
Technicians and associate professionals (iii)	–.524*	.104	25.20	.592
Clerical support workers (iv)	–2.79*	.310	80.83	.062
Service and sales workers (v)	.529*	.069	58.09	1.696
Skilled agricultural, forestry and fishery workers (vi)	4.10*	.129	1 010.88	60.154
Craft and related trade workers (vii)	.983*	.069	204.54	2.674
Plant & machine operators and assemblers (viii)	.423*	.074	32.86	1.526
Constant	–.389*	.085	20.66	.678
Correctly predicted cases (at step 0)	64.7 percent			
Correctly predicted cases (at step 1)	77.8 percent			
Model fitting criteria: –2 log likelihood	–			
Intercept only (initial –2 log likelihood)	24 732.525 (estimated at iteration number 3 because parameter estimates changed by less than .001)			
Final model	17 763.752 (estimated at iteration number 7 because parameter estimates changed by less than .001)			
Omnibus tests of model coefficients ($\chi 2_{12}$)	6 968.773*			
Nagelkerke R^2	.422			

Notes: * indicates significant at 1 percent level of significance, ** indicates significant at 5 percent level of significance, a detailed description about occupational groups is given in endnotes.
Source: Authors

CONCLUSIONS

This research delves into the multifaceted factors that influence the decision of young adults to pursue self-employment as opposed to traditional wage employment in India. The study identifies personal, family, and occupational factors as key influencers in this decision-making process. Specifically, the research finds that higher levels of education, a non-agricultural upbringing, non-traditional gender roles, and a more gender-equal labour market make individuals more inclined to opt for wage employment. Conversely, a lack of access to education and skills training, a reliance on the agricultural economy, family and childcare responsibilities, and a need for flexible work hours are drivers of self-employment. Furthermore, the study highlights those certain professions, such as sales, service, crafts, and machine operation, are more likely to attract young people to self-employment due to their alignment with personal preferences. These findings underscore the importance of addressing the systemic barriers that limit individuals' access to education and skills training, as well as the need for more flexible and inclusive labour market policies that better accommodate the needs of workers.

Policy Implications and suggestions for future research

To effectively promote entrepreneurship in India, a multifaceted approach is essential. The first prong of this approach involves fostering an entrepreneurial culture, especially in urban areas. This can be achieved through mentorship programs, networking events, and educational campaigns to promote an entrepreneurial mindset. The second prong of this approach involves providing incentives to encourage young people to pursue entrepreneurship as a viable career option rather than simply out of necessity. These incentives could be tax breaks, subsidies, and grants for start-ups. Moreover, it is important to educate individuals about the various self-employment opportunities available, including the financial and training requirements and the long-term profitability of these ventures. This could be accomplished through workshops, seminars, and online resources that guide starting and managing a successful business.

In addition, future research should focus on longitudinal data and other social and economic factors that may impact the labour market in India. This could include factors such as access to capital, government policies, and cultural attitudes towards entrepreneurship. By better understanding these factors, policymakers can design more effective programs and initiatives to promote entrepreneurship and support the growth of small businesses in India.

References

AHITUV, A., LERMAN, R. I. (2005). How do marital status, wage rates, and work commitment interact? *Urban Institute and IZA Bonn Discussion Paper*, No. 1688.

ALANIZ, E., ESPINO, A., GINDLING, T. H. (2019). *Self-Employment and Labour Market Dynamics of Men and Women in El Salvador and Nicaragua* [online]. WIDER Working Paper 2019/97, Helsinki: UNU-WIDER. <https://doi.org/10.35188/UNU-WIDER/2019/733-0>.

ARA, S. (2016). Gender and jobs: Evidence from urban labour market in India [online]. *The Indian Journal of Labour Economics*, 58(3): 377–403. <https://doi.org/10.1007/s41027-016-0027-2>.

BARBULESCU, R., BIDWELL, M. (2013). Do women choose different jobs from men? Mechanisms of application segregation in the market for managerial workers. *Organization Science*, 24(3): 767–756.

BERTOLA, G., BOERI, T., CAZES, S. (1999). Employment protection and labor market adjustment in OECD countries: Evolving institutions and variable enforcement. *Employment and Training Papers*, No. 48, Geneva: International Labor Organization.

BIANCHI, S. M. (2000). Maternal employment and time with children: Dramatic change or surprising continuity? *Demography*, 37(4): 401–414.

BLACKBURN, R. M., MANN, M. (1979). Stratification within the manual working class [online]. In: *The working class in the labour market, Cambridge studies Sociology*, London: Palgrave Macmillan, 245–276. <https://doi.org/10.1007/978-1-349-16097-6_9>.

BLANCHARD, O., PORTUGAL, P. (2001). What hides behind an unemployment rate: Comparing Portuguese and US unemployment. *American Economic Review*, 91: 187–207.
CASPER, L. M., O'CONNELL, M. (1998). Work, income, the economy, and married fathers as child-care providers. *Demography*, 5(4): 243–250.
CEA, S., CONTRERAS, M. I., MARTINEZ, C., PUENTES, E. (2008). Transiciones laborales: Evidencia para datos de panel, mimeo [online]. *Working paper 3*, Department of Economics, University of Chile. <https://ideas.repec.org/p/udc/wpaper/wp308.html>.
CORRELL, S. J. (2001). Gender and the career choice process: the role of biased self-assessments [online]. *American Journal of Sociology*, 106(6): 1691–1730. <http://www.jstor.org/stable/10.1086/321299>.
DAWSON, C., ANDREW H., PAUL L. L. (2014). Individual Motives for Choosing Self-employment in the UK: Does region matter? *Regional Studies*, 48(5): 804–822.
DESAI, S. (2009). *Measuring Entrepreneurship in Developing Countries*. Working Papers UNU-WIDER Research Paper, World Institute for Development Economic Research.
DIBEH, G., FAKIH, A., MARROUCH, W. (2019). Employment and skill mismatch among youth in Lebanon [online]. *International Journal of Manpower*, 40(8): 1438–1457. <https://doi.org/10.1108/IJM-02-2018-0073>.
FAGGIO, G., SILVA, O. (2014). Self-employment and entrepreneurship in urban and rural labour markets [online]. *Journal of Urban Economics*, 84(11): 67–85. <https://doi.org/10.1016/j.jue.2014.09.001>.
FERRARA, R., GALLO, G., MONTANARI, A. (2018). Young Italians in European countries: Social and economic perspective [online]. *BELGEO*, 3: 1–18. <https://doi.org/10.4000/belgeo.29516>.
FIASCHI, D., TEALDI, C. (2021). *A General Methodology to Measure Labour Market Dynamics*. IZA DP No. 14254, Discussion Paper Series, IZA Institute of Labour Economic.
FILIPOVICH, D. (1997). An alternative approach to the theory of employment: Self-employment versus wage employment. *Economia Mexicana*, *NUEVA EPOCA*, 6(2): 177–195.
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. *Demography*, 46(2): 341–369.
GOVERNMENT OF INDIA. (2021). *Economic Survey 2020–21* [online]. New Delhi: Ministry of Finance, Department of Economic Affairs. <https://www.indiabudget.gov.in/economicsurvey>.
HAMID, S. (1991). Determinants of the supply of women in the labour market: A micro analysis [online]. *The Pakistan Development Review*, 30(4): 755–766. <https://doi.org/10.30541/v30i4IIpp.755-766>.
HARTUNG, P. J., PORFELI, E. J., VONDRACEK, F. W. (2005). Child vocational development: A review and reconsideration [online]. *Journal of Vocational Behavior*, 66(3): 385–419. <https://doi.org/10.1016/j.jvb.2004.05.006>.
HYYTINEN, A., ROUVINEN, P. (2008). The labour market consequences of self-employment spells: European evidence. *Labour Economics*, 15(2): 246–271.
ILO. (2012). *International standard classification of occupations*. Geneva: ILO.
ILO. (2015). *Global Wage Report. Wage and income inequality* [online]. Geneva: ILO. <https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_324678.pdf>.
KANGASHARJU, A., PEKKALA, S. (2002). The role of education in self-employment success in Finland [online]. *Growth and Change*, 33(2): 216–237. <https://doi.org/10.1111/0017-4815.00188>.
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. *International Journal of Economic Policy in Emerging Economies* (in press).
MANDELMAN, F. S., GABRIEL, V. M. S. (2009). Is self-employment and micro-entrepreneurship a desired outcome? *World Development*, 37(12): 1914–1925.
MANNING, A., SWAFFIELD, J. (2008). The gender gap in early-career wage growth [online]. *Economic Journal*, 118(530), 983–1024. <https://doi.org/10.1111/J.1468-0297.2008.02158.X>.
MAZZAROL, T. (1999). Factors influencing small business start-ups. *International Journal of Entrepreneurial Behaviour & Research,* 5(2): 48–63.
MITZE, T., JAVAKHISHVILI-LARSEN, N. (2020). Graduate Migration and Early-career Labour Market Outcomes: Do Education Programs and Qualification Levels Matter? [online]. *Labour*, 34(4): 477–503. <https://doi.org/10.1111/LABR.12177>.
MOR, S., MADAN, S., GEOFFREY R. A., 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>.
MOR, S., MADAN, S., CHHIKARA, 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>.
MSIGWA, R., KIPESHA, E. F. (2013). Determinants of Youth unemployment in Developing Countries: Evidence from Tanzania. *Journal of Economics and Sustainable Development,* 4(14): 67–76.

MULLER, C., POSEL, D., (2008). Is there evidence of a wage penalty to female part-time employment in South Africa? *South African Journal of Economics*, 76(3): 467–479.
NAPARI, S. (2008). The early-career gender wage gap among university graduates in the Finnish private sector [online]. *Labour*, 22(4): 697–733. [cit. 13. 7. 2021]. <https://doi.org/10.1111/J.1467-9914.2008.00429.X>.
NTULI, M. (2007). Exploring gender wage 'discrimination' in South Africa, 1995–2004: A quantitative regression approach. *IPC Working Paper Series*, No. 56, University of Michigan, Ann Arbor, MI.
OLENEVA, K. (2020). Self-Employment as a Viable Solution for Youth Integration in the Labor Market: The Case of Albania. *International Journal of Economics, Commerce, and Management*, 8(9): 376–391.
PANTEA, S. (2020). Self-employment in the EU: quality work, precarious work, or both? [online]. *Small Business Economics*. <https://doi.org/10.1007/s11187-020-00423-y>.
PARDO, C., JAIME R. T. (2011). *Self-employment versus wage-employment: a dynamic choice model and evidence* [online]. <https://www.researchgate.net/publication/254648330_Self-Employment_Versus_Wage-Employment_A_Dynamic_Choice_Model_and_Evidence>.
PATTON, W., CREED, P. A. (2001). Developmental issues in career maturity and career decision status. *Career Development Quarterly*, 49(4): 336–351.
PERIODIC LABOUR FORCE SURVEY. (July 2017–June 2018). *Annual Report*. Government of India, National Statistical Office, Ministry of Statistic and Programme Implementation, India.
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/10.1002/j.2161-0045.2001.tb00961.x>.
PUNIA, K. (2020). *Future of unemployment and the informal sector of India* [online]. New Delhi: Observers research Foundation. [cit. 23.5.2020]. <https://www.orfonline.org/expert-speak/future-of-unemployment-and-the-informal-sector-of-india-63190>.
REHAM, R., SALEMI, C. (2019). *Own account workers in Jordan profile and work characteristics* [online]. <https://doi:10.1093/oso/9780198846079.003.0004>.
RESHID, A. A. (2019). The Gender Gap in Early Career Wage Growth: The Role of Children, Job Mobility, and Occupational Mobility [online]. *Labour*, 33(3): 278–305. <https://doi.org/10.1111/LABR.12148>.
ROBINSON, P. B., SEXTON, E. A. (1994). The effect of education and experience on self-employment success [online]. *Journal of Business Venturing*, 9(2): 141–156. <https://doi.org/10.1016/0883-9026(94)90006-X>.
SARIDAKIS, G., MENDOZA GONZALEZ, M. A., MUÑOZ TORRES, R. I., HAND, C. (2019). Do Self-employment Rates Converge? Evidence from European OECD Countries [online]. *JCMS: Journal of Common Market Studies*, 57(3): 551–562. <https://doi.org/10.1111/jcms.12834>.
STOKKE, H. E. (2021). The gender wage gap and the early-career effect: the role of actual experience and education level [online]. *Labour*, 35(2): 135–162. [cit. 13.7.2021]. <https://doi.org/10.1111/LABR.12191>.
STOREY, D. J. (1994). *Understanding the small business sector*. London: International Thomson Business Press.
TODD, S. (2005). Entering Employment. In: *Young Women, Work, and Family in England 1918–1950* [online]. Oxford University Press, 85–112. <https://doi.org/10.1093/acprof:oso/9780199282753.003.0004>.
UUSITALO, R. (2001). Homo entrepreneurs? *Applied Economics*, 33(13): 1631–38.
VAN STEL, A., VAN DER ZWAN, P. (2020). Analyzing the changing education distributions of solo self-employed workers and employer entrepreneurs in Europe [online]. *Small Business Economics*, 55(2). <https://doi.org/10.1007/s11187-019-00243-9>.
VINDIGNI, A., SCOTTI, S., TEALDI, C. (2015). Uncertainty and the politics of employment protection. *Journal of Labor Economics*, 33: 209–67.
WANG, Z. C., YANG, W. G. (2013). Self-employment or wage-employment? On the occupational choice of return migration in rural China [online]. *China Agricultural Economic Review,* 5(2): 231–247. <https://doi.org/10.1108/17561371311331115>.
WOLDETENSAY, A. (2012). *Factors affecting self-employment decision in Addis Ababa*. Master thesis, Addis Ababa University, Ethiopia.
XIA, S. (2021). *Labour Market Dynamics of Self-employment in China: The Interplay of Institutions and Gender* [online]. Doctoral thesis. <https://doi.org/10.17863/CAM.83792>.
YAKUBU, Y. A. (2010). Factors influencing female labor force participation in South Africa, *The African Statistical Journal*, 11: 85–104.
YEE, K. M. (2007). Work orientation and wives' employment careers. *Work and Occupations*, 34(4): 430–462.
YOU, J. (1995). Small firms in economic theory. *Cambridge Journal of Economics*, 19: 441–462.

ANNEX

Table A1 Description of variables used in the study

Name of variable	Nature of variable*/categories	Label for variables/treatment	Number of observations
State of labour market (response variable)	Nominal: self-employment (i) and regular wage/salaried employment (ii)	1 for self-employment	6 729
		0 for regular wage/salaried employment (reference category)	12 309
Education	Randomized continuous variable	Treated as continuous variable (measured in years of formal education)	19 038
*Structure of occupation**	Categorical nine broad occupational groups: managers[1] (i); professionals[2] (ii); technicians & associate professionals[3] (iii); clerical support workers[4] (iv); service and sales workers[5] (v); skilled agricultural, forestry & fishery workers[6] (vi); craft and related trade workers[7] (vii); plant & machine operators and assemblers[8] (viii); elementary workers[9] (ix)	1 for managers and 0 for any other profession	1 378
		1 for professionals and 0 for any other profession	1 798
		1 for technicians and associate professionals and 0 for any other profession	1 863
		1 for clerical support workers and 0 for any other profession	1 089
		1 for service and sales workers and 0 for any other profession	3 668
		1 for skilled agricultural, forestry and fishery workers and 0 for any other profession	1 900
		1 for craft and related trade workers and 0 for any other profession	3 085
		1 for plant & machine operators and assemblers and 0 for any other profession	2 404
		1 for elementary workers and 0 for any other profession	1 853
Nativity	Nominal: rural (a) and urban (b)	1 for rural workers and	8 510
		0 for urban workers	10 528
Marital status	Nominal: never married (a) and currently married (b)	1 for never married workers and	10 949
		0 currently married workers	8 089
Gender	Nominal: male (a) and female (b)	1 for male workers and	15 536
		0 for female workers	3 502

Notes: * Occupations are classified in accordance with International Standard Classification of Occupations-08 (ILO, 2012).

[1] Chief executives, senior officials, legislators, administrative & commercial managers, production & specialized services managers, hospitality, retail and other services managers.

[2] Science & engineering professionals, health professionals, Teaching professionals, business & administration professionals, Information & communications technology professionals, legal, social & cultural professionals.

[3] Science & engineering associate professionals; health associate professionals; business and administration associate professionals; legal, social, cultural, and related associate professionals; Information and communications technicians.

[4] Occupation as general & keyboard clerks; customer services clerks; numerical & material recording clerks and other clerical support workers.

[5] Personal service workers; sales workers; personal care workers and protective services workers.

[6] Market-oriented skilled agricultural workers; market-oriented skilled forestry, fishery, and hunting workers; subsistence farmers, fishers, hunters & gatherers.

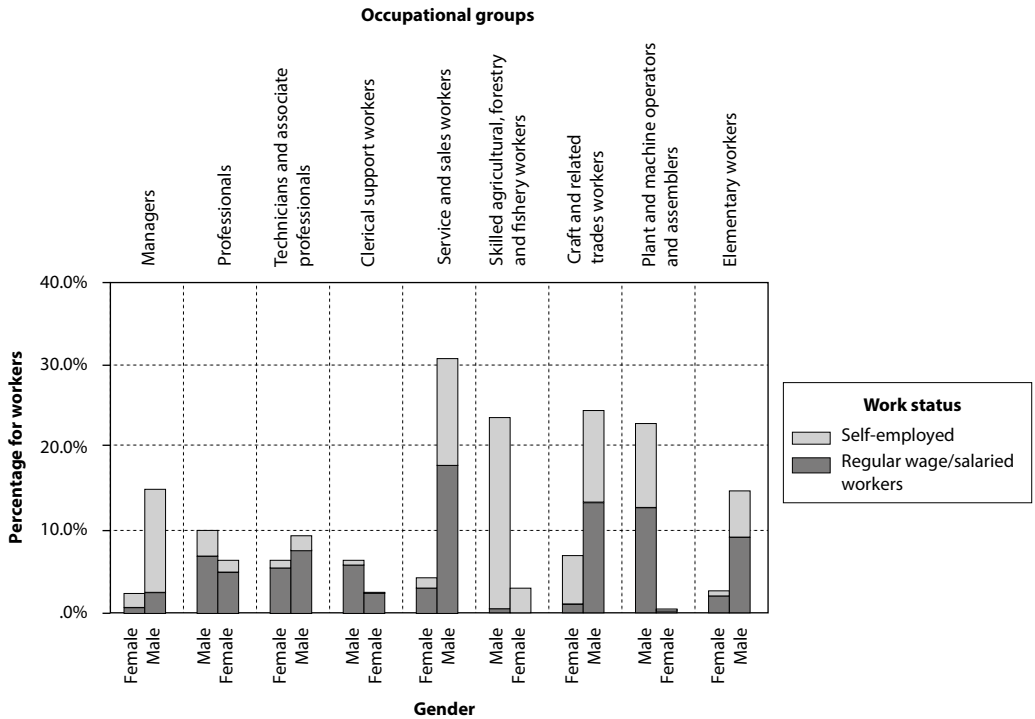
[7] Building and related trades workers, excluding electricians; metal, machinery and related trades workers; handicraft & printing workers; electrical & electronic trades workers; electronics and telecommunications installers and repairers; food processing, wood working, garment and other craft and related trades workers.

[8] Stationary plant & machine operators; Assemblers; drivers and mobile plant operators.

[9] Cleaners & helpers; agricultural, forestry and fishery labourers; labourers in mining, construction, manufacturing, and transport; food preparation assistants; preparation assistants; street and related sales and service workers; refuse workers and other elementary workers.

Source: Periodic Labour Force Survey (July 2017–June 2018)

Figure A1 Distribution of workers as per work status, occupation, nativity and gender



Source: Authors

Consumer Price Index in the Czech Republic – New Sources and Data Processing

Jaroslav Sixta[1] | *Prague University of Economics and Business, Prague, Czech Republic*
Petr Musil[2] | *Prague University of Economics and Business, Prague, Czech Republic*

Received 1.8.2023, **Accepted** (reviewed) 3.11.2023, **Published** 15.3.2024

Abstract

Consumer price index has been in the centre of interest for many years, since being published in 1990s in the Czech Republic but recent price growth raised more questions on methodology and data sources used in price statistics. Users are interested not only in the figures itself but also in statistical issue influencing interpretation and the quality of consumer price index that is often used as an approximation of inflation rate. The paper introduces price statistics compiled by the Czech Statistical Office and it specifically focuses on data sources and in particular scanner data. The paper explains how advanced statistical methods such as machine learning are implemented in official statistical production. We think that the official statistics is being on the historical junction where modern methods are going to be implemented. Our paper shows the usage of machine learning procedures applied on scanner data within consumer price index. Used method is based on logistic regression and powerful Python solution and that provides fast and high quality results.

Keywords	***DOI***	***JEL code***
Consumer price index, scanner data, machine learning, logistic regression	*https://doi.org/10.54694/stat.2023.37*	*C 43, E 31*

INTRODUCTION

Price statistics produces various indicators describing either price development or price level in a given economy. A palette of available indicators differs among countries depending on a level of development of price statistics. A minimum set of indicators is laid down by the EU regulation for EU members. It can be said that the European data set exceeds standards in other countries in terms of coverage, timeliness and frequency. In addition, Eurozone countries are obliged to compile flash estimate of a consumer price index that is being published at the end of a month.

[1] Department of Economic Statistics, Faculty of Informatics and Statistics, Prague University of Economics and Business, W. Curchill Sq. 1938/4, 130 67 Prague, Czech Republic. Corresponding author: e-mail: sixta@vse.cz. The author is a Vice-President of the Czech Statistical Office, Na Padesátém 81, 100 82 Prague 10, Czech Republic.

[2] Department of Economic Statistics, Faculty of Informatics and Statistics, Prague University of Economics and Business, W. Curchill Sq. 1938/4, 130 67 Prague, Czech Republic. E-mail: petr.musil@vse.cz. The author is also working at the Czech Statistical Office, Na Padesátém 81, 100 82 Prague 10, Czech Republic.

Price indices are used for two main purposes: to describe price fluctuation of products or a group of products and to deflate nominal indicators to volume terms (process of statistical deflation). Application of price indices in a deflation process is well described in respective manuals, e.g. Handbook on price and volume measures in national accounts (European Commission, 2016). To this end several types of prices indices are required: Producers price indices (PPIs) measuring price development of supplies at basic prices to a domestic market, prices of foreign trade and derived terms of trade, price indices of real estates and land and consumer price index (CPI). The latest is the most famous and well known by not only statisticians and economists but also general public. CPI enables a comparison of household income and expenditures constituting the main component of GDP in developed countries. We can say, CPI is an approximation of inflation rate and one has to consider its limitations. Last but not least consumer price index is an indicator used for inflation targeting.

Price indices are normally produced and published by National Statistical Institutes (NSIs). Users in the EU benefit from a high level of harmonization laid down in the EU legislation. Harmonised Index of Consumer Prices (HICP) is well-known but other types of price indices are also subject to harmonization. On one hand, price indices are still based on the same principles: Laspeyres formula, monthly frequency for most indices etc. On the other hand, most statistical institutes are creative and have been continuously improving quality of price statistics. For instance, revisions of a weighting scheme are more frequent to capture changes in consumption habits. In addition, new data sources have been acquired and used in the production process. The most valuable source is scanner data that allow us to make a jump in quality. Obviously, scanner data have been processed in different way than data collected within a field survey. Our paper describes a particular detail that is very innovative and significantly improved the quality and possibility of price comparisons, scanner data. The process of construction of CPI based on scanner data cannot be fully automatized up to now but machine learning processes (MLP) radically increased the scope and efficiency of such statistical process. Since 2019, when scanner data were incorporated into statistical production of CPI the amount of data has been continuously rising reaching about 500 thousands records per month. About 10 thousand products are new every month and need to be classified into the classification of consumption by purpose (COICOP). Machine-learning is connected with artificial intelligence (AI).[3] The situation usually stands as the computers are learning from training data and later derived algorithms and parameters that are used for predictions. It is supposed that such algorithms are still improving by both supervised and non-supervised learning. The third existing approach is reinforcement training nearly completely automatic. Up to now, the most prevailing in official statistics in supervised learning. It is obvious that at the beginning, human work is necessary but later on we can use supervised machine-learning processes. The incentives for changes and switchover to partly or fully automatized process grow both from internal or external environment. Most of European statistical office are facing cost reductions and the pressure on efficiency. Optimal allocation of scarce resources – qualified staff is very necessary. Statisticians working for state statistical agencies and offices are usually public clerks, conservative ones but not blind. Enormous spread of modern techniques couldn't stay be overlooked. External environment also determines newcomers from universities with excellent knowledge of modern IT tools such as R and Python programming languages. Time to time, some of these activities could be associated with the term Big Data but it is not our case. The Czech Statistical Office receives twice a month a batch of large amount of pre aggregated data by individual products (later described in a detail), quantities sold and sales from this product. This allows the Czech Statistical Office to compute average prices that completely respects the real demand on particular product, CPI.

[3] Google very nicely explains the connection points and differences, see: <*https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning*>.

The main purpose of the paper is to describe and explain modern statistical methods that have recently been implemented in processing of scanner data. We believe that the methods are inspirational for official statisticians across statistical domains. The first part introduces price statistics with the focus on consumer price index. Next, data sources and data processing are described. Last section is devoted to machine learning based on logistic regression that is applied for data processing.

1 METHODOLOGY

A number of products that are produced and consumed in national economy is finite but hardly countable. For each single product, it is possible to estimate a price index either describing price development over time or comparing prices between regions. However, products are quite heterogeneous, consequently an average price cannot be calculated (e.g. average price of 1 kilogram of vegetable). In order to estimate an 'overall' aggregated index, it is necessary to introduce weights and averaging individual price indices. Laspeyres' formula is applied for the most of price indices. All of published price indices are estimates of 'theoretical' price indices, which are not observable. For example, Cost of living index (CLI) is a theoretical price index for household expenditure defined as 'A ratio that measures impact of price change on consumer well-being' (Fixler, 1993). Laspeyres index using weighting scheme from previous period is upper bound of CLI while Paasche index using weighting scheme from current period is lower bound of CLI (Schultz, 1939). It should be noted that several assumptions need to be held, in particular a typical (decreasing) demand curve that is depicted for instance in Samuelson, Nordhaus (2009). The relationship between Laspeyres and Paasche indices is also described by so-called Bortkiewicz formula (Lippe, 2012). Superlative indices overcome caveats of Laspeyres and Paasche indices by various types of averages of weights (Diewert, 1976).The most famous ones are Fisher, Törnqvist.

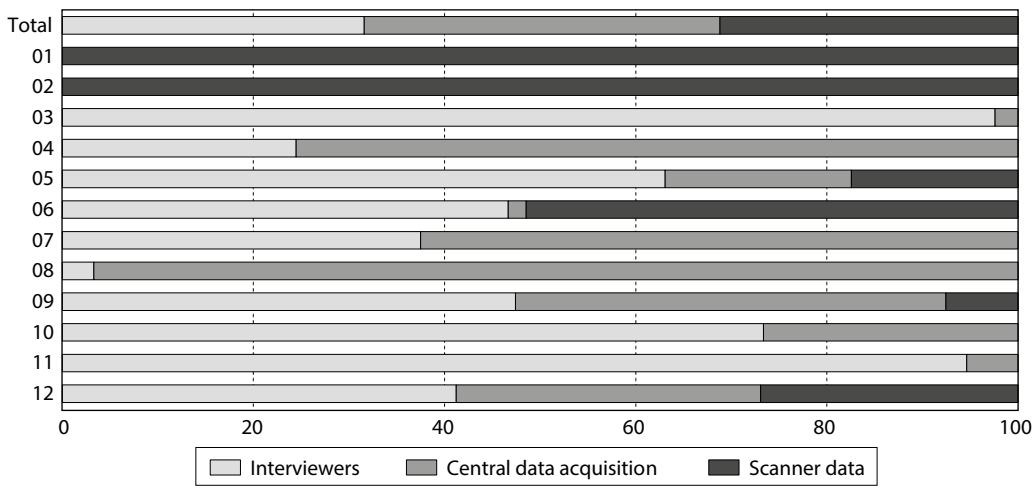
It is well-known that almost all price indices are Laspeyres types. In other words superlative price indices, which are closer to CLI, are not produced and published by National Statistical Institutes. Most price indices including consumer price index are published monthly while weights are based on annual structures. Price statistics is considered as short-term statistics showing emerging trends in national economy. For instance, consumer price index is usually released in the middle of the following month; moreover, flash estimate of HICP is published at the end of a given month. At that time no data source that provides information on products being transacted (produced, consumed) is available. In addition, weights are derived from annual structures in order to eliminate seasonal effect, short-run fluctuation. It does not mean that weights are outdated as regular update is carried out, e.g. HICP weights of ECOICOP categories are updated annually, CPI weights once in 2 or 3 years including detailed breakdown by price representatives. Similarly, weights of other price indices are regularly updated. Consequently, neither Paasche formula nor superlative price indices can be applied.

In practice, the weights are estimated using several data sources. Among them, the most relevant ones are national accounts and household budget survey. National accounts depict comprehensively national economy including the household sector nevertheless the level of detail is not sufficient to derive weights for representatives. As consequence national accounts data adjusted to methodological differences such as non-monetary transactions for instance agriculture self-supply are deployed to estimate weights of COICOP groups. Household budget survey data serve as a supplementary source to derive detailed weights of price representatives within a given COICOP group. It should be noted that household budget survey is completely replaced by scanner data for relevant product groups. In order to keep the weighting scheme relevant they are updated once every two years while the HICP regulation (2016/792) requires annual update. Due to the Covid-19 outbreak and related containment measures that substantially affect a consumer basket the current weights applied from January 2022 are based on household expenditure average in 2019–2021. It is believed that a change in the structure of household expenditure caused by the Covid-19 outbreak is temporary and therefore should not be fully reflected in the weighting scheme.

2 DATA COLLECTION

As mentioned above, all price indices are based on sampling i.e. prices of selected products are surveyed only. Traditionally, prices had been collected in outlets by interviewers. At later stage central data acquisition was introduced especially for products that are supplied centrally e.g. energy, gasoline, communication service. The most modern and sophisticated method is a collection of electronic records ('scanner data'). Scanner data were firstly applied by Dutch Statistical Office in 2002, some other countries started to use this data source in 2010 (Bialek, 2020). The method is triggered by the EU Regulation (2016/792) that lays down obligation of statistical units to provide scanner data to NSIs. The Czech Statistical Office launched voluntary collection of scanner data a couple of years before the Regulation came into force. Since then the collection became binding for retailers whose revenues exceeded a given threshold. The first product group was foodstuff next ones were drug products and drugstore products. Currently, hobby markets have been asked to provide scanner data. It is planned to extend a coverage substantially (Bookstores, Fashion & Clothing retailers) in the following years. The Czech Statistical Office plans to replace a price collection in outlets by scanner for all products for which benefits surpasses costs. Figure 1 depicts methods of data collection used in consumer price index by main COICOP categories.

Figure 1 Data collection methods by COICOP categories



Note: 01 – Food and non-alcoholic beverages, 02 – Alcoholic beverages and tobacco, 03 – Clothing and footwear, 04 – Housing, water, gas electricity and other fuels, 05 – Furnishings, household equipment and routine maintenance of the house, 06 – Health, 07 – Transport, 08 – Communications, 10 – Education, 11 – Restaurants and hotels, 12 – Miscellaneous goods and services.
Source: Own elaboration based on Rojíček and Sixta (2022)

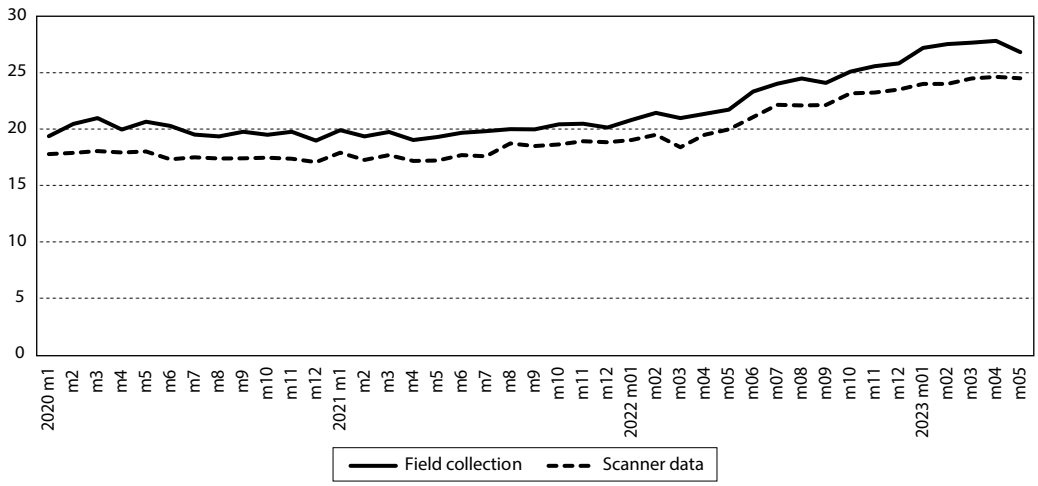
Scanner data represent a substantial improvement in the quality of consumer price index. Firstly, number of products whose prices are surveyed have jumped from hundreds to approximately 100 000. Secondly, actually realized prices instead of spot prices are used. Last but not least, comprehensive data on household expenditure are available and can be deployed in statistics.

Prices of about 700 products of which approx. 150 food products were surveyed before the implementation of scanner data. It should be noted that interviewers collected prices at the moment of collection, i.e. spot prices. Indeed, Nielsen reports that more than 50%[4] of products, in particular

[4] *<https://www.seznamzpravy.cz/clanek/ekonomika-byznys-trendy-analyzy-cesi-jdou-jeste-vic-po-slevach-v-akci-kupuji-uz-60-procent-zbozi-216642>.*

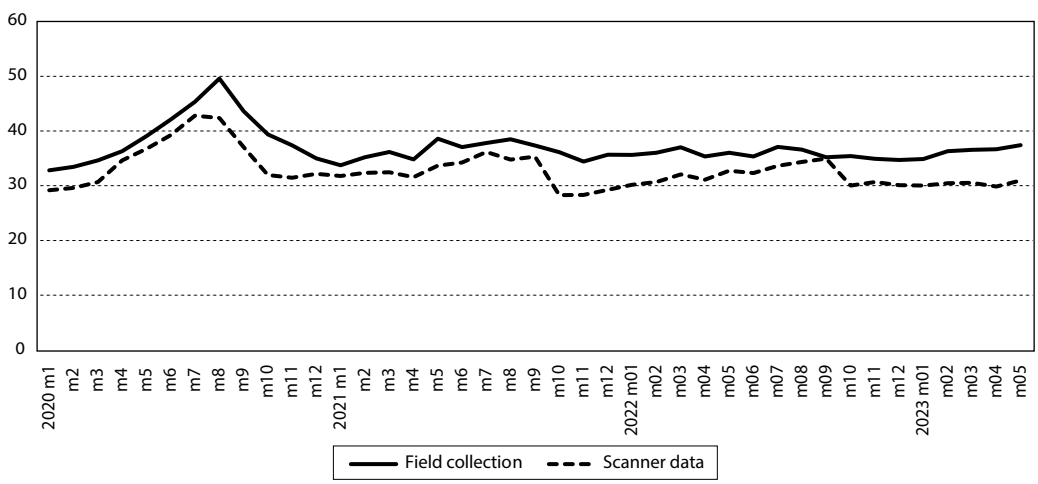
food stuff, are sold in promotion. Spot prices do not necessarily reflect actual (realized) prices affected by promotion. It does not necessarily distort price indices as it can be assumed that price evolution of spot prices and realized prices is the same. On the other hand it may affect international comparison as price level in a country with a higher share of promotion can be overestimated. The Czech Statistical Office stopped field collection of prices once scanner data of products became available. However, there is an exception of weekly survey of food stuff (about 10 items). Users have requested weekly prices for basic food stuff that cannot be gathered from scanner data. Actual prices from scanner data are higher than prices collected in shops by interviewers for all items but prices evolution does not necessarily vary. A difference in average price of semi skimmed milk can be seen stable over time while apple prices indicate also dissimilar evolution in certain time period.

Figure 2 Price of semi skimmed milk (CZK/liter)



Source: Own elaboration

Figure 3 Price of apples (CZK/kg)



Source: Own elaboration

On one hand, implementation of scanner data cuts of costs on a price collection in outlets. On the other hand, data protection and processing require additional ICT costs and highly qualified staff. Processing of scanner data poses a challenge to National Statistical Institutes. Statistical units (retailers) provide data sets that are not fully standardized due to different accounting software and internal classifications that proved to be very useful for further data processing. Even though EAN/GTIN codes have been introduced and are in fact useful for data processing those codes themselves are not sufficient for classifying items (products) to statistical classifications mainly the product classification (CPA) and classification of individual consumption by purpose (COICOP). There is no bridge table between EAN/GTIN codes and statistical classifications. Above that, EAN/GTIN codes are not entirely unique and stable during the time.

Scanner data also allow estimation of superlative and multilateral indices. Bialek (2021) presented comparison of indices for dairy products. He came to the conclusion that data filtering is extremely important with substantial impact on the results.

3 DATA PROCESSING

Utilization of scanner data poses a challenge on National Statistical Institutes. Scanner data represent enormously rich data source for official statistics, namely price, retail trade and national accounts statistics. At the same time scanner data need to be processed into information. Even though scanner data received by the Czech Statistical Office are not real big data as individual records (transactions) are not submitted to the CZSO. In other words, data are pre-processed by reporting statistical units. Only monthly aggregates in breakdown by products defined by EAN/GTIN or internal codes are sent. It means that data on individual transactions or higher frequency data are not available to the CZSO that limit amount of transferred data. In addition, monthly frequency suffices as no more frequent statistics are produced. However, respondents are requested to send data twice a month. The first transmission includes partial (incomplete) data for the first three weeks of a given month, which are used for consumer price index. The second (complete) data transmission is utilized in other statistical domains.

It should be noted that processing of scanner data is challenging and advances statistical methods need to be deployed. Data structure is not fully standardized as each respondent provides the data from its database. While certain variables are the same (number of products sold, revenue with/without VAT) other may differ (internal classification, description). In addition, product variety gradually changes as about 5 000 new products are identified every month. Those items need to be classified into statistical classifications such as ECOICOP, CPA. Obviously, manual data processing would be very resource consuming. Internal classifications are very helpful for data processing nevertheless statistical classifications, especially ECOICOP, are very detailed.

4 MACHINE LEARNING

The crucial point of all the effort leading to automatic coding or classification is to find suitable base that will be used for supervised training. Further on, selection of appropriate statistical method is necessary. With respect to that, process started detailed studying of the products sold at selected retail chain. Experts used up all the information hidden in shopkeepers' information system ranging from typical words to the position of the goods in the shop. It was soon recognised that this demanding work can be used only for a short time in testing period. During the regular production, it had to be found something else.

After several different attempts for implementation of some type of partially or fully automatized procedures, it was decided to select logistic regression model inbuilt into Python big data and scientific environment Pandas and Scikit-learn library. Logistic regression provides easy and reliable solution with

very limited set of assumptions[5] and it is applicable on categorical variables (characters). It was found that universal GTIN codes (bar codes) are not easily translatable into statistical classification (COICOP) and the description of the product by both letters, numbers and words would have to be used. Even though that all data providers use their own specific internal system and coding system for sold products that was deeply used up during preparatory phase.

Finally, the process of classification of newly observed products that are classified by logistic regression and the results lead to continuously spreading knowledge database that is used for further classification. Obviously, the quality of classification can be tested only ex-post which we allow in this paper. For illustration purposes, we present the process of computation on reduced example counting 100 000 records and test classification quality on the randomly selected 5% sample. Further on, we also present actual information based on complete set of data.

Standard model of logistic regression for binomial variable (belongs to the group 1 and does not belong to group 0) can be described by Formula (1). Dependent variable (y) expresses 0 or 1 depending on the group belonging. It is rewritten as share of probability belonging to the group (P(y_i=1) or π) to opposite case, the share represents the chance – odd. Right side of the equation express exponent of linear regression with unknown vector of parameters ($\boldsymbol{\beta}$) and explanatory variables (x_1 to x_k):

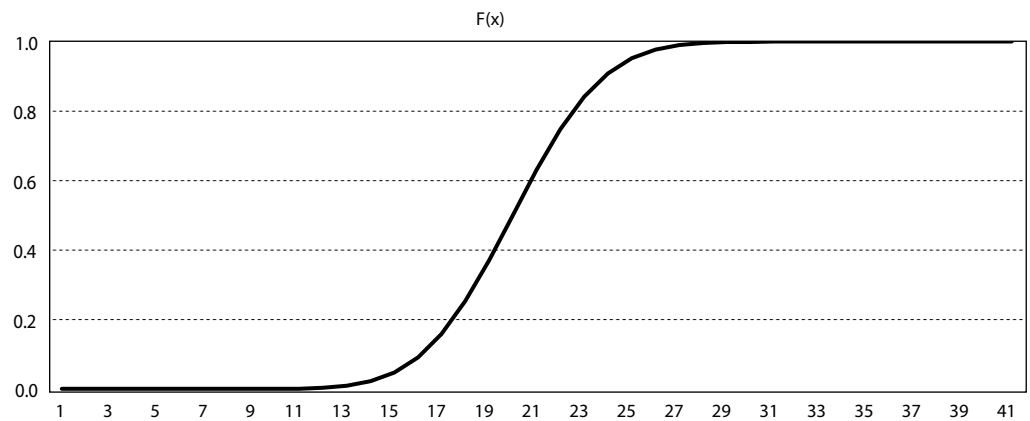
$$\text{odd } y_i = \frac{P(y_i = 1)}{1 - P(y_i = 1)} = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}. \tag{1}$$

For transformation purposes, the model is rewritten for logit (left side of equation) as:

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \tag{2}$$

Generally, the probability of classification in logistic regression has very suitable form and in optimal case leads to relatively strict separation between probability of belonging and non-belonging to group, e.g. sample logistic function see Figure 4.

Figure 4 Probability of classification in logistic regression



Source: Own elaboration

[5] Despite logistic regression does not require the same set of assumptions as linear regression, some assumptions are necessary, mainly dependent variable to be binary, independent variables should not be too highly correlated, independent variables are linearly related to the log odds and usually large sample size is recommended.

The description was provided on the basis of binomial dependent variable but in practice we use 157 COICOP groups, we need 156 equations. This multinomial comparison is based on the one vs all method, where for each product is estimated probability of belonging to tested group versus all other. It means that a product is being tested where it belongs to foodstuff, beverages, tobacco, …domestic appliances, …sport equipment, etc. The product is classified according to the highest probability in corresponding group.

The set of dependent variables correspond to the probability of belonging to 157 categories (given by specific COICOP group). All explanatory variables are derived from the product description used by the shopkeepers. This is represented by letters, numbers and words in the description and therefore all explanatory variable are categorical. For better explanation, see following Table 1 representing very short part of the explanatory variable, called dictionary (contains full or shortened description of the products). Due to computational issues, the Czech Statistical Office limits the number of explanatory variables to 150 000.

Table 1 Sample of the dictionary

Number	Text
1	0
…	…
24	00ml
…	…
203	100x200cm
…	…
17 778	Irsai
…	…
41 905	Zweigeltrebe
…	…

Source: Own elaboration

The explanatory variables cannot be used in the presented form, they have to be translated into a Boolean type of variables, representing the occurrence of concrete variant of the text. Processing of the text is very modern and Python environment allows many possibilities such as text recognising. The way that is used for logistic regression is based on transformation of explanatory variables into Boolean type, 0 or 1 when exact character/letter/word presents. A Python tool Vectorizer is used. In this way, variables are processed in a form of unit-zero matrix, see Table 2.

Table 2 Boolean representation of the matrix of explanatory variables

	Explanatory variables alphabetically ordered									
	1	**…**	**24**	**…**	**203**	**…**	**17 778**	**…**	**41 905**	**…**
Observation	0	…	00ml	…	100x200cm	…	Irsai	…	Zweigeltrebe	…
1	0	…	0	…	0	…	1	…	…	…
2	0	…	0	…	0	…	0	…	1	…
…	…	…	…	…	…	…	…	…	…	…
55 000	1	…	1	…	0	…	0	…	0	…
…	…	…	…	…	…	…	…	…	…	…
195 000	0	…	0	…	1	…	0	…	0	…
…	…	…	…	…	…	…	…	…	…	…

Source: Own elaboration

The estimates of parameters are completely done in Python. Since we have 150 equation with more than 150 000 parameters, it is not possible to simply present all coefficients and their p-values. Only synthetic quality information can be easily interpreted. Table 3 brings the sample of estimated parameters,

$\hat{\beta}_0, \hat{\beta}_1, .. \hat{\beta}_{150000}$.

Table 3 Illustration of fitted model

	b0	b1(500)														
		10	100	100g	100ml	12	140	15	150	150g	150ml	16	1l	20	200	200g
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	–0,0	0,3	–0,5	–1,2	–0,2	–0,0	–0,0	–0,1	–0,2	–0,2	2,7	–0,0	–0,4	–0,0	–0,4	–0,8
2	–0,0	–0,1	–2,0	–1,7	–0,0	–0,1	–0,0	–0,6	–0,8	–1,0	–0,2	–0,0	–1,9	–0,1		
3	–0,0	–0,2	–0,3	–0,4	–0,0	0,1	–0,0	–0,0	–0,8	0,1	–0,0	–0,0	–0,1	0		
4	–0,0	–0,2	0,9	0,5	–0,1	–0,8	0,1	–0,4	0,2	0,5	–0,0	–0,1	–1,1	0,8		
5	–0,0	–0,1	–0,1	–0,3	–0,1	–0,9	–0,4	–0,0	0,8	–1,2	–0,0	–0,0	–0,1	–0,0		
6	–0,0	–0,3	0,4	–0,2	–0,2	–0,1	–0,1	–0,2	1,3	–0,8	–0,3	–0,0	–0,7	–0,0		
7	–0,0	–0,8	0,8	0,8	–0,0	–0,1	–0,6	–0,5	1,1	0,0	–0,0	–0,0	–1,6	–1,9		
8	0,0	–1,0	0,8	0,5	–0,2	–0,4	1,1	0,5	–0,7	–1,1	–0,2	–0,1	–1,6	–2,1		
9	–0,0	1,2	1,7	2,4	–0,0	–0,0	–0,0	–0,0	–0,1	–0,1	–0,5	–0,0	–0,0	1,4	–0,7	2,3
10	–0,0	0,8	–0,4	–0,2	–0,0	–0,0	–0,0	–0,0	–0,1	–0,1	–0,7	–0,0	–0,0	–0,7	–0,8	–1,2
11	–0,0	–0,0	–0,1	–0,1	–0,0	–0,0	–0,0	–0,0	–0,1	–0,0	–0,1	–0,0	–0,0			
12	0,0	–0,6	–0,6	–0,2	–0,0	–0,0	–0,0	–0,0	–0,1	–0,1	–0,9	–0,0	–0,0	–		
13	–0,0	–0,1	–0,1	–0,1	–0,0	–0,0	–0,0	–0,0	–0,1	–0,0	–0,1	–0,0	–0,0			
14	–0,0	–0,6	–0,2	–0,1	–0,0	–0,0	–0,0	–0,0	–0,1	–0,0	–0,4	–0,0	–0,0			
15	0,0	0,8	0,1	1,5	–0,0	–0,1	–0,3	–0,7	0,3	–1,4	–0,2	–0,0	–0,1	–0,1	–0,0	–1,1
16	0,0	–1,2	–0,4	0,7	–0,0	1,2	–0,0	0,7	1,2	0,9	–1,8	–0,0	–0,2	1,5		
17	0,0	–0,1	–0,6	–0,2	–0,0	–0,0	–0,0	–0,0	–0,1	–0,3	–0,3	–0,0	–0,0	–		
18	–0,0	–0,0	–0,3	–0,5	–0,0	–0,4	–0,0	–0,0	–0,1	–0,3	–0,0	–0,0	–0,0	–		
19	0,0	–0,0	–0,2	–0,1	–0,0	–0,0	–0,0	–0,0	–0,1	–0,1	–0,1	–0,0	–0,0			
20	–0,0	–0,0	–0,1	–0,1	–0,0	–0,1	–0,0	–0,0	–0,1	–0,1	–0,1	–0,0	–0,0			
21	0,0	–0,0	–0,2	2,0	–0,0	–0,2	–0,0	–0,0	–0,1	1,4	–0,1	–0,0	–0,0	0		
22	0,0	–0,1	0,2	–0,6	–0,1	–0,4	–0,0	–0,1	0,6	1,5	–0,5	–0,0	–0,1	–0,0		

Source: Own elaboration

The interpretation of regression coefficients is not very easy since they are reflecting transformed variables. For the reflecting quality of the model, the CZSO estimates the rate of correctly classified products, sensitivity and specificity and ROC. These tests are done ex-post by splitting the data randomly into testing (95% of the sample) and prediction group (5% random sample). The results can be classified according to the belonging to a particular COICOP group (belongs – 1 positive, does not belong – 0 negative), see Table 4.

Table 4 Possible classification of outcomes

		Actual	
		False (0)	**True (1)**
Predicted	False (0)	True negative (TN)	False negative (FN)
	True (1)	False positive (FP)	True positive (TP)

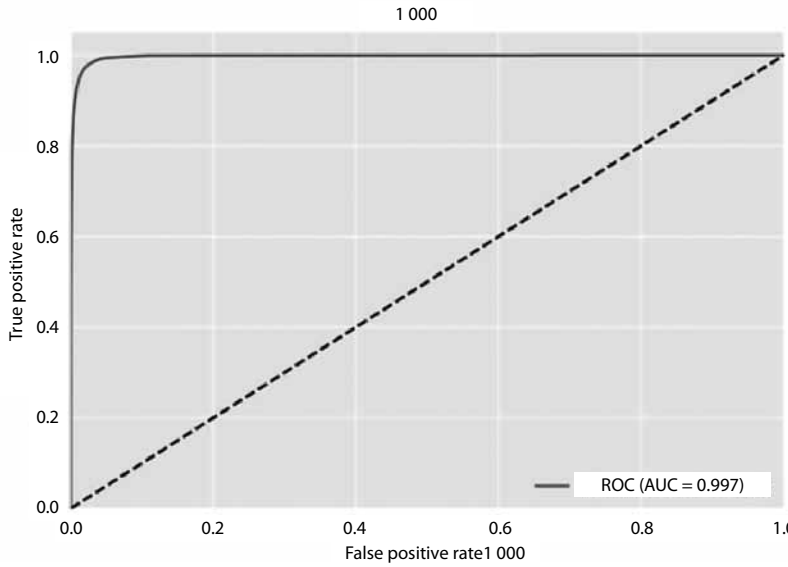
Source: Own elaboration

Standard measures in logistic regression such sensitivity and specificity are used as well as ROC curves. Since one to all approach is applied, the data for different COICOP groups had to be finally aggregated for overall assessment. Python environment and libraries Pandas and Scikit-learn are used. The overall success rate exceeds 95% when classified new product. ROC[6] is perfect for illustration of binomial classification,

[6] ROC is receiver operating characteristic curve and it is used for assessing the quality of classification with respect to sensitivity and specificity.

in our case it is obtained as an average of classification in detailed COICOP level. Since, the sub-groups are not well balanced, ROC provides just indicative picture of the quality of classification. For ROC of food and non-alcoholic beverages, see Figure 5. Despite the limitation of ROC for aggregated data, the quality of classification by logistic regression is enormous. The development of the true positive rate is very good and the Area Under the Curve (AUC) reached 0.997 when 1 is perfect fit.

Figure 5 Estimation of ROC for food and non-alcoholic beverages, COICOP 10



Source: Own elaboration based on the estimates of M. Král (CZSO)

The quality of classification of products (rate of truly classified) reaches very high numbers, exceeding 0.9, 90% of products are correctly classified. When studied for seven possible COICOP groups where these products belong, results are higher than expected. The probability of correct classification is presented in the Table 5. It is influenced by extreme observation and non-weighted average (i.e. all products without

Table 5 Probability of classification into COICOP group

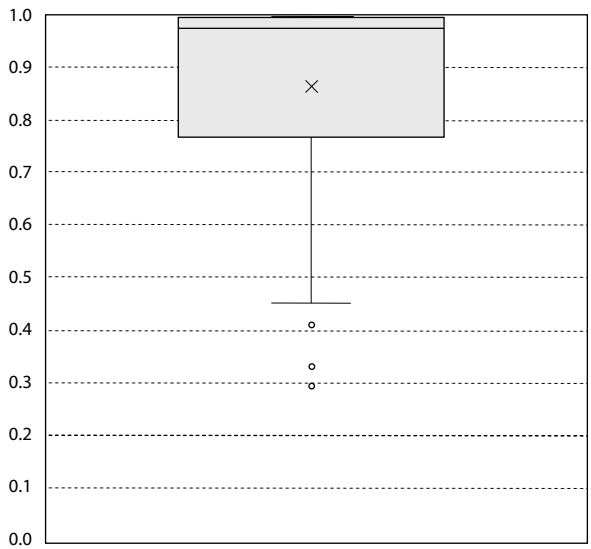
True classification		Predicted classification								
		01	02	03	04	05	07	09	12	Mean
1	Food and non-alcoholic beverages	0.68								0.68
2	Alcoholic beverages, tobacco		0.80							0.80
3	Clothing and footwear			0.95				0.41		0.92
5	Furnishings, household equipment and routine household maintenance				0.70	0.84				0.82
7	Transport						0.98			0.98
9	Recreation and culture							0.92		0.92
12	Miscellaneous goods and services								0.98	0.98

Source: Own computation on selected subset of data

respecting the amount of sales) does not fall below 0.68.[7] The figures in the tables come from aggregation of maximal probabilities belonging to the group. It must be taken into account that these aggregations do not play role in practice since products are classified on the lowest possible detail. The only problematic group seems to be clothing and footwear where significant misclassification was found. It means that for some products, the highest probabilities pointed at incorrect group, from clothing to recreation and culture. The distinction between sport tools and clothing is not easy in practical statistics and moreover this group is not currently fully covered by specialised shops. It means that these products come from retailers with high variety of products.

Similar view on the quality is presented on the Figure 6, this box-plot provides a distribution of probabilities of correct classification within our sample. It is clear that lower quartile reaching nearly 0.8 and median is close to 0.97. In the sample of 100 thousands of products, just three extremes were observed, from 0.3 to 0.5.

Figure 6 Distribution of probabilities in a sample



Source: Own elaboration

With respect to finds presented above, the overall classification of products is very successful and it allows very fast large data processing. Currently, the efficiency reaches upper limit of CZSO's possibilities. The knowledge database that being developed is continuously enlarging.

5 FURTHER DEVELOPMENT

The issue of machine-learning processing of scanner data will be supplemented by web-scrapped data in the near future. Some countries, such as Austria has been deployed scrapped data in price statistics (European Commission, 2020). Web-scrapping has lots of advantages but also limitations. Web scraping does not require cooperation with respondents that are just notified or discussed at the beginning of such process that is usually continuous and repeats with regular frequency and lasts long time.

[7] Due to computational issues, these results were estimated on basis of subset of all scanner data (100 000 observation) for illustration purposes.

The most important limitation is the sole existence of a spot price of product without quantity sold. Firstly, it is not possible to calculate true average price in a given period. Secondly, weights for aggregation into consumer basket are missing. However, practical solutions may be adopted to overcome missing quantities (European Commission, 2020). For example, it is recommended to Jevons index or more dynamic approaches (e.g. multilateral methods such as GEKS Jevons).

The Czech Statistical Office has experience with web-scraping in tourism statistics while price statistics in particular consumer price index rely on scanner data, central data acquisition and field collection. We believe that scanner data are more comprehensive data source than web-scraping especially for products whose price are volatile or their price elasticity of demand is high. Nevertheless, web-scraping may be applied in the future for companies that do not have electronic records at disposal.

CONCLUSION

The paper summarizes data sources and statistical methods used in price statistics notably consumer price index and brings readers overview of the current state of art of this new high quality statistics. Ongoing digitalization is a unique opportunity for modernization of (not only price) statistics using new data sources and data processing. The Czech Statistical Office started exploiting scanner data once respective legal act was adopted. Undoubtedly, scanner data have considerably improved quality of statistics. At the same time, new automatized procedures have to be applied to process huge data datasets within couple of days. Machine learning based on logistic regression fits very well.

Machine learning applied within the CZSO belongs to set of changes that should increase the quality and efficiency of statistical production. The possibilities of modern tools such as R and Python are huge but the most important for a successful application of ML procedures are high qualified statisticians and data analysts. The overall success rate exceeding 95% was not expected at the beginning since optimistic estimates were around 80%. This was a very big step into the new field and its possibilities. The success with scanner data motivated statisticians in other fields and now there can be found other projects at the CZSO being developed suited well for automatized and machine-learning classification. At the same time, machine learning itself it not a tool for quality adjustments or substitution of products. Highly qualified statisticians need to be employed by NSIs to develop and correctly apply modern statistical methods.

We are of the opinion that Laspeyres formula works well and the results are easily interpretable to the general public. Obviously, weights should be regularly updated to reflect changes in consumption habits. Nevertheless, scanner data allows to calculate also superlative price indices that will be subject of our future research.

ACKNOWLEDGEMENT

The support of the Technology Agency of the Czech Republic within the Project No SS04030013 (Center for Socio-Economic Research on Environmental Policy Impact Assessment) is gladly acknowledged.

Authors would like express special thanks to Michal Kral, expert from the Czech Statistical Office, who estimated ROC curves for the purposes of this paper.

References

BIALEK, J. (2020). Remarks on Price Index Methods for CPI Measurement Using Scanner Data [online]. *Statistika: Statistics and Economy Journal*, 100(1): 54–69. <https://www.czso.cz/documents/10180/125507867/32019720q1_54_bialek.pdf/f4ee19a0-75fd-41bf-b1fd-b192d177e125?version=1.2>.

BIALEK, J. (2021). Price Indices – a New R Package for Bilateral and Multilateral Price Index Calculations [online]. *Statistika: Statistics and Economy Journal*, 101(2): 122–141. <https://www.czso.cz/documents/10180/143550797/32019721q2_bialek.pdf/3cd5bf11-22f4-4ee5-b294-1d7d5909e4b4?version=1.2>.

DIEWERT, W. E. (1976). Exact and Superlative Index Numbers. *Journal of Econometrics*. 4(2): 115–145.
EUROPEAN COMMISSION. (2016). *Handbook on Price and Volume Measures in National Accounts*. Eurostat.
EUROPEAN COMMISSION. (2020). *Practical guidelines on web scraping for the HICP*. Eurostat.
FIXLER, D. (1993). The Consumer Price Index: Underlying Concepts and Caveats. *Monthly Labor Review*, 116.
LIPPE, P. (2012). Covariances and relationships between price indices: Notes on a theorem of Ladislaus von Bortkiewicz on linear index functions [online]. *MPRA Paper No. 38566*. <https://mpra.ub.uni-muenchen.de/38566>.
ROJÍČEK, M., SIXTA, J. (2022). *Machine Learning Process in the Production of Statistics*. [online]. <https://lsvdv-my.sharepoint.com/:w:/g/personal/ingabal_stat_gov_lt/EZZ09VOQmelOoaKsHPst3DUBpGVtY0d_emsGpmTmfBPQAw?e=L4WKjx>.
SAMUELSON, P., NORDHAUS, W. (2009) *Economics*. 19th Ed. NY: McGraw-Hill.
SCHULTZ, H. (1939). A Misunderstanding in Index-Number Theory: The True Konüs Condition on Cost-of-Living Index Numbers and Its Limitations. *Econometrica*, 7(1): 1–9.

Non-Profit Institutions in National Accounts

Ludmila Vebrová[1] | *Prague University of Economics and Business, Prague, Czech Republic*
Václav Rybáček[2] | *University of Economics and Management, Prague, Czech Republic*

Received 17.6.2023, **Accepted** (reviewed) 9.8.2023, **Published** 15.3.2024

Abstract

Since the outset of economic transformation in the 90th of the 20th century, the sphere of non-profit institutions in the Czech economy has been expanding. This evolution poses a challenge for macroeconomic statistics and its compilers to reflect adequately the operation of non-profit segment as well as to treat non-profit institutions properly in terms of their sector classification. The paper thus aims to discuss not only the treatment of non-profit institutions in national accounts but also the way non-profit institutions are identified in the regulatory environment of the Czech economy and classified following the rules laid down in the national accounts methodology. The paper also highlights major obstacles the compilers are facing when dealing with this specific economic sphere.

Keywords	***DOI***	***JEL code***
Non-profit institutions, national accounts, sector classification	*https://doi.org/10.54694/stat.2023.25*	*L30, E20*

INTRODUCTION

The transformation of the Czech economy into a market economy (since 1990 and 1993 respectively) has brought about a number of significant changes. The transformation concerned not only the change of ownership relations in the economy and their legal background, but also changes in theory and practice of economic policy and, not least, changes in the statistical system describing the development of economy. The system of statistics, as well as other areas of social life, later reacted to the accession of the Czech Republic (2004) to the community of states integrated in the European Union by continuously implementing regulations, directives and other recommendations of the EU institutions – the Council and the European Parliament – into the national legal system.

The introduction of the system of European national accounts as a system (ESA), whose aim is to show overarching relations between economic agents at the national economic level, internationally comparable flow and stock of economic indicators, macroeconomic aggregates, has become an integral part of these changes. The ESA represents a system that depicts past economic events in the national economy and,

[1] Faculty of Economics, Prague University of Economics and Business, W. Churchill Sq. 1938/4, 120 00 Prague 3, Czech Republic. Also Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic. E-mail: ludmila.vebrova@czso.cz.
[2] Department of Economics, University of Economics and Management, Nárožní 2600, 158 00 Prague 5, Czech Republic. E-mail: vaclav.rybacek@vsem.cz.

as an important tool of macroeconomic analysis, enables the quantifications feeding into predictions of the future economic development thereby serving as a basis for the decision-making process of the supreme government units. It has also become a basic tool for formulating proposals about the future economic and social goals at the level of national economy.

In principle, national economy can be described as a system in which number of different economic entities with different goals are acting in their role of consumers or producers of goods and services. These entities have different economic functions and behavior that, in legal terms, are expressed by their legal form. In general, they can have either market (business) or non-market character, suggesting the conditions or incentives defining their economic operation. While attention of economic analysts is mostly paid to the market sphere where most of the labor force are employed and which dominantly drives the business cycle, this paper focuses on the non-market sphere composed of non-profit institutions (organizations).[3]

The term non-profit institution (organization) is laid down in a number of national legal acts, e.g., on the tax or accounting system or, not least, in the acts defining legal forms of economic entities. In some cases, a non-profit institution is defined only by some essential features, the rules for the use of any financial revenue or even the prescribed accounting method (according to the Decree No. 504/2002 Coll., as amended for accounting units whose main activity is not business (if their account in the double-entry bookkeeping).

At the beginning of the transformation of the Czech economy, the activities of non-profit institutions were regulated by a number of legal documents determining their establishing and termination, registration, organizational structure, management, accounting system, etc. Currently it can be mentioned Civil Code Act No. 89/2012 Coll., as amended), Act No. 424/1991 Coll., on Association in Political Parties and Political Movements as amended, or Act No. 3/2002 Coll., on Churches and Religious Societies, as amended etc.

Over the transformation period, the non-profit institutions have been gaining importance which materialized in creating a specific executive government body. In 1992, the Council for Foundation was established that has been later transformed (in 1998) into the Government Council for Non-Governmental Non-Profit Organizations. In connection with this process, the term non-state (or non-government) non-profit institutions (NSO, NGO) has been gradually introduced. This term covers units established by non-government units only, i.e., units not managed by a ministry, central offices or municipalities. Their aims are generally beneficial goals (e.g., within healthcare, social services, sports, etc.) or to support members of the unit.[4]

Currently, the category of non-profit institutions is defined in the international standard "No.549/2013, Regulation of European Parliament and of the Council (EU), on the System of National and Regional Accounts in the European Union" (ESA), which has become an integral part of the Czech Republic legal system since 2004 (when the Czech Republic became part of EU). The Czech Statistical Office has been following the corresponding treatment of non-profit institutions since it started compiling the national accounts of the Czech Republic.

It should also be noted that a more detailed description of non-profit institutions has been set in another internal standard of national accounts – the System of National Accounts, UN (SNA), which is broadly compatible with the ESA. It is of note that while social benefits from the operation of non-profit institutions are generally substantial, they are inadequately reflected in the accounts due to non-monetary nature of many transactions in which non-profit institutions are involved.

---

[3] Both terms will be treated as synonyms in this article.

[4] However, the "definition" is not an official one in a sense that it is not legally established definition. It significantly differs from an internationally comparable category established in the international standard of the system of national accounts back in the 1950s.

Attempts to provide more realistic picture in terms of size and expediency of non-profit institutions in economies have been provided in a number of studies, not least exploring the value created by volunteering.[5] The aim of this paper is instead to deal with the basic features of a non-profit institution, with the notion of a non-profit institution in the system of national accounts, how non-profit institutions are identified, analyzed and classified against the background of the Czech legal system and data sources available to compilers. The aim is thus to reveal the complexity of proper identification of non-profit institutions as well as their sector classification which are not always straightforward.

1 BASIC FEATURES OF NON-PROFIT INSTITUTIONS

In the system of national accounts, non-profit institutions are recognized as an institutional unit if certain conditions common for all economic entities are fulfilled (see par. 2.12 ESA 2010). However, for non-profit institutions, there is a number of peculiarities that reflects a specific nature of their non-market orientation. The basic characteristic features of non-profit institutions, as laid down in the national accounts standards, are the following:

1. Non-profit institutions are economic entities (legal or natural persons) producing goods and providing services; they are established under the law (s) or special legislation under which they also carry out their economic activities and are (or not) recognized as an institutional unit in the system of national accounts. Similarly, to other economic entities operating in the national economy, non-profit institutions:
   (a) have a target (purpose) function;
   (b) dispose of (technical, financial, human) resources to fulfil their basic function, acquire non-financial or financial assets and create resources (liabilities);
   (c) are accounting entities that keep accounting records of all transactions, assets, receivables and payables, etc. i.e., they compile a complete set of accounts (or would be able to compile such accounts if necessary);
   (d) have autonomy in decision-making; are entitled to make economic decisions, enter into commitments for which they are legally responsible; in this case, such units are designated as institutional units in the national accounts system;
   (e) are not considered an institutional unit if they have limited (or no) autonomy in decision-making, i.e., economic decisions are made practically by their owners, who also have de facto legal responsibility;
   (f) are not always necessarily separated from their owner(s).
2. Specific reasons behind their foundation and the functions to which non-profit institutions serve:
   (a) philanthropic or charitable activities, i.e., activities related to efforts to provide humanitarian aid;
   (b) satisfying public or common interests:
      – in social life areas, such as health, education, culture, sport or social affairs, leisure or politics,
      – of certain group of people in business.

Due to a wide range of purposes which non-profit institutions may pursue, the founders of non-profit institutions are various economic entities from all economic sectors. Among others, government institutions routinely ensure, by way of establishing of non-profit institutions, the performance of some of their functions in the interest of the general public (for instance in the area of health services, culture or social security). Churches and religious societies also commonly establish and manage charitable organizations providing health, social or educational services to people in need. In case households, establishing of a non-profit institution is mainly the case of private associations such as those of gardeners,

[5] See e.g. Škarabelová (2005), Dostál (2020), Orlowski-Wicker (2016), Rybáček et al. (2017), Mook et al. (2007).

fishermen, sports clubs, etc. in the area of leisure, or with the aim of protecting property (fire brigades), or to preserve historical monuments or areas. Finally, business entities (industrial enterprises, agricultural cooperatives, banks, insurance companies, etc.) frequently establish non-profit institutions to serve in their interest which may range from lobbying to supporting research, education, or to centralize certain activities such as accounting or supplying. To this end, non-profit institutions serving business are jointly financed and controlled by the business which have established them.

3. Principal resources are primarily voluntary contributions (cash or in kind) from households, payments from public budgets, contributions or subscriptions from the group of businesses to whom they provide services.

4. Compared to market-oriented entities, non-profit institutions are not seeking profit. However, making a profit (financial gain) is not legally prohibited. Profit can be achieved, for example, through carrying out secondary activities (if agreed by the founder and stated in the instrument of incorporation – in the articles of association).

5. If a profit is achieved, a unit is not legally entitled to transfer the profit to the founder/owner. Profit shall be used for further development of the institution, for the improvement of the services provided or the goods produced. The inability to transfer profits to entities that set it up implies that no entity is in the position of "shareholder" or "copartner" as is the case of market entities. As it becomes clear, the notion "non-profit" does not imply an inability to make a profit but inability to pass a profit on another entity in form of primary income such as dividends. For these reasons, non-profit institutions may also be exempt from taxes of various kinds.

6. The management of a non-profit institution is entrusted to a group of persons (or a person). The officials are usually appointed by a majority of the founding members (or by a single founder). This group of officials and its activities are similar to the board of profit-oriented corporation. This means that it decides on the implementation of the activity of the unit, on the use of technical and economic resources, including the involvement of other paid employees or volunteers in the activities of a non-profit institution.

2 NON-PROFIT INSTITUTIONS AND INSTITUTIONAL SECTORS

The needs for quantification and recording of the economic entities activities, including non-profit institutions, their mutual technical-economic relations and positions, require grouping of economic entities (institutional units) into so-called institutional sectors (resident units), i.e., non-financial corporations (S.11), financial institutions (S.12), general government (S.13), households (S.14) and the sector of non-profit institutions serving households (S.15).

For non-profit institutions applies that they can appear in all institutional sectors.[6] The following table illustrates the sector classification of non-profit institutions in the Czech national accounts along with the number of units being thus classified.

Proper and correct classification of non-profit institutions into institutional sectors is not a minor issue as it affects the way their value added is calculated, thus the extent the value added creation in individual sectors as well as redistribution of income within the economy, the size of individual economic sectors, the deficit and debt of government institutions and therefore the findings of economic analyses. To decide on the proper sector classification, a sufficient information is needed on economic behavior, financing or the way of decision-making of the governing bodies, including the extent of control from the side of owner/founders.

[6] For S14 holds that only those non-profit institutions (serving households) not having independent legal status or those of minor importance can be classified in this institutional sector (par. 2.119 ESA2010).

Table 1 Sectoral structure of non-profit institutions, the Czech Republic, end-2022

Legal form	Institutional sector			
	S.11	S.12	S.13	S.15
Total	**628**	**9**	**17 979**	**149 968**
of which: central and local budgetary organizations, including state funds	x	x	7 348	x
semi-budgetary organizations	x	x	10 436	x
political organizations, churches, trade unions	x	x	x	11 623
public benefit societies	131	0	29	2 421
registered institutions	33	0	7	47
interest association of legal persons	228	8	0	880
registered associations	206	1	5	673
foundations and endowment fund	x	0	4	149

Source: czso.cz

Economy entities that are operating in the Czech Republic and which can be classified as non-profit institutions (based on the above features) are the following:

- Generally benefit company,
- Schools legal entity,
- Societies,
- Institutes,
- Political party, political movement,
- Registered church (legal entities and religious companies),
- Trade unions,
- Professional organization – professional chamber,
- Chamber,
- Foundation and endowment fund,
- Homeowner´s associations,
- Interest association of legal persons,
- Associations and unions that became society or institute in accordance with the civil code (No. 89/2012 Coll.); many of them retained the name "union" or "association" in the name),
- Other non–profit organizations – different funds and other units,
- Various funds and other institutions having, for example, in the articles of association the features of a non-profit organization or accounting according to the rules for entities for which the main activity is not business.

Organizational units of the state (ministries and central authorities) and state funds, territorial self-governing units, voluntary associations of municipalities, (i.e. "budgetary" organizations), all contributory organizations and health insurance companies managing general (compulsory) health insurance, public universities, public research institutions, etc.

2.1 Types of non-profit institutions

In the system of national accounts (that is implemented in the legislation of the Czech Republic) and other international statistical documents, non-profit institutions are generally divided into two groups, depending on their founders, namely private and public non-profit institutions.

A private non-profit institution is, by definition, established by private entities (resident or non-resident units). According to ESA conventions, some of these institutions are automatically classified in the sector of non-profit institutions serving households (S.15). These are political parties and movements, churches and religious societies, trade unions, professional chambers, consumer associations, social, cultural, recreational, sporting organizations and charitable organizations financed by voluntary contributions from households.

In providing products and services to private non-financial or financial corporations (entrepreneurs), such units are under the control of the respective founders and financed by their contributions. They are classified in S.11002 – non-financial national private enterprises, such as the Agricultural Union of the Czech Republic and its territorial organization or the Czech-Moravian Association of Agricultural Entrepreneurs, or in subsector S.12602 – private national financial institutions (e.g., The Czech Insurance Association or Czech Banking Association).

In summary, private non-profit institutions are classified into the following institutional sectors/ subsectors:

- non-financial corporations (S.11/S.11002), resp. of the sub-sectors 'private national non-financial corporations' (S.11002), or in the subsector 'foreign-controlled corporations' (S.110032), insofar as they provide services to foreign-controlled corporations (i.e. excluding public enterprises in sub-sector S.11001),
- financial institutions (S.12 / S.12602) – into the sub-sectors 'private national institutions' and 'foreign-controlled corporations', excluding the sub-sectors of public financial institutions,
- and non-profit institutions serving households (S.15).[7]

When it comes to the second group, i.e. a public non-profit institution, it is a unit usually established by a law or a government institution, e.g., a ministry, a territorial self-governing unit or a semi-budgetary organization (school etc.) or a public university, or according to a special legal regulation. It is a unit performing certain government functions, such as education, culture, health and social services, public administration and the municipal economy, or collets funds for the benefit of a certain government institution – school, museum, hospital, etc. or provides services only to government institutions. They are non-profit institutions, which are under a government institution control.

Control means the right of a government institution to exercise an influence over economic policy of the unit, to appoint managers (e.g., directors, chairmen, to act on the board), to formulate certain legal instruments or contractual arrangements (e.g., in the possibility of accepting loans and other payables, or in decision on investment activities). An important feature is the level of risk that the government unit bears on behalf of the corresponding unit (e.g., guarantee for the entity's liabilities) and, last but not least, the level of financing, i.e., the size of reimbursing operating costs of the non-profit institution (whether revenues from sales of products or services predominate or budget subsidies government institutions). These characteristics than determine[8] whether a public non-profit institutions is treated as a government non-profit institution (and therefore classified in S.13) or a non-profit unit can be classified elsewhere in national accounts (S.11, S.12).

In this respect, all semi-budgetary organizations, public universities, public research institutions or the Vine-grower Fund, Voluntary Associations of Municipalities established by the Act on Municipalities or as interest associations of legal persons, societies established by the Civil Code are all classified in S.13, as they are treated public non-profit institutions operating on the non-market basis.

---

[7] E.g. § 2.51, 2.53, 2.65 d) etc. *European System of Accounts* (ESA 2010), Luxembourg: Publication Office of the European Union, ISBN 978-92-79-31242-7.

[8] For more details, please see part 2.2.

In this context, it is worth mentioning the term "non-state non-profit institutions" encompasses institutions from both groups, private and public.[9] Under the term "state non-profit institutions" budgetary and semi-budgetary institutions, grouped in S.13 as government non-profit institutions, are commonly understood. As described above, also non-profit institutions of other legal forms might be controlled by government institutions. This means that the group of non-state non-profit institutions represent only a sub-group of public non-profit institutions or, from another perspective, the terms "private non-profit institutions" and "non-state non-profit institutions" do not overlap.

In this sense, the term "non-state" institutions includes private and public non-profit institutions which might be classified in S.11, S.12, S.13 and S.15, while state non-profit institutions are, by definition, classified in S.13. The logic behind the decision on the eventual classification of an individual units is discussed in the following paragraphs.

2.2 Assessment of economic behavior of non-profit institutions

Non-profit institutions are not primarily meant to seek and generate profit. However, it might be the case that the operation of a non-profit institutions turns out to be profitable. From the point of view of national accounts´ methodology, it is to be assessed first who controls a unit and, second, whether a unit, including non-profit one, is a market or non-market economic agent. Concretely, whether the actual operation of a unit is akin to that of private market entities.

Concerning the former, i.e. who is the controlling agent, non-profit institutions can be founded by:

(a) market producers (corporations) to provide services to their founders; if so, non-profit institutions are also considered as market producers and are classified in the sector of the controlling party; an example is an Association of agricultural companies (classified in S.11), or Association of banks (classified in S.12) etc.;

(b) non-market producers (i.e. a government institution or some of non-profit institution serving households other than households to provide services to their founders, which are also considered as non-market producers and are classified in the same sector/subsector);

(c) households and to produce products or provide services.

It is of note that the rules just mentioned apply only in case a non-profit institution is recognized as institutional unit, i.e. it fulfills the conditions prescribed in par. 2.12 ESA2010. If not, a unit is automatically consolidated with its founder. In the methodology, this is commonly referred to as "qualitative criteria" which are to be applied in the assessment of sector classification of any unit. Practically all of the existing legal forms of non-profit institutions can be, or not, considered an institutional unit (according to the ESA), depending on the decision-making autonomy. Decision-making autonomy is normally limited in case of the organizational units of the state, whose income and expenses are directly determined through the state budget (which is approved by the government and Parliament), as well as contribution organizations that manage the assets of the founder and can only carry out a number of activities with the consent of the founder (e.g., receive loans).

If the outcome of "qualitative analysis" is such that a unit is indeed recognized as an independent institutional unit, it still remains to be determined whether it operates on the market or non-market basis in order to decide on a proper sector classification. This is where so-called quantitative criterion comes into play. Although established as not profit-seeking agent, the actual operation of a non-profit institution may show signs of market behavior where its output is provided at economically significant prices (par. 3.19 ESA2010).

[9] Which also reflects the fact that the term "government" in the national accounts methodology is much a broader term than the term "state", see par. 20.08 ESA2010 and the followings.

Determination of the market and non-market nature of a public unit is based on quantifying the relationship between sale of the production (revenue) and production costs for their achievement (par. 20.29 ESA2010). Specifically, a market producer sells its products and services at so-called economically significant prices, i.e. prices that should from more than 50% covered production costs. Expressed by the ESA2010 codes, it is therefore the relationship of market output P.11 to the sum of production costs, i.e. compensation of employees (D.1) + intermediate consumption (P.21) + consumption of fixed capital (P.51c) + net interest (calculated as the difference between interest expense and income; D.41p–D.41r) and reduced by production for own use ((–) P.12).

The test relating to non-profit institutions is calculated with regard to market and non-market production (P.11 + P.131), because non-profit institutions report (both in accounting and in statistics) main and secondary (economic) activities separately; secondary activities are usually treated as market.

In addition, there are other rules specified in the Manual on general government deficit and debt which specifies conditions under which a operation of an unit can be likened to that of market agent. For example, if a public non-profit institution sells products or provides services to government units only, e.g. Union of Towns and Municipalities of the Czech Republic, then this unit is treated as effectively controlled by government and therefore classified in the same sector as the controlling units.

To generalize this concept, if no actual competition between private or public producers exists, a public producer is classified in S.13 (in compliance with the qualitative criterion of government units' classification). If the public non-profit institutions sells products or services to other costumers (not only to the government) and although its costs are more than 50% covered by revenue for products or services (sales), such unit is treated as a market public producer (a non-government institution), classified in S.11 or S.12.

Quantitative criterion (distinction between market and non-market producer), together with the qualitative criterion, are therefore among the basic criteria for classifying non-profit institutions into institutional sectors. However, according to ESA2010, the basic criterion is considered to be a qualitative criterion, which means that non-profit institutions providing their products and services at economically significant prices and whose operating costs are more than 50% covered by sales can still be classified in the non-market producers sector (S.13 or S.15).

As already mentioned, non-profit institutions (public or private) are classified in different institutional sectors (international classification ESA2010). Considering their main economic activity nature, they are also represented in various industries of the national economy (CZ-NACE classification) and the expenditures of non-profit institutions classified in S.15 are monitored according to the COPNI classification (classification of services of non-profit institutions serving households).

2.3 Legal form of non-profit institutions

As became clear in previous paragraphs, in order to classify non-profit institutions properly a wide range of relevant information is needed in order to determine whether a non-profit institution:

- is a private or a public unit,
- and whether its intended function is of a market or non-market nature.

In the former, statisticians need to assess whether public or private bodies control a non-profit institutions. In the latter, it is to be assessed whether a non-profit institutions provides its output at economically significant prices or at prices which are seen as economically significant. Obviously, the determination of the sector classification is not as trivial task as it might seem at the first sight. In the following paragraphs, major directions of consideration on the sector classification and data sources are outlined.

First of all, the key element is the legal form of a non-profit institutions as these are laid down in the Czech legal system. However, the legal form is not decisive for classification into the institutional

sector as it does not clearly indicates whether a unit operates on the market or non-market basis, in the national accounts´ sense. In addition, the legal form set out in the Register of Persons (based on the designation of editing centers) is implemented (unchanged) in the Register of Economic Entities administered by the CZSO. The legal form is only the basis for the sectoral classification of the newly established unit, as there is no further information needed on its predominant activity. New established unit might be reclassified subsequently, based on a more detailed information on its actual operation, which becomes available after a certain period of time.

In addition, some legal forms determined by the editing center (for the purposes of the basic Register of Persons and taken over by the Czech Statistical Office into the Business Register) may not exactly correspond to the purpose function of the relevant unit. An example was the Hedge Fund, which manages contributions from Health Insurance Companies (not from the insured), which received the legal personality (form) of a "Health Insurance Company", although its purpose function was different. Another example is determination of the legal form of foundations and endowment funds, where number of foundations received (from the editing center) the legally personality of the endowment fund and vice versa.

For assessment of the proper sector classification of a non-profit institution, the information from the Public Register has important position as it provides an information on the field of activity, business, information from the financial statements and relevant documents such as deed of establishment, Statute of the unit, etc. However, in many cases such as interest associations of legal persons, society, or institution etc., this information is not sufficiently specific to allow for proper classification in terms of sector or branches.

Yet, the characteristics of the subject of the activity are very often of a general nature (e.g. "support of education, culture, sports, science and research", or "production, trade and services not classified elsewhere", etc.) that it is inconclusive in choosing a basic activity. It is also often the case that the main activity is, in a separate list of activities, listed as the last one (instead as the major function). In the case of interest associations of legal entities, the subject of activity is usually not reported at all. It is of note that the name of the unit is not a sufficient guide for their classification.

Moreover, two legal forms (personalities) are reported in the Public Register (in a number cases), although the Civil Code states that the name of the entity "must distinguish a legal entity from another entity and contain an indication of its legal form" (§ 132 of the Civil Code No. 89/2012 Coll.). In the event that the unit has changed an «interest association of legal person» to an «association» and approved new Statutes, the name of the institution has the correctly stated abbreviation «z.s.", i.e. a registered association. At the same time, this unit has however stated, in the list of basic ("valid") features, in addition to the data of creation and registration, identification number (verbal) legal form as "interest association of legal entities". Due to that, the Register of Associations and the Register of Persons (ROS) are not accessible to verify the correct legal form and find out the reason for this approach, including the legal form stated in the Register of Economic Entities managed by the CZSO based on the information from the Register of Persons.

As it became clear, more information on the actual operation is needed. Relevant information are retrieved from publicly available information such as deed of proof or establishment, the articles of association or accounting statements of a unit. Taking all these elements into account, only then the decision on the sector classification can be made and may also change in the course of time if a situation has changed.

CONCLUSION

The aim of the paper was to discuss the specific case of non-profit institutions, i.e. units taking several legal forms and being classified in several sectors in national accounts. The paper outlined the basic ideas behind their identification, based broadly on the legal form, and their sector classification

in national accounts, following the rules laid down in the relevant methodology. For proper classification, it is necessary to determine whether a unit in question acts as an independent institutional unit, i.e. independently of its founder/owner, and whether the actual operation meets the criteria of market or non-market unit. Although established as non-profit, such a unit can conceivably fulfill the criteria for a market unit as specified in the national accounts methodology. The paper also pointed to some drawback in the existing information on the non-profit institutions in public databases. From the perspective of the national accounts methodology, it is of note that the methodological approach toward non-profit institutions has been rapidly developing, especially in case of public non-profit institutions, and further development can be expected with the new generation of national accounts manual currently discussed and expectedly applied from 2029 onwards.

References

CZECH STATISTICAL OFFICE. (2022). *Satelitní účet neziskových institucí* [online]. Prague: CZSO. [cit. 20.5.2023]. <https://apl.czso.cz/pll/rocenka/rocenka.indexnu_sat>.

DOSTÁL, J. (2020). Revealed value of volunteering: A volunteer centre network [online]. *Annals of Public and Cooperative Economics*, 91(2): 319–345. <http://dx.doi.org/10.1111/apce.12271>.

EUROPEAN COMMISSION. (2023). *Manual on Government Deficit and Debt – Implementation of ESA 2010*. Luxembourg: Publication Office of the European Union. ISBN 978-92-76-98618-8

EUROPEAN COMMISSION. (2013). *European System of Accounts (ESA2010)*. Luxembourg: Publication Office of the European Union. ISBN 978-92-79-31242-7

FRIČ, P. (2001). *Neziskový sektor v ČR: výsledky mezinárodního srovnávacího projektu Johns Hopkins University*. Prague: Eurolex Bohemia. ISBN: ISBN: 80-86432-04-1

KOČÍ, P. (2005). *Nevýdělečné organizace – daně a právo.* Prague: SPI.

MOOK, L., HANDY, F., GINIENIEWICZ, J., QUARTER, J. (2007). The value of volunteering for a nonprofit membership association: The case of ARNOVA. *Nonprofit and Voluntary Sector Quarterly*, 36: 504–520.

ORLOWSKI, J., WICKER, P. (2016). The monetary value of voluntary coaching: An output-based approach. *International Journal of Sport Finance*, 11: 310–326.

SKALÁKOVA, Š., FOŘTOVÁ, J., RYBÁČEK, V. (2017). Valuation of Volunteer Work in Satellite Account of Non-Profit Institutions [online]. *Statistika: Statistics and Economy Journal*, 91(4): 4–17. <https://www.czso.cz/documents/10180/45606529/32019717q4016.pdf/72615c60-24d2-4cc2-a21a-11cbb57e930e?version=1.0>.

ŠKARABELOVÁ, S. (2005). *Definice neziskového sektoru: sborník příspěvků z internetové diskuse CVNS.* Brno: Centrum pro výzkum neziskového sektoru.

UNITED NATIONS. (2008). *System of National Accounts 2008*. European Commission, International Monetary Fund, Organization for Economic Co-operation and Development, United Nations, World Bank. ISBN 978-92-1-161522-7

UNITED NATIONS. (2003). *Handbook on Non-profit Institutions in the System of National Accounts*. New York: United Nations. ISBN 92-1-161461-9

Recent Events

Conferences

The ***26th Applications of Mathematics and Statistics in Economics Conference (AMSE 2024)*** will take place ***from 28th August to 1st September 2024*** in ***Wroclaw, Poland***. The conference aims to acquaint its participants with the latest mathematical and statistical methods that can be used in solving theoretical and practical problems and challenges of economics. More at: <*https://www.amse-conference.eu*>.

The ***32nd Interdisciplinary Information Management Talks (IDIMT 2024)*** will be held ***during 4–6 September 2024*** in ***Hradec Králové, Czechia***. IDIMT-conferences became interdisciplinary international forum for the exchange of concepts and visions in the area of software intensive systems, management and engineering of information and knowledge, social media, business engineering, and related topics. More at: <*https://idimt.org*>.

The ***18th International Days of Statistics and Economics (MSED 2024)*** will take place ***during 5–6 September 2024*** in ***Prague, Czechia***. The aim of the conference is to present and discuss current problems of statistics, demography, economics and management and their mutual interconnection. More at: <*http://msed.vse.cz*>.

The ***23rd*** biannual conference and joint summer school in statistics ***ROBUST 2024*** will take place ***from 8th to 13th September 2024*** in ***Bardějov, Slovakia***. This conference is organized by the Union of Czech Mathematicians and Physicists, Czech Statistical Society and Slovak Statistical and Demographic Society. Its programme is focused on theoretical and applied statistics, probability, optimization and data analytics. Special section will be focused on impact of AI and open sources like ChatGPT, Wikipedia on teaching statistics and work of statisticians. More at: <*www.karlin.mff.cuni.cz/~antoch*>.

The ***42nd International Conference on Mathematical Methods in Economics (MME 2024)*** will be held ***from 11th to 13th September 2024*** in ***Ústí nad Labem, Czechia***. The conference is a traditional meeting of professionals from universities and businesses interested in the theory and applications of operations research and econometrics. More at: <https://mme2024.ujep.cz>.

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
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

104th 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