PaKSoM 2024
DOI: https://doi.org/10.5281/zenodo.14693416

Original Scientific Article

Modeling Education and Internet Usage: PCA and Linear Regression Approaches

Ivana Petkovski[1], Petar Vranić[2]

[1,2]Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, Serbia

[1]ivana.v.93@gmail.com, [2]petarvvv@mi.sanu.ac.rs

Abstract—**The frequency of Internet use is significantly influenced by the formal and non-formal education of individuals. Acquiring education enables the acquisition of basic skills in the use of digital technology and influences the reduction of the digital gap created in this way. In this research, the influence of a number of education parameters on the rate of Internet adoption in 27 countries of the European Union was modeled. The principal component analysis (PCA) technique was used to identify the parameters that contribute the most to the development of education through the formation of two factors. The statistically significant score of factor one was used to form a linear regression (LR) model. The results show that the percentage of explanation of the variability of internet users is explained with 54.5% through the value of factor one. The correlation coefficient between real and projected values (0.738) shows a strong positive correlation between these values.**

Keywords - insert education, internet, PCA, linear regression, digitalization

I. INTRODUCTION

The Internet's rapid rise and transition to digital world in recent years has had a significant influence on today's society. It has resulted in changes in all fields and has had a wide-ranging impact on our daily life. This implies that teaching individuals essential skills for using digital technologies has become critical. Schools play an important role in enabling more individuals to utilize the Internet and develop digital skills. Digital literacy involves being able to find, handle, grasp, and assess information utilizing digital technology. This talent enables people to convert information into knowledge. It is about educating people ICT concepts and how to utilize digital technologies effectively [1,2]. This not only helps people get decent employment and create their own enterprises, but it also helps them integrate into society.

The OECD report "Skills Outlook 2019: Thriving in a Digital World" stresses how digital skills matter in many industries as AI, automation, and data analytics change their structure. The report also shows how digital changes shape new skill needs and affect education systems [3]. In the same vein, "The Future of Jobs Report 2020" by the World Economic Forum estimates that by 2025, machines will replace about 85 million jobs. At the same time digital shifts may create 97 million new roles [4]. Digital shifts in business, drive economic growth and social progress. They cut transaction costs, boost worker knowledge and skills, and affect inflation, wages, job markets, and output. All these factors add to the ongoing structural and tech changes in the world economy. Results presented in McKinsey & Company (2018), estimates that improvement in digital skills is likely to contribute up to additional $2 trillion to global GDP by 2030, whereas, nations with higher level of digital literacy are likely to have more prominent growth [5].

Regardless of the intense propagation of Internet access and digitalization, there is a prominent digital divide between the regions and nations. The concept of the “digital divide”, often known as the “digital gap”, is closely linked to the possibility of accessing and utilizing

ISBN: 978-86-82602-04-0

information and communication technology in one country or between countries [6]. The emergence of a digital divide within the nation hinders the ability of individuals to actively engage in social processes and access information that is important in an information age [7]. Education is important in bridging this gap by ensuring fair access to digital tools and information, allowing all people to benefit from the digital transformation. International Telecommunication Union conducted a study that shows that various digital literacy programs conducted in rural regions reflects on increase in income among participants for 20%, emphasizing the potential of such educational initiative [8].

The dynamics of digitalization with constantly emerging of new tools and online platforms calls for continuous education and skill update, not only in form of formal schooling, but also though life long learning programs for employability in the digital age. Studies shows that individuals who are undergoing continuous learning are less likely to loss their jobs i.e., have more chance for upward mobility [9]. Besides formal education, community-based non-formal learning has significant potential for developing digital literacy, especially among social groups that are less presented in the digital economy like low-income families, rural population or elderly. Study "The Role of Community-Based Programs in Closing the Digital Divide" shows that community-based digital literacy program may uplift not only digital skills of participants, but also their self-confidence in using technology that increase use of Internet in further education [10].

The quality education is undoubtable key factor for the expansion of the Internet and digitalization. However, among plethora of indicators that describes education system it is important to understand which one have dominant influence when it comes to use of Internet. Thus, this paper main goal is to model this relation in order to offer better understanding of which segments of education have significant influence on use of Internet.

II. Literature Review

A number of studies in broader sense attempted to model the relationship between the education, students and the use of Internet and digitalization, as well as their implication for employability. For instance, through an empirical study, [11] modeled relationships between traditional literacy (reading, writing, and understanding text), medium related internet skills (operational and formal skills) and content related Internet usage (information and strategic skills) and Internet usage types (information, career or entertainment directed use). The study applied structural equation modeling to test developed conceptual model, and Principal component analysis (PCA) with varimax rotation to determine two underlying usage clusters. Results point out that traditional literacy is a pre-condition for the employment of Internet skills, i.e. traditional literacy has a direct effect on formal and information Internet skills and an indirect effect on strategic Internet skills. Another study, presented in [12] attempted to determine the effect of the RMS teaching approach on students' digital and mathematical literacy. The study used a quasi-experimental approach, with two experimental courses and one control class. Data analysis analyze statistical distribution, including percentage, mean, standard deviation, correlation, and Analysis of Variance (ANOVA). In order to explore the relationships between variables authors applied structural equation modeling (SEM). The results shows that understanding mathematical principles can help students better appreciate and utilize digital technologies effectively, i.e. math education can enhance digital literacy. In developing a value-based online learning model to predict learners' reactions to internet entrepreneurship education [13] using partial least squares structural equation modeling (PLS-SEM). The findings revealed that instructional assistance, instructor excitement, and instructor preparation are important drivers of the perceived value of utilizing online learning platforms to develop internet entrepreneurship knowledge and skills, which favorably increases reuse intention. A study presented in [14] modeled the antecedents of artificial intelligence usage and effects on 21st century employability skills among postgraduate students in Ghana using partial least squares structural equation modelling (PLS-SEM) techniques for the quantitative data analyses and thematic pattern matching for the analysis of qualitative data. For checking the item loading of five selected variable, authors employed Confirmatory Factor Analysis (CFA). The PLS-SEM paths modelling showed that both AI usage and AI challenges were important predictors of postgraduate students' acquisition of 21st century employment skills. Also, the use of advanced online tools as for instance chabot's increases digital literacy

even more. In [15] authors tried to understand the higher education students' usage of ChatGPT. Applying a modified technology acceptance model, the findings of this study, which used a stratified random sampling approach to recruit 1389 higher education students from 11 universities in Vietnam, revealed that effort expectancy not only directly affected students' actual use of ChatGPT, but also serially indirectly increased their actual use of ChatGPT via performance expectancy and intentions to use ChatGPT. To evaluate the reliability and validity of the constructs being researched, and to test the proposed hypotheses, a three-step analysis is used. The first step considered Confirmatory factor analysis (CFA) and Cronbach's alpha to assess the reliability and validity of the constructs. Secondly, multiple linear regression analysis was utilized to investigate the predicted relationships in the developed model. And thirdly, for mediation coefficients testing, authors applied models 4 and 6 of the PROCESS macro approach with 5000 bias-corrected bootstrapping samples.

Having an understanding of the causal relationship between education and the use of Internet can help policymakers in developing strategies towards more digitally inclusive society.

III. Data and Methodology

The relationship between education and internet use was modeled on the case study of EU 27 countries such as Belgium, Bulgaria, Czech Republic, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, and Sweden. The data for this investigation were gathered from open-access databases, including the World Data Bank and the European Commission for the reference year 2022 [16,17]. The initial database considered data on pupils and student's enrollment distributed at all levels of education, data on pupils out of school, number of classroom teachers, number of graduates at different level of education, non - formal education and training of working age population, early leavers from education and training and share of GDP invested in education. In total, seventeen variables of education and training indicators gathered by the European Commission for the EU27 were evaluated in the first iteration of the analysis for dimensionality reduction, choosing twelve indicators to form a regression model. The Table I reports the overall independent parameters (x1-x12) considered for the model establishment, along with the dependent parameter (Int_use). The diversity in measures was solved by adopting percentage (%) as a referent unit.

The methodological part of the paper considered Principal Component Analysis (PCA) as a dimension reduction technique for selecting the most optimal parameters as indicators of education progression. The precondition for implementing PCA is to perform the Kaiser-Meyer-Olkin (KMO) test and Bartlett's test of sphericity to ensure that the data is appropriate for using this type of technique [18]. The PCA's results will identify factors that provide the most valuable information for explaining data variability [19]. The PCA groups

TABLE I. Initial parameters.

Indicator name	**Label**	**Unit**	**Source**
Individuals - internet use	Internet use	Percentage of population (%)	World data bank [16]
Gross domestic Expenditure R&D in higher education	x_1		European Commission [17]
Pupils enrolled in primary education	x_2		
Pupils enrolled in lower secondary education	x_3		
Pupils enrolled in upper secondary education	x_4		
Students enrolled in bachelor education	x_5		
Students enrolled in master education	x_6		
Students enrolled in doctoral education	x_7		
Formal education and training of population ages 18 to 64	x_8		
Non -Formal education and training of population ages 18 to 65	x_9		
Graduate pupils tertiary education	x_{10}		
Graduates students bachelor education	x_{11}		
Graduates students master education	x_{12}		

the indicators that share the highest correlation coefficients and establishes a factor or component [18]. The number of components is selected based on an eigenvalue that is higher than threshold 1 or on the basis of a scree plot examination [19]. For easier interpretation of the results, the several rotation methods can be used [20]. The linear regression (LR) analysis will use the outcome values of the factors as input parameters. Linear regression analysis is typically used to investigate the relationship between independent and dependent variables through creating regression equation [21]. The results of the LR summarize the value of the coefficient of determination (R2), ANOVA analysis, and coefficient statistics [21].

IV. Research Results

The Table II presents the values obtained from testing the suitability of the data set for PCA analysis. The results show that after several iterations, the obtained value of the KMO test and the *p*-value of Bartlett's test of sphericity reach acceptable values [18].

TABLE II. Testing data adequacy for the PCA.

Test name	Indicator	Reference value	Calculated value
Kaiser-Meyer-Olkin test	KMO index	> 0.6	0.710
Bartlett's test of sphericity	*p*-value	< 0.05	0.000

TABLE III. Rotated component matrix.

Extracted parameter	Component	
	Factor 1	Factor 2
Formal education and training of population ages 18 to 64	.898	
Pupils enrolled in upper secondary education	.783	
Gross domestic Expenditure R&D in higher education	.759	
Pupils enrolled in primary education	.653	
Students enrolled in bachelor education		.828
Students enrolled in doctoral education		.800

TABLE IV. Pearson's correlation matrix.

		Individuals - internet use	REGR factor score 1 for analysis 1	REGR factor score 2 for analysis 1
Pearson Correlation	Individuals - internet use	1.000	.738	-.003
	REGR factor score 1 for analysis 1	.738	1.000	.000
	REGR factor score 2 for analysis 1	-.003	.000	1.000
Sig. (1-tailed)	Individuals - internet use	.	.000	.494
	REGR factor score 1 for analysis 1	.000	.	.500
	REGR factor score 2 for analysis 1	.494	.500	.
N		27	27	27

TABLE V. Model summary.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.738	.545	.526	3.42398	1.583

The next step in modeling the relationship between education and internet use is to employ PCA to extract the most valuable parameters [22]. Table reports reduced dimensions of education progression expressed as components of two factors. The first factor accounts for 46.88% of the data variability, while the second factor accounts for 19.90% of the variability. Factor I contributes to the internet adoption level through the share of the formally educated population, lower-level educated individuals, and the government's investments in higher education. While Factor II contributes to the internet adoption rate through highly educated populations, such as bachelor and doctoral-level educated populations. The first factor refers to basic educational prerequisites for the use of digital technology, while the second factor includes more sophisticated indicators of Internet use. The Table III reports factor loading results higher than 0.65 for each extracted parameter. The results were calculated in three iterations using the Varimax rotation method.

The obtained scores of the two factors were utilized as independent parameters for the LR model. In order to evaluate their correlation with the dependent parameter Int_use, Person's correlation coefficients were calculated (Table IV). The results show a high positive correlation value between the regression factor score for factor one and internet users ($r = 0.738$), with a statistical significance of $p = 0.000$. The results for factor two failed to prove a statistically significant relationship between the observed parameters ($p > 0.05$) so it was excluded from the LR analysis.

The LR model was built based on the scores obtained by factor one. The computational outcome reveals a high positive correlation between real and predicted values ($R = 0.738$) of the internet users. The coefficient of determination value equals 54.5%, with a Durbin-Watson value in the acceptable range, indicating there is no autocorrelation between residuals (Table V). The ANOVA test results (analysis of variance) confirm the LR model's quality and statistical significance, with f-statistics of 29.904 and a p-value of 0.000.

The Fig. 1 illustrates the difference between the actual and projected value of the dependent parameter *Int_use* using the PCA-LR model in the EU 27.

V. DISCUSSION

The modeling results highlight two key characteristics of the observed set of countries. The first characteristic emphasizes that the percentage of the population with primary education, non-formal education, and training, along with GDP investment in higher education, largely determines the rate of internet use in the EU 27. The later has been also highlighted by [23], arguing that investment in higher education significantly boosts innovation capacity, particularly in regions with strong research universities, contributing to overall technological development in US. This argument is supported by [24]. Study finds that regions with more universities experience higher levels of innovation and technological advancement, making investment in higher education critical for fostering technological progress. Focusing on extensive insights on European countries [25] shows on how higher education levels correlate

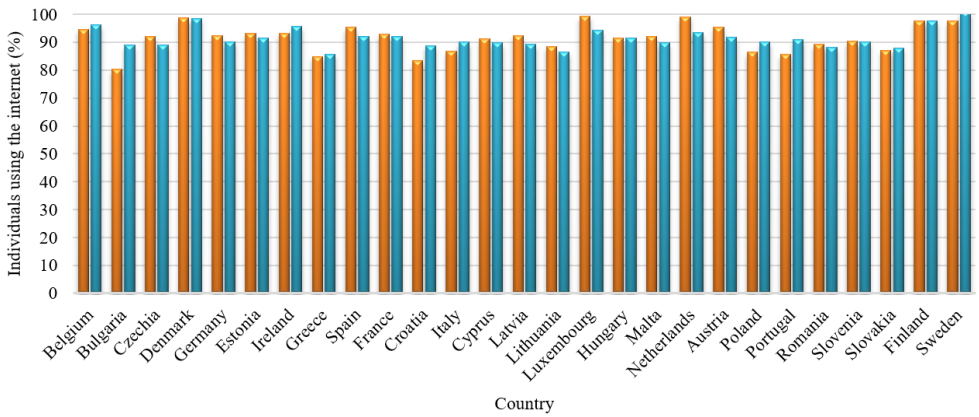

Figure 1. Real and predicted value of internet users in EU27 using PCA-LR model.

with technological advancements, particularly in Western and Northern Europe, providing an empirical basis for the link between education investments. This is also portrayed in [26], through detailed analysis of how universities in several European countries, including Italy, Germany, and the UK, contribute to technological development.

In this study, the mentioned parameters are marked as Factor 1. The next characteristic of the set is that the parameters pertaining to the highly educated segment of the population do not statistically significantly influence the number of internet users, as indicated by Factor 2. The share of the highly educated individuals is minor compared to the individuals with lower education, so the result of this relationship is also reflected in the rate of Internet use.

In order to increase the rate of Internet users, it is necessary to plan the activities of promotion and support of the use of digital technology towards the low-educated population. Furthermore, this study suggests that we should provide support through both formal educational institutions and non-formal educational channels. ICT specialists organize trainings to acquire digital knowledge through non-formal education [27]. However, the young population can also offer non-formal education to the elder population to help them gain basic digital skills [28]. Simultaneously, we should concentrate on the population segment experiencing the most severe effects of digital inequality, specifically the marginalized groups. Individuals with a lower educational background frequently have reduced financial earnings, which directly impacts their life standard. These population groups are often subject to exclusion from the digital society due to limited access to digital technologies. Non-formal trainings and courses for the working-age population are important for acquiring digital competencies that address jobs generated in the digital age [29].

VI. CONCLUSION

The PCA-LR model's results, which evaluate the influence of a variety of education indicators on Internet adoption in 27 EU member states, have led to the development of the following conclusions:

- Investing in education increases the internet adoption rate by enabling individuals to develop the necessary skills to use digital technology, thereby increasing the rate of digital literacy. This strategy should also be directed towards the digitalization of education system.
- Providing public funding for higher education of socially sensitive groups that have limited access to education and facilitating their education and digital inclusion.
- Investment in higher education develops the highly educated segment of the population, responsible for the country's technological development due to its frequent use of ICT.
- Non-formal education plays an important role because it follows the fulfillment of specific market needs from the perspective of gathering additional knowledge and skills of individuals or organizations. It aids in developing a set of values that are not explicitly addressed in formal education.

The main limitation of the study is its exclusion of various socio-economic factors that can impact the access and level of education across Europe. Future research can specifically address indicators such as household income levels, as [30] has proven their impact on the level of digital adoption. They suggest that individuals with lower incomes are at risk of digital exclusion. The authors will address this limitation in their future research to determine whether economic factors are impacting education equality and digital divide.

ACKNOWLEDGMENT

This work was supported by the Serbian Ministry of Science, Technological Development and Innovation through the Mathematical Institute of the Serbian Academy of Sciences and Arts.

REFERENCES

[1] International Telecommunication Union, World Telecommunication/ICT Development Report Monitoring the WSIS Targets.

[2] UNESCO Institute for Statistics (2018), A Global Framework of Reference on Digital Literacy Skills for Indicator 4.4.2.

[3] OECD. (2019). OECD Skills Outlook 2019 Thriving in a Digital World. OECD Publishing.

[4] World Economic Forum. (2020). The Future of Jobs Report 2020.

[5] Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., & Sanghvi, S. (2017). Jobs lost, jobs gained: Workforce transitions in a time of automation. McKinsey Global Institute.

[6] Jamil S. 2021. From digital divide to digital inclusion: Challenges for wide-ranging digitalization in Pakistan. Telecommun. *Policy*, *45*(8), 102206.

[7] Van Dijk JA. 2013. A theory of the digital divide 1. In The digital divide (pp. 29-51). Routledge.

[8] International Telecommunication Union. (2021). Digital skills insights.

[9] Cedefop (2020). Empowering adults through upskilling and reskilling pathways: Vol. 1: adult population with potential for upskilling and reskilling. Luxembourg: Publications Office. Cedefop reference series, No 112.

[10] Pew Research Center. (2022). The Role of Community-Based Programs in Closing the Digital Divide.

[11] Van Deursen, A. J., & Van Dijk, J. A. (2016). Modeling traditional literacy, Internet skills and Internet usage: An empirical study. *Interacting with computers*, *28*(1), 13-26.

[12] Komarudin, K., Suherman, S., & Vidákovich, T. (2024). The RMS teaching model with brainstorming technique and student digital literacy as predictors of mathematical literacy. *Heliyon*, *10*(13).

[13] Tseng, T. H., Wu, T. Y., Lian, Y. H., & Zhuang, B. K. (2023). Developing a value-based online learning model to predict learners' reactions to internet entrepreneurship education: The moderating role of platform type. *The International Journal of Management Education, 21*(3), 100867.

[14] Segbenya, M., Bervell, B., Frimpong-Manso, E., Otoo, I. C., Andzie, T. A., & Achina, S. (2023). Artificial intelligence in higher education: Modelling the antecedents of artificial intelligence usage and effects on 21st century employability skills among postgraduate students in Ghana. *Computers and Education: Artificial Intelligence, 5*, 100188.

[15] Duong, C. D., Vu, T. N., & Ngo, T. V. N. (2023). Applying a modified technology acceptance model to explain higher education students' usage of ChatGPT: A serial multiple mediation model with knowledge sharing as a moderator. *The International Journal of Management Education, 21*(3), 100883.

[16] World data bank (2024). Indicator database. Available at: https://data.worldbank.org/

[17] European Commission (2024). Indicator database. Available at: https://ec.europa.eu/eurostat/data/database

[18] Lakhotia, R., Nagesh, C. K., & Madgula, K. (2019). Identifying Missing Component in the Bechdel Test Using Principal Component Analysis Method. *arXiv preprint arXiv:1907.03702*.

[19] Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.

[20] Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A. & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100.

[21] Weisberg, S. (2005). Applied linear regression (Vol. 528). John Wiley & Sons.

[22] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37-52.

[23] Aghion, P., Boustan, L. P., Hoxby, C. M., & Vandenbussche, J. (2019). The contribution of higher education to innovation and productivity: Evidence from U.S. states. *Journal of Economic Growth*, *24*(1), 35-82.

[24] Valero, A., & Van Reenen, J. (2019). The economic impact of universities: Evidence from across the globe. *Economics of Education Review*, *68*, 53-67.

[25] Barro, R. J., & Lee, J. W. (2013). A new data set of educational attainment in the world, 1950–2010. *Journal of Development Economics, 104*, 184-198.

[26] Bonaccorsi, A., & Daraio, C. (2007). *Universities and strategic knowledge creation: Specialization and performance in Europe*. Edward Elgar Publishing.

[27] Pihlainen, K., Korjonen-Kuusipuro, K., & Kärnä, E. (2021). Perceived benefits from non-formal digital training sessions in later life: views of older adult learners, peer tutors, and teachers. *International Journal of Lifelong Education*, *40*(2), 155-169.

[28] Creech, A., & Hallam, S. (2014). Critical geragogy: A framework for facilitating older learners in community music. *London Review of Education*, (in pre), XX.

[29] Ferreira, L. S., Infante-Moro, J. C., Infante-Moro, A., & Gallardo-Pérez, J. (2020, December). Continuous Training in Digital Skills, saving gaps between the needs and the training offer in the field of non-formal education for European Active Citizenship. In *2020 X International Conference on Virtual Campus (JICV)* (pp. 1-6). IEEE.

[30] Mulyaningsih, T., Wahyunengseh, R., & Hastjarjo, S. (2021). Poverty and digital divide: A study in urban poor neighborhoods. *J. Ilmu Sos dan Ilmu Polit*, *24*(2), 189-203.