PREDICTING THE NUMBER OF NEWBORNS IN THE REGION OF WESTERN SERBIA AND SUMADIJA USING THE REGRESSION METHOD

Andrijana Gaborović[1], MSc; Katarina Karić[2], MSc; Jelena Plašić[3] , MSc; Nikola Stanić[4], MSc

[1] University of Kragujevac, Faculty of Technical Sciences, Čačak, SERBIA, andrijana.gaborovic@ftn.kg.ac.rs
[2] University of Kragujevac, Faculty of Technical Sciences Čačak, Čačak, SERBIA, katarina.karic@ftn.kg.ac.rs
[3] University of Kragujevac, Faculty of Technical Sciences Čačak, Čačak, SERBIA, jelena.plasic@ftn.kg.ac.rs
[4] University of Kragujevac, Faculty of Technical Sciences Čačak, Čačak, SERBIA, nikola.stanic@ftn.kg.ac.rs

Abstract: *By gathering different information, studying reality, and approaching different data, in order to find patterns and templates in large data sets, the field of data mining uses various techniques for analysis and prediction. One of the techniques is the regression method, whose application will be presented in this paper. The aim of this paper is to analyze the data set on the topic of the number of newborns in the territory of Western Serbia and Šumadija, from 2011 to 2021, in order to predict the number of newborns in the next period. The results of the analysis conducted in this paper, as well as related research, predict a decrease in the number of newborns in the Republic of Serbia in the coming years, which unfortunately only confirms the continuation of negative statistics in the previous three decades.*

Keywords: *regression, newborns, predicting, data mining*

1. INTRODUCTION

The official projection of the Statistical Office of the Republic of Serbia, according to the high variant, predicts 5,174,654 inhabitants, according to the low variant 4,737,920 inhabitants, and according to the medium (most likely) variant 5,026,638 inhabitants [7]. There are numerous studies on predicting the birth rate and total population in the Republic of Serbia. Study [1] deals with the prediction of fertility, mortality, and birth rate until 2050. The author predicts, using the empirical error method, a decrease in the total population in Serbia below 7 million by 2020 and below 6.5 million by 2050. According to data from the Statistical Office of the Republic of Serbia, in 2020, the Republic of Serbia had 6,899,126 inhabitants, with every fifth inhabitant being over 65 years of age or retired.

In study [2], the author predicts, using the stochastic projection method, that Central Serbia will have between 4,688,097 and 5,179,095 inhabitants in 2032, with a 95% confidence interval, and the estimated mean value is 4,933,089 inhabitants. Stanic [3], in a paper from 2013, examines the relationship between birth rate and living standards. Using extrapolation, the author predicts that Serbia will have slightly over 6.5 million inhabitants (6529788) in 2030, with a total of 41,650 newborns in that year. This study sheds light on the potential impact of living standards on population growth, and the findings can be useful for policymakers seeking to promote population growth while improving living conditions.

The aim of this paper is to conduct a comprehensive analysis of a dataset pertaining to the number of newborns in the territory of Western Serbia and Sumadija, situated within the Republic of Serbia, covering the period from 2011 to 2021. The data analysis was carried out using the simple linear regression technique to predict the number of newborns in the forthcoming years.

2. REGRESSION TECHNIQUES AND METHODS

Regression analysis is a fundamental technique in predictive analytics that uses statistical and data mining methods to analyze past and present data to predict future events. By analyzing the relationship between the dependent and independent variables, regression analysis approximates the regression function and evaluates the functional dependency between them. The number of independent variables in the regression model determines its type, which can be:

- Simple regression, in which there is one dependent and one independent variable.

- Multiple regression involves one dependent variable but multiple independent variables. The goal of multiple regression is to uncover as many factors (independent variables) that influence the dependent variable. This model provides the best possible prediction of the dependent variable values based on the values of the independent variables if all assumptions are met. Based on the magnitude of the standardized regression coefficients, we can infer the relative influence or importance of each independent variable.

Based on the nature of the relationship between the dependent and independent variables, regression can be:
- Linear regression is a type of regression analysis that is characterized by the presence of a linear relationship between the independent variables and the dependent variable, expressed as the sum of the first-degree independent variables in the model. It refers to any approach to modeling the relationship between one or more response (dependent) variables labeled Y and one or more independent variables labeled X, in such a way that the model is linearly dependent on unknown parameters estimated from data [5].
- Nonlinear regression in statistics is a form of regression analysis in which experimental data is modeled by a function that is a nonlinear combination of model parameters and depends on one or more independent variables.

The concept of regression is implemented in almost every statistical software package, enabling the examination of the functional relationship between variables, and as such, it lies at the core of many modern statistical techniques. Therefore, the application of regression analysis can be found in almost all academic fields or applied sciences today.

3. RESEARCH TOOLS AND METHODOLOGY

The analyzed data in this paper were obtained from the open data portal provided by the government of the Republic of Serbia, from the following web address [6]. The selected category is population, specifically the number of live births in the period from 2011 to 2021 in the territory of Western Serbia and Šumadija. Regression analysis was conducted using NCSS software, which provides a complete and user-friendly collection of hundreds of statistical and graphical tools for data analysis and visualization.

Preprocessing of data is one of the most important tasks in data mining, involving data preparation and transformation into a suitable format for research methods. The techniques of cleaning, integrating, transforming, and removing redundant data are included in the preprocessing process and have been overcome using Microsoft Excel 2019 software. The data after transformation can be seen in Figure 1.

1	idindikator	godina	IDPol	nPol	Region Sumadije i Zapadne Srbije
2	18010402IND01	2011	0	Укупно	17378
3	18010402IND01	2012	0	Укупно	17931
4	18010402IND01	2013	0	Укупно	17108
5	18010402IND01	2014	0	Укупно	17579
6	18010402IND01	2015	0	Укупно	17279
7	18010402IND01	2016	0	Укупно	17168
8	18010402IND01	2017	0	Укупно	16743
9	18010402IND01	2018	0	Укупно	16589
10	18010402IND01	2019	0	Укупно	16662
11	18010402IND01	2020	0	Укупно	16186

Figure 1: Display of the dataset for analysis after transformation

After installing the software, collecting, preprocessing, and transforming data from the open data portal, analysis and discussion were carried out based on the achieved results using regression methods.

4. RESULTS AND DISCUSSION

The results obtained by applying simple linear regression method are presented below.

Figure 2 shows the number of newborns in the Western Serbia and Šumadija region from 2010 to 2021, using the simple linear regression method. The analysis result shows that the number of newborns is decreasing over time, which is also predicted in the future period.
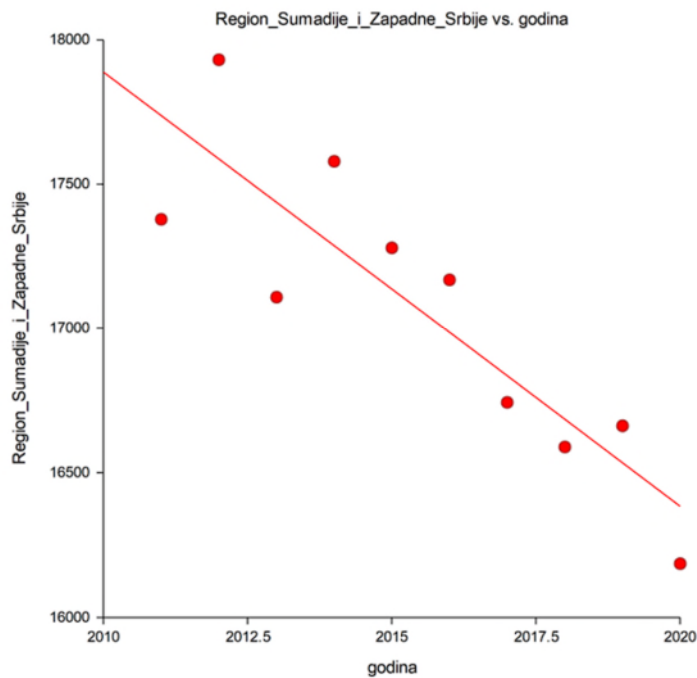
Figure 2: The number of newborns in the region of Western Serbia and Šumadija from 2010 to 2021.

Based on the previous conclusions and the established simple linear regression model, the NCSS software can predict the number of newborn babies. It is predicted that in 2022 there will be 16,084 births in the region of Western Serbia and Šumadija, and 15,933 births in 2023 (Figure 3 and Figure 4).

Predicted Values and Confidence Limits Section				
godina (X)	Region_Sumadije_i_Zapadne_Srbije	Predicted Region_Sumadije_i_Zapadne_Srbije (Yhat|X)	Standard Error of Yhat	Lower 95% Confidence Limit of Y|X	Upper 95% Confidence Limit of Y|X
2022.0000		16084.3455	208.4895	15603.5677	16565.1232

Figure 3: Prediction of the number of newborns in 2022 in the region of Western Serbia and Šumadija using simple linear regression method.

Predicted Values and Confidence Limits Section				
godina (X)	Region_Sumadije_i_Zapadne_Srbije	Predicted Region_Sumadije_i_Zapadne_Srbije (Yhat|X)	Standard Error of Yhat	Lower 95% Confidence Limit of Y|X	Upper 95% Confidence Limit of Y|X
2023.0000		15933.8909	235.6234	15390.5424	16477.2395

Figure 4: Prediction of the number of newborns in 2023 in the region of Western Serbia and Šumadija using simple linear regression method.

Based on the established linear regression model, the NCSS software can predict the number of newborns in the upcoming years. As shown in Figure 5, a decline in the number of newborns is predicted for 2022 - 16084, as well as in the following years (15934 newborns in 2023 and 15783 newborns in 2024).

	godina	Region_Sumadije_i_Zapadne_Srbije	Predicted_Region_Sumadije_i_Zapadne_S
1	2011	17378	17739.345455
2	2012	17931	17588.890909
3	2013	17108	17438.436364
4	2014	17579	17287.981818
5	2015	17279	17137.527273
6	2016	17168	16987.072727
7	2017	16743	16836.618182
8	2018	16589	16686.163636
9	2019	16662	16535.709091
10	2020	16186	16385.254545
11	2021		16234.8
12	2022		16084.345455
13	2023		15933.890909
14	2024		15783.436364

Figure 5: Predictions of the number of newborn babies in the region of Šumadija and Western Serbia for 2022, 2023 and 2024

5. COMPARATIVE ANALYSIS OF THE RESULTS OBTAINED WITH THE ANALYSIS OF RELATED RESEARCH RESULTS

After conducting an analysis using regression methods on a dataset of the number of newborns in the region of Western Serbia and Šumadija from 2011 to 2020, as well as predicting the number of children in the upcoming years, a comparison can be made between the methods used and the results obtained in this study with the results of related studies. Analysis using linear and multiple linear regression methods predicts a decline in the number of newborns in each of the upcoming years. Stanić also predicts a decline in fertility in his research [3], using extrapolation methods. It is predicted that in 2030, the number of newborns in the Republic of Serbia will be just over 40,000. In the 2007 study [2] using the method of stochastic projection, it is predicted that Central Serbia will have between 4688097 and 5179095 inhabitants in 2032, with a 95% confidence interval, with the mean estimate being 4933089 inhabitants. The official projection of the Statistical Office of the Republic of Serbia, according to the high variant, predicts 5,174,654 inhabitants, according to the low variant 4,737,920 inhabitants, and according to the medium (most likely) variant 5,026,638 inhabitants [7].

6. CONCLUSION

With the rapid advancement of technology, various software programs are being developed that allow for the analysis of large amounts of data, the identification of patterns and regularities in data, as well as the prediction of results in future periods. Thanks to software programs like NCSS, used in the research conducted in this paper, and data analysis using various regression methods, future data values can be predicted.

The results of the analysis conducted in this study, as well as related research, predict a decline in the number of newborns in Serbia in the coming years, which unfortunately confirms the continuation of negative statistics over the past three decades. Poor socio-economic status, unstable political situation, youth emigration, the COVID-19 pandemic, delayed childbirth, obstetric violence, are just some of the reasons for the decreasing number of newborns every year.

The future course of research could involve further data collection and continuous monitoring of the current situation, as well as comparison with previous data, in order to analyze on an annual basis which factors are key to stimulating fertility, and which population policy measures could have a positive impact on addressing this issue.

ACKNOWLEDGEMENTS

REFERENCES

[1] V. Nikitović, „Demografska budućnost Srbije na drugi način", UDK 303.092.5:314.8(497.11) DOI: 10.2298/STNV1302053N. Dostupno na: http://www.doiserbia.nb.rs/img/doi/0038-982X/2013/0038-982X1302053N.pdf, 2013. godine

[2] V. Nikitović, „Stohastička projekcija stanovništva Centralne Srbije na osnovu empirijskih projekcionih grešaka". Stanovništvo, god. XLV, br. 2, str. 7-31. Dostupno na: https://www.researchgate.net/publication/265825141_STOHASTICKA_PROJEKCIJA_STANOVNISTVA_CENTRALNE_SRBIJE_NA_OSNOVU_EMPIRIJSKIH_PROJEKCIONIH_GRESAKA, 2007. godine

[3] N. Stanić, M. Šekarić, „Uticaj ekonomskih faktora na natalitet u Srbiji", 14. međunarodni skup Sinergija 2013., strana 219. Dostupno na: https://www.researchgate.net/profile/Dusan-Regodic/publication/321475033_GLOBALIZACIJA_I_SAVREMENI_KONCEPT_E-PROIZVODNJA_Globalization_and_the_modern_concept_of_e-production/links/5a22cca1aca2727dd87c9553/GLOBALIZACIJA-I-SAVREMENI-KONCEPT-E-PROIZVODNJA-Globalization-and-the-modern-concept-of-e-production.pdf, 2013. Godine

[4] Javatpoint: Data Mining Techniques. Available at: Data Mining Techniques - Javatpoint

[5] H. L. Seal, „The historical development of the Gauss linear model". Biometrika. 54 (1/2): 1—24. Str, 1967. godine

[6] https://opendata.stat.gov.rs/odata/

[7] The Republic Statistical Office, Available at: https://data.stat.gov.rs/Home/Result/180203?languageCode=sr-Latn