sciendo

Polish Cartographical Review
Vol. 51, 2019, no. 3, pp. 109–116
DOI: 10.2478/pcr-2019-0010
Received: 31.07.2019
Accepted: 13.09.2019

PIOTR MICHALAK
Aalborg University
Geoinformatics
Aalborg, Denmark
orcid.org/0000-0002-4407-1682
pmicha19@student.aau.dk

Application of artificial neural networks in predicting voter turnout based on the analysis of demographic data

Abstract. The author presents the results of research on the use of artificial neural networks in predicting voter turnout. He describes the principles of operation of artificial neural networks, as well as detailed results of two machine learning methods used to predict voter turnout. The research resulted in creation of a functional model that allows for prediction of voter turnout results with a considerable degree of accuracy. The entire research process was carried out using the cartographic research method.

Keywords: artificial neural networks, voter turnout, machine learning, cartographic research method

1. Introduction

The aim of the research was to verify the possibility of using machine learning methods to predict voter turnout on the basis of demographic data. Two machine learning methods were analysed – Random Forest Regressor (RFR) and Artificial Neural Networks (ANN) – and the results of both analyses are presented in this article below. The research was carried out by means of practical application of the cartographic research method proposed in 1955 by K.A. Salishchev (1955) and developed by A.M. Berlant (1973).

The method is defined as the use of maps for describing, analysing and gaining scientific understanding of various phenomena, discovering new patterns of their distribution and mutual dependence, as well as forecasting changes. Application of this method has made it possible to approach voter turnout as a time-space phenomenon and present results and analyses in the form of cartographic models.

2. Technologies and software

Artificial neural networks are getting increasingly popular and slowly becoming inextricably linked with our daily lives. They are responsible for personalised advertising, suggested videos on YouTube, and operation of commonly available real-time translators (Y. Wu 2016). In SNN cartography, they are used in such fields as identification of land use violations or updating land cover maps (M. Golenia 2015). In a somewhat simplified version, it can be said that there are three main types of networks:

1) ANN (Artificial Neural Networks) which are used for regression and classification,

2) CNN (Convolutional Neutral Networks) which are used in computer vision,

3) RNN (Reccurent Neural Networks) which are used to conduct time series analyses.

Operation of the networks is shown in fig. 1. Input data are described in the visible input layer. It is followed by a series of hidden layers whose goal is to extract the relations between data. These layers are referred to as hidden because their values are not given in the data, and the model itself must determine which concepts are useful for explaining the relationship between the observed input data (I. Goodfellow et al. 2009). Internal network parameters are determined during neural network training. The information transmitted through the network is subjected to repeated weighting of the neuron's mathematical operation and activa-

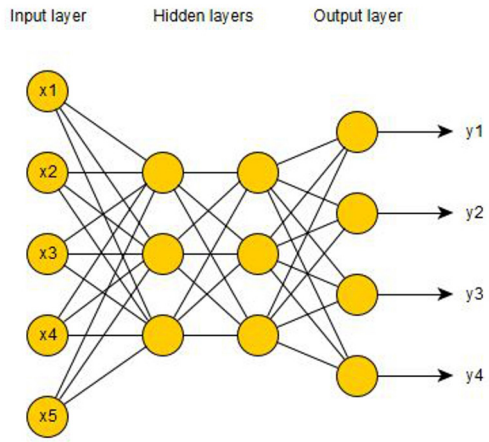

Fig. 1. An example of a network with two perception network layers (author's own work)

tion function of the hidden layers until the output function is finally executed.

Two methods, the decision tree method and the method using regular ANN, were used to examine the impact of demographic factors. In both cases, it was assumed that earlier values of voter turnout are irrelevant. The training set consisted of demographic data for communes recorded in 2005, 2007 and 2011. The data were treated independently, so data concerning the commune in 2011 constituted a separate training sample from data concerning the same commune in 2007.

During the research, the Python programming language was used together with keras and tensorflow deep learning libraries. Python is an extremely popular high-level open source programming language with many different uses ranging from web applications to artificial neural networks and big data analyses (M. Lutz 2009). The main network structure was developed in the keras library, which is an application programming interface for high-level neural networks, written in Python and capable of working within the TensorFlow library. The keras library was developed with the goal of allowing for quick experimentation in accordance with the belief that the ability to move as swiftly as possible from idea to result is the key to good research.

In addition, many of the necessary preprocessing operations were performed with the help of FME software. The Feature Manipulation Engine (FME) is a platform that streamlines the transformation of spatial data from geometric to digital formats and vice versa. It has been designed to support and cooperate with geographic information systems (GIS), as well as computer aided design (CAD) and raster graphics software. The platform was developed by Safe Software, Inc. From Surrey, located in Canadian British Columbia.

3. Training and test data

The network was built on the basis of the turnout data for the Polish Sejm election received from the National Electoral Commission and the demographic data obtained from the Central Statistical Office. The demographic data used are:

1) Number of marriages per 1000 citizens
2) Commune income per capita
3) Registered unemployment
4) Feminisation coefficient
5) Age dependency ratio
6) Population per 1 pharmacy
7) Usable floor area per 1 inhabitant
8) Population per 1 library
9) Enrolment ratio
10) Economic entities entered into the REGON register per 10,000 citizens
11) Population density
12) X-coordinate of the commune's centroid
13) Y-coordinate of the commune's centroid

RFR and ANN training data were from 2005, 2007, and 2011. Due to the nature of the research and the limited possibilities of expanding the training dataset, it was necessary to create an appropriate network architecture and select the most diverse parameters characterising communes. All 13 parameters were analysed in detail before they were incorporated into the network to isolate the specific features of individual communes and thus also of their inhabitants. Selected data define such areas of life as education, family life, sex ratio, health care, and unemployment. An autocorrelation plot was prepared for the pre-selected data (fig. 2). On the basis of this plot, we can conclude that the most important variables for determining the level of voter turnout include the feminisation coefficient, the number of economic entities entered into REGON register per 10,000 people, the number of inhabitants per one

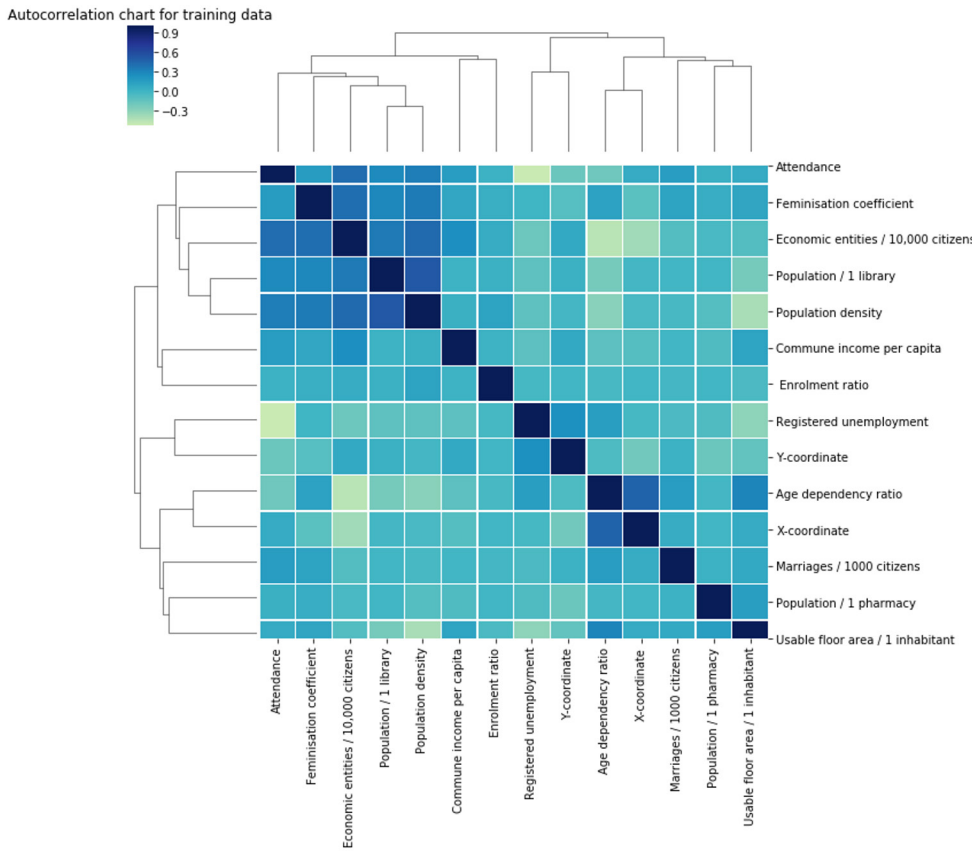

Fig. 2. The autocorrelation plot for the training data (author's own work)

library and the population density. However, the scope of the training data set was not narrowed down, because the network achieved the best results with all thirteen variables.

4. Random Forest Regressor

The first stage of research was the use of the Random Forest Regressor method which

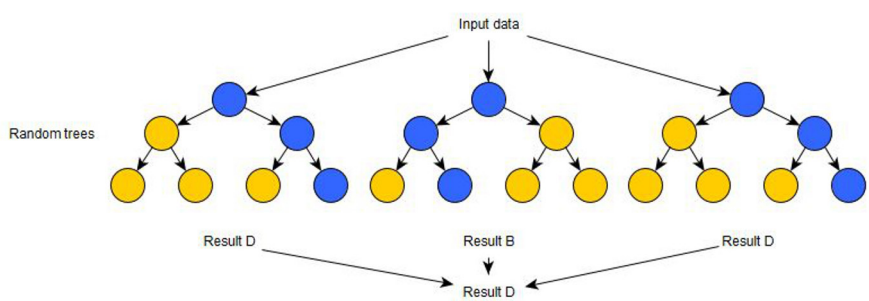

Fig. 3. The principle of operation of the decision tree method (author's own work)

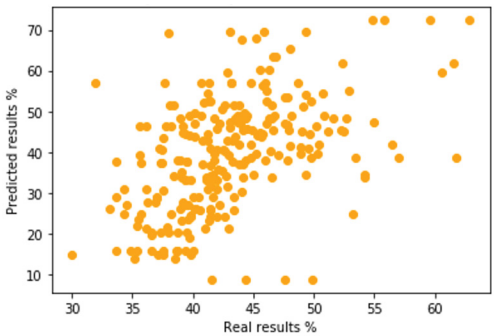

Fig. 4. Comparison of the predicted and real results for each commune in the RFR method (author's own work)

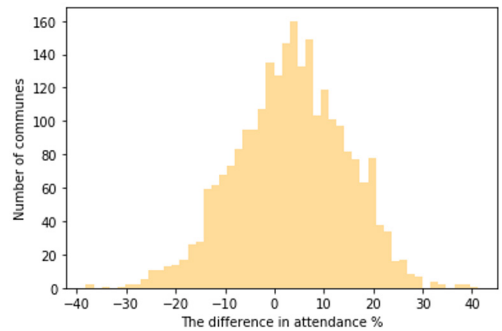

Fig. 5. The difference between the predicted and real values in the RFR model (author's own work)

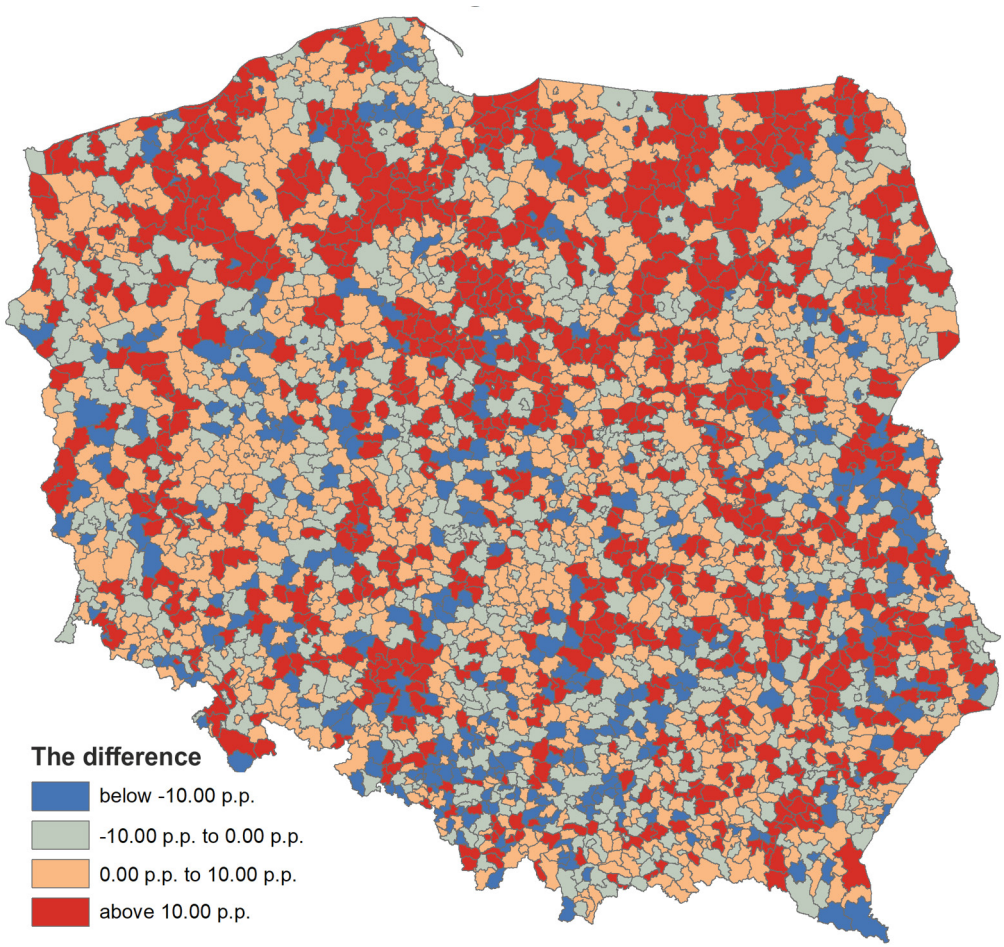

Fig. 6. The difference between the real and predicted turnout in 2015 using the RFR method (author's own work)

was built into the sklearn library. This method is based on the use of decision trees. A single decision tree asks a series of data questions, where each subsequent question narrows the range of possible values. The series of questions continues until the model reaches the confidence level which allows it to make a forecast. The order of the questions, as well as their content, is determined by the model. A single tree method multiplication, so-called random forests, is used to increase the model's independence and prevent overfitting. As can be seen in figure 3, the model uses random forests to formulate questions in the true/false formula and presents its forecast on the basis of the answers. In the case presented in figure 3, the model receives D answers for two trees, and a B answer for one tree. Therefore, the use of multiple decision trees allows to average the results and make the model independent (A. Géron 2018).

The use of the RFR method results in a mean absolute error[1] of 10.6 percentage points. This means that the average prediction of the model differs by 10.6 pp from the actual voter turnout. The results for a sample of 250 communes are presented in figure 4. The network overestimates the attendance values to a small degree, and it does not reach anywhere the value of prediction lower than 30 pp.

Presentation of the result of the RFR method in the form of a histogram (fig. 5) allows to note that it adopts a close-to-normal distribution, only slightly shifted towards positive values. Showing the difference between the actual and predicted value on a choropleth map (fig. 6) presents the model's results in the spatial dimension. The spatial distribution of clusters is largely random, which confirms the normal distribution of the histogram.

5. Artificial Neural Network

The next stage of research consisted in using a regular ANN. As already mentioned, due to the small training set, it was necessary to create the right network architecture to maximise the potential of the demographic data. The number of neural layers and network parameters, such as batch size[2], epochs[3] or learning rate[4], that allows to obtain the best results was determined in the course of the research. Figure 7 presents an overview diagram of the network architecture.

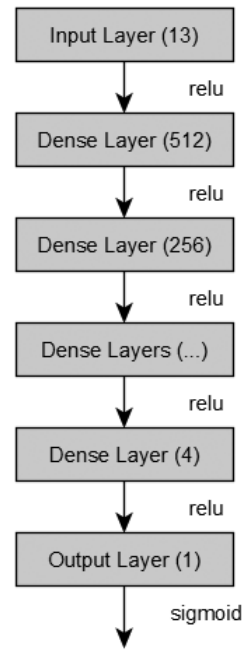

Fig. 7. Overview of architecture of a network used to predict election turnout (author's own work)

Ultimately, the learning rate that changed during the network training process was used, which increased accuracy, reducing the mean absolute error from 9.9 pp to 9.6 pp. ReLU was used as the activation function, and the output function was sigmoid. Table 1 presents the most important network parameters.

The predicted and actual attendance chart for each commune (fig. 8) indicates that the network largely understood the input data and did better than the RFR method.

[1] Mean absolute error – the average difference between the expected and actual values.

[2] Batch size – the number of samples which is processed independently by the network.

[3] Epoch – it is generally defined as "one pass through the entire data set" and is used to divide the training into different phases, which is useful for its recording and periodic evaluation.

[4] Learning rate – it determines to what extent newly acquired information overrides old information.

Table 1. The most important network parameters

Batch size	30
Epochs	1000
Learning rate	0.1 (initial)
Loss	Binary crossentropy
Optimizer	SGD

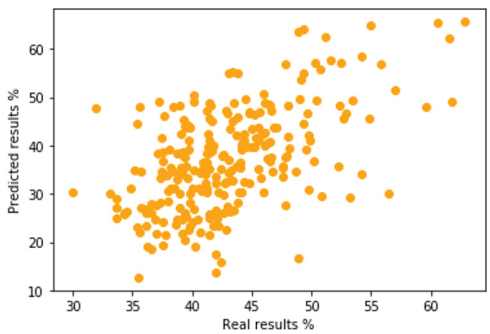

Fig. 8. Comparison of the real and predicted results with the ANN method for a sample of 250 communes (author's own work)

Preparation of the histogram (fig. 9) reveals that it also adopts a normal distribution when ANN is used. At the same time, just as in the case of the RFR model, the difference is shifted a few percentage points towards positive values.

The choropleth map (fig. 10) indicates that the network did well in the areas of Western Poland and Subcarpathia region. There are also large spatial clusters in which turnout is underestimated, including in central Poland. In the case of large agglomerations such as Warsaw or the Tri-City (Gdańsk, Sopot, Gdynia), SNN also achieves worse results than those obtained by the RFR method. In the future, it may be beneficial to combine both methods and use them where they will allow for the highest accuracy of prediction.

6. Conclusions

The research conducted by the author demonstrated that it is possible to prepare an artificial neural network that makes it possible to predict voter turnout with an average absolute error of

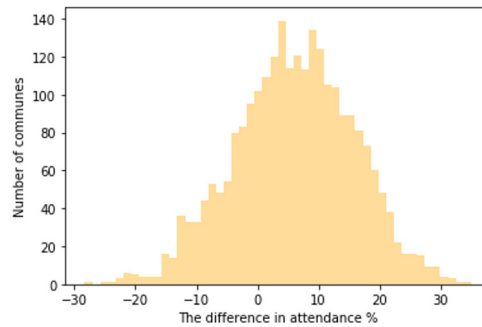

Fig. 9. The real and predicted difference in the ANN model (author's own work)

9.6 percentage points. The achieved accuracy level is therefore comparable to the levels obtained through similar applications of SNN (S.R. Khaze et al. 2013). In this case, the RFR method turned out to be less accurate and resulted in a mean error of 10.6 percentage points. The results can therefore be described as satisfactory given that the voter turnout is also determined by non-geographical factors, such as the political situation in the country and the world, new political trends or even the climate change (K. Dośpiał-Borysiak 2015). Such factors are extremely difficult to model, which makes it virtually impossible to incorporate them into the network.

In 2018, a report was issued whose purpose was to verify the survey data by comparing them with the actual results of local government elections in 2018. Comparison with the results provided by the National Electoral Commission revealed that as many as 18 surveys which were carried out in cities had the cumulative error[5] of 30 pp or more. Only the best 7 polls achieved the cumulative error of below 9.6 pp (S. Ossowski, M. Kilian 2018). The above-presented results indicate that in the future it should be possible to replace costly and inaccurate surveys with an artificial neural network which is constantly increasing its accuracy as the amount of training data increases in subsequent years.

[5] Cumulative error – the sum of differences between the support level registered in a survey and the actual result of a specific candidate reported by the National Electoral Commission expressed in percentage points.

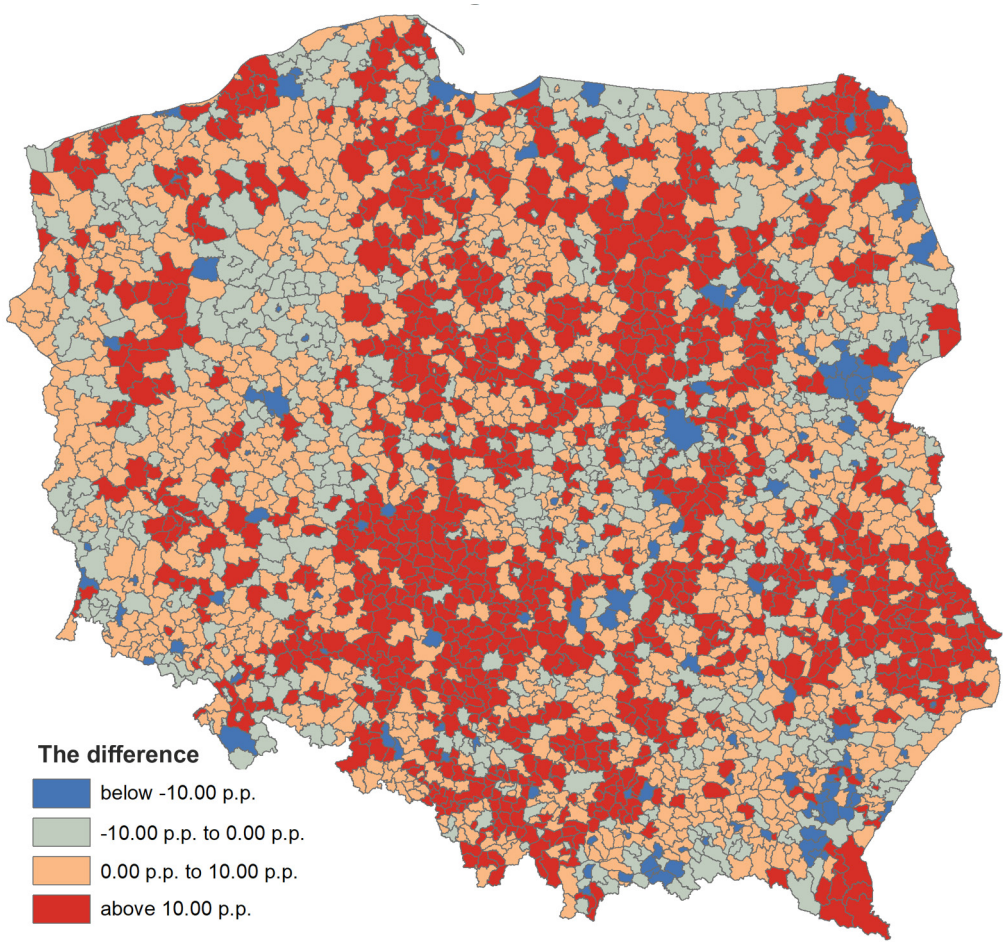

Fig. 10. The difference between the real and the predicted frequency in 2015 using ANN (author's own work)

7. Recapitulation

The on-going development of artificial neural networks means that they interdisciplinary use can be gradually broadened, and that they can be also applied in cartography. Together with political sciences, they will make it possible to manage the state more effectively and better predict e.g. voter turnout, or results of elections or referendums. It can also create new threats, e.g. target marketing[6]. It will be necessary to consider such questions as the following – "Can X Party, knowing that it has a 45% chance of winning the election in Y poviat, focus its campaign in this area to increase its chances?" To what extent are such activities compliant with the democratic principles? Where will we set the boundary between an election campaign and a case of manipulation of election results? In the coming years, we will have to answer such questions and find effective solutions within our democratic system.

[6] Target marketing – activities aimed at increasing the popularity of a company or party by directing personalised messages to a selected group of recipients.

Literature

Bartman J., Bajda K., 2014*, Wykorzystanie sztucznych sieci neuronowych do prognozowania wyników meczów piłkarskich*. "Edukacja-Technika-Informatyka" T. 5, pp. 425–431.

Berlant A.M., 1973*, Problemy teorii wykorzystania map w badaniach naukowych*. In: *Kartograficzna metoda badań w geografii*. "Przegląd Zagranicznej Literatury Geograficznej" z. 3/4, pp. 39–50.

Dośpiał-Borysiak K., 2015*, Norweskie partie polityczne wobec problemu zmian klimatu i polityki klimatycznej*. Warszawa: Szkoła Główna Handlowa.

Géron A., 2018*, Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Helion.

Golenia M., Zagajewski B., Ochtyra A., 2005*, Zastosowanie sztucznych sieci neuronowych do aktualizacji map pokrycia terenu Corine*. "Polski Przegląd Kartograficzny" T. 47, nr 3-4, pp. 257–266.

Goodfellow I., Bengio Y., Courville A., 2009*, Deep Learning*. Boston*:* Massachusetts Institute of Technology (MIT).

Khaze S. R., Mohammed M., Hojjatkhah S., 2013*, Application of artificial neural networks in estimating participation in elections*. "International Journal of Information Technology, Modelling and Computing (IJITMC)" Vol. 1, no. 3.

Lutz M., 2009, *Learning Python*. 5th edition, Helion.

Ossowski S., Kilian M., 2018*, Trafność sondaży przed wyborami samorządowymi 2018*. Poznań: Wydział Nauk Politycznych i Dziennikarstwa UAM.

Salishchev K.A., 1955, *O kartograficheskom metode issledowania*. „Vestnik Moskovskogo Universiteta, Ser. fiziko-mat." no. 10, pp. 161–170.

Wu Y., Schuster M., Chen Z., Le Q., Norouzi M., 2016, *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. Cornell University.