The Use of Self-organizing Maps (SOM) in Demographic Analysis

Nagy Balázs*, Majdán Márk*, Dr. Varga Valéria*, Dr. Nagyné Dr. Hajnal Éva*
* Óbuda University, Alba Regia Technical Faculty

***Abstract*—The present study investigates Hungary's higher education student numbers based on demographic as well as education system statistics. The following data sources were used: Central Statistical Office (Hungary), education information portal felvi.hu and Educational Authority (Hungary). The applied method, Kohonen's self-organizing map (SOM). In the interval between 2000 and 2012, the following statistics were used: total population, students in secondary education, students in higher education, those completing secondary education, fresh graduates and the number of students admitted to higher education; in the interval between 1982 and 2012 statistics of infant mortality and live births were used. We created two statistical scenarios to predict the number of years spent in higher education. After importing data into a SOM data structure and training the map, two clusters were recognizable, the separating line between the two being the year 2008. Low quantization and topographic error values indicate the high quality of data processing. After traing the map, we made a prediction for the interval 2001-2030. Predicted values were retrieved after calculating the mean value of neurons linked to values in the weight vector matrix. The results suggest that by 2030, higher education student numbers will decrease by a percentage 30 to 35, as compared to the numbers in 2011.**

I. INTRODUCTION

The research and development (commonly abbreviated as R&D) sector is a noteworthy factor in an economy due to its role in a country's competitiveness. The sector can be described with many indicators, however, higher education numbers are of special importance, as universities provide a new generation of research staff. The further advantage of this approach is that higher education statistics are widely available.

Higher education student numbers are determined by demographics and legal changes in education policy. In Hungary, decreasing demographic tendencies result in decreasing student numbers. A suitable method is necessary in order to be able to estimate the exact tendency of this change.

Although not primarily used for prediction and estimation, the method we chose is widely used for data analysis. In the present study we give an overview of how SOMs may be used for prediction and estimation, through establishing and evaluating a statistical model capable of estimating higher education student numbers from demographic data with good approximation.

Higher education institutions, apart from providing academic education, participate in various R&D activities. The border between universities as degree-giving instiutions and research institutes is hard to tell. Students wishing to obtain a PhD degree are required to have done a certain amount of research work, from which further publications and experimental results originate, even if students themselves do not stay on a researcher's career path.

In the past twenty years it has become common to build so-called knowledge centres, where representative offices of companies, state-funded or privately owned research institutes work on joint projects at the same place. This type of research cooperation is not fully developed in Hungary.

The structure and quality of Hungary's higher education is essentially determined by education policy reforms, such as the introduction of the so-called two-level exam system, replacing university entrance exams.

The Bologna Process, introduced in Hungary in 2006, has a significant effect on the structure and quality of academic education. The formerly five-year-long university courses (apart from a few) were divided into a three-year-long bachelor (BSc/BA) and a two-year-old master's (MSc/MA) course. Similarly the variety of university majors were also reduced: compared to 400 majors in 2005, applicants could only choose from 108 majors in 2006.

The decade between 2000 and 2010 was characterized by an increase in higher education student numbers. Based on higher education students per total population ratio, more people enrolled in university courses than ever before. Some distorting factors, however, should be taken into account. Applicants for master's courses after 2006 should also be counted in statistics of university applicant numbers.

Despite the above mentioned reforms, the most significant factor that affected higher education student numbers was demographics. In an analysis of long time series statistics, decreasing population and decreasing student numbers are clearly to be seen. These tendencies are only slightly changed by fluctuations caused by education reforms. Similarly to how age groups (demographic cohorts) can be tracked through the statistics of a population, a relatively good estimation can also be given for the ratio of students being admitted to higher education.

II. DATA AND METHODS

A. Data selection

Data from various sources were used in the present study. The main data sources were Hungary's Central Statistical Office (KSH) [1][2], education portals felvi.hu [3] and Education Authority [4]. In total, 152 data values

were used as well as data derived from the original data set.

The following from KSH were used: total population, mortality and birth rates, education statistics. Demographic data included birth rate (1982-2012; a total of 31 data values), total population (2000-2012; 13 data values) and infant mortality (1982-2012; 31 data values).

Education statistics included the following time series data: students in secondary education (2000-2012; 13 data values), students in higher education (2000-2012; 13 data values), students completing secondary education (2000-2012; 13 data values), higher education students receiving a degree (2000-2012; 13 data values).

Higher education data were retrieved from the website felvi.hu and the Education Authority. From felvi.hu's statistics, the following data sets were used: statistics of recent years, number of students applying for a university and number of those admitted to a university, from which we retrieved the number of students admitted to a university (2001-2012, 12 data values).

We obtained the number of first-year university students using Education Authority's higher education data (2000-2012, 13 data values).

B. An overview of the model

We created the prediction using artificial neural networks. Artificial neural networks are widely used to solve mathematical problems, also in statistical data processing, especially in the case of large and complex data sets. Pattern recognition neural networks are also used in face recognition [5][6][7] and optical character recognition, but also in robot navigation [8] (to avoid obstacles in a robot's path).

The method of choice is Kohonen's self-organizing map, abbreviated as SOM, created by Teuvo Kohonen [9].

This type of artificial neural network belongs to neural networks with unsupervised learning, this means that only an input is given without any reward signal. In a neural network, neuron connections are structured in a two-layer topology that consists of an input and an output layer. Some neural networks have further layers, also known as hidden layers, but Kohonen SOMs only have the two basic neuron layers. The output layer is two-dimensional and map-like, hence its name. Neurons in the output layer are interconnected and are located in a hexagonal or rectangular lattice.

Input signals activate neurons in certain areas of the output layer in case of similar activation patterns. It is important to know how neurons are interconnected in order to be able to optimalise the activation patterns. The optimalisation process creates a topological map of the input signals and transforms similarities between input signals into a neighbourhood indicator between neurons. This results in the map being able to ignore disturbing factors to take into account only the most predominant similarities (connections). In short, Kohonen's SOM compares input signals (data values), then uses these numeric values to calculate the neighbourhood of output neurons.

Each neuron's position in the lattice can be described with a weight vector assigned to it. Similarly to biological neurons, artificial neurons are also in a network of synaptic connections. This similarity can even be extended to the two basic types of synapses, inhibitory and excitatory, the principle of which also exists in artificial neural networks.

Kohonen's SOM algorithm can be broken down into the following main steps. Let x_k denote a training sample of n dimensions and w_{ij} the neuron to be selected. The distance between neurons is defined by the neighbourhood function $h(w_{ij}, w_{mn})$. The neighbourhood function returns high values if neurons are close to each other in the output space. The selected winner neuron from the input data is w_{winner} (best-matching unit, abbreviated as BMU).

The following steps are completed for each input data value:

1. The distance between the input data value and all the other neurons is calculated one by one.

$$d_{ij} = \| x_k - w_{ij} \|$$

2. The closest neuron (the one with the smallest distance) is the winner neuron.

$$w_{nyertes}(w_{ij} : d_{ij} = \min(d_{mn}))$$

3. After selecting the winner neuron, the values of all the other neurons are updated.

$$w_{ij} = w_{ij} + \alpha \cdot h(w_{nyertes}, w_{ij}) \cdot \| x_k - w_{ij} \|$$

4. The process restarts after the third step and and the distance between each input value and the winner neuron is calculated. The process is repeated until it reaches a previously set condition such as the number of iterations.

The distance between weight vectors can be calculated with different mathematical methods, but Euclidean metric is the most often used and we also used this type of metric.

The SOM updates weight vector values one by one until they get close to the BMU. This type of algorithm is called an adaptive algorithm. Learning rate (α) is a numerical value that determines how close neurons can get to the winner neuron in each learning epoch. Training can be iterative, when the entire data set is processed through iterations (steps) in the above mentioned way.

Another approach is batch training when the entire data set is partitioned into Voronoi cells based on similarities between data values. Voronoi cells are grouped around a neuron. By default, most self-organizing maps use iterative training.

After training the SOM, patterns close to each other in the input space are close to each other in the output space and vice versa. Thus, the SOM training algorithm preserves topology (neighbourhood values). Regardless of the number of training epochs, there is always a difference between the input value and the value assigned to a neuron. This is known as quantization error. It is worth to note that SOM training algorithm is similar to vector quantization (VQ), a type of unsupervised learning used in digital signal processing and data processing.

The output space can be described with topographic error. This number is calculated by measuring the distance of an input vector's BMU and its second BMU.

Aside from visualizing the output layer itself, a SOM also visualizes the U-matrix, which denotes with colour tones the Euclidean distance of weight vectors of neighbouring neurons. It provides useful information related to the entire data set. Clusters are denoted by light colours, whereas borders between clusters are denoted by dark colours.

SOMs, as clustering methods, are widely used for processing economic data. Pablo-Martí [10] demonstrated the regional distribution of economic activities in the Spanish economy; Collan, Eklund and Back [11] used a SOM to create a map that groups the countries of the world by their economic development. The above examples mostly take advantage of the two-dimensional, map-like visualization, as this type of visualization can be projected onto a geographical map.

Several examples can be found in bibliography for SOMs used as statistical estimation methods [12], often in various areas of sciences. Al-Shayea et al. [13] created a data prediction model from the financial data of sixty-six Spanish banks that is capable of predicting bank insolvency. Ultsch and Röske [14] created a sea level prediction model from historic meteorological data using a SOM.

C. Creating the model

We used Microsoft Office 2010 Excel to make estimations from the data available. Thus, we obtained the number of the 18-year-old age group by shifting the number of live births (2000-2030, 31 data values), the number of higher education students minus the number of those applying for a university (2000-2012, 13 data values) and the number of years spent in higher education (2001-2012, 12 data values). We obtained this by taking the number of students in higher education minus the number of those applying for a university and dividing it by the number of freshly graduated students.

The majority of estimations were conducted with Microsoft Excel. As for the final prediction, however, we used MATLAB and the MATLAB function library SOM Toolbox. We used an Excel estimation (2013-2030) to obtain the number of those receiving a university degree (18 data values), the number of years spent in higher education in two scenarios. One scenario is that the number of years spent in higher education follows an exponential trendline. The second scenario is that the number of years is at a constant average. Thus, we obtained the number of students minus those applying for a university (18 data values), the number of students * years and the number of students.

Sixty data values were used to conduct the MATLAB prediction. These were the following: the number of students * years spent in higher education, the number of the 18-year-old age group, the number of years spent in higher education, the number of those receiving a university degree and the number of students completing secondary education. In order to reveal connections between data values we used SOM Toolbox, which visualizes clusters and gradients to analyse data connections.

We used various settings to train the SOM. All the 60 available data values were imported into a sample matrix of the dimensions 12×5. A data struct ‘D’ was created from the sample matrix, then the data struct was normalized, using linear ‘range’ type normalization. ‘Range’ type means that data values fall into an interval between 0 and 1, proportional to the biggest value of the data series. After this, we trained the SOM from the normalized data using Euclidean metric to measure distances. Euclidean metric is based on Pythagorean theorem and bibliography suggests it is a robust metric to be used in this case.

To create the prediction model, we used a SOM similar to the previous one, but in this case with data in the the interval between 2001 and 2030. The model was created in two scenarios. In the first one we supposed that the number of years spent in higher education follows an exponential trendline, whereas in the second one the number of years was a constant mean value. After training the map, the four neurons (BMUs) closest to each weight vector were determined. Then BMUs of each row in the weight vector matrix were placed in the variable ‘index’. This variable helped us to find values in the weight vector matrix associated to the index. The next step was creating a matrix in order to store the mean value of weight vectors. The final step was data denormalization, after which we obtained the results.

III. RESULTS AND DISCUSSION

In the first step, we provided a common number format for all the data series available.

We investigated the number of participants in secondary education, those enrolled in higher education, those completing secondary education and those receiving a degree.

The number of those completing secondary education, those enrolled in secondary education and those receiving a degree began stagnating in 2006 and after 2008 a drop in their numbers can be observed. However, in the years after 2008, the numbers reflect the values before the rapid decrease and are constant apart from minor fluctuations.

In the higher education, however, different tendencies can be observed. One of the reasons inducing an increase in student numbers was an 1985 law on higher education and its modification imposed in 1990, as these caused large numbers of students to be admitted to higher education. The subsequent laws on higher education also increased student numbers.

The increase reached its peak point in 2006, when the new law on higher education was imposed, thus introducing the Bologna Process into Hungary’s education system. This caused a drop in student numbers in Hungary. One reason for this was that by the new law, students had to pay for subjects not included in the university curriculum. Another reason is that stricter regulation was introduced concerning language proficiency as prerequisite of graduation and also for university admission.

The most significant decrease starts from 2008. This is partly due to the effects of the economic crisis and also due to growing tuition fees.

Our aim was to give an estimation for higher education student numbers twenty years ahead. As the first step, we determined the number of the 18-year-old age group by shifting birth numbers 18 years ahead in time. As infant mortality is low, it is not a significant factor that would distort our estimation.

Knowing the number of the 18-year-old age group, the second step was determining what percentage of this age group completed secondary education and was enrolled in higher education. In the last decade, the number of students completing secondary education stagnated at approx. 71.13%. We then applied this ratio to the number of the 18-year-old age group.

After this we determined the number of students in higher education by subtracting the number of recently admitted students from the total higher education student number. We took into account not just bachelor students, but also those enrolled in master's and PhD courses. We then divided the resulting value by the number of those receiving a degree, thereby giving an estimated mean value of 5.5 years for the number of years spent in higher education.

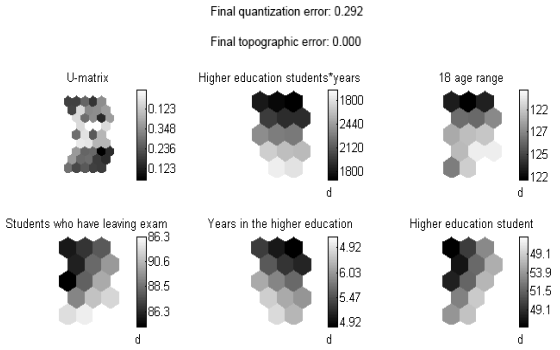

Fig. 1: A SOM after training with the higher education sample data. Five different variables were used to train the SOM. The U-matrix shows two, clearly distinguishable clusters, the separating value between the two clusters is the year 2008.

In order to estimate the number of higher education students we had to determine the number of those receiving a degree and those having spent at least one semester in higher education. We created a diagram from these data and drew a linear trendline to determine the number of those receiving a degree twenty years ahead. In the period between 2005 and 2008, the number of those receiving a degree showed a decreasing tendency, followed by a temporary increase between 2008 and 2011, when it started to decrease again. If this decreasing tendency continues, then student numbers are likely to drop to 40.000 in 2030. However, this is still a rough estimation, which needs to be improved via the SOM method.

Now all the necessary data were available to us to train a SOM and use it for prediction. In order to create the neural network we used the number of students*years spent in higher education, the number of the 18-year-old age group, the number of years, the number of those receiving a degree and those completing secondary education, all in the period between 2001-2011.

The map was created using 55 data values that were imported into a MATLAB variable. After creating the map, we visualized the results. At this stage there was no need to make any distinction between the first and the second scenarios, as in the period between 2001-2011 all the data values were available. Two clearly distinguishable clusters were present in the visualization, with the year 2008 separating them.

Quantization error shows the mean distance between input vectors and their BMUs. Topographic error shows the percentage of data values where related BMUs are not neighbours to each other. In this case both error values were considerably low. Quantization error equals to 29%, whereas topographic error has a value of 0 %, which means related BMUs are in each other's neighbourhood.

We then estimated student numbers with a linear trendline from 2011 to 2030.

To estimate the student numbers we used two different methods. The first one was drawing a linear trendline with Excel.

Student numbers in 2030 were given using the equation of the Excel trendline. According to the results, student numbers are likely to decrease by 50% from present day to 2030. However, a linear trendline does not fit perfectly to the data series, as it would drop into the negative domain.

To give a more exact estimation for the one conducted with a linear trendline, we also made a prediction using SOM. As a first step, we multiplied the results from linear trendline estimation by the mean value of years spent in higher education (first scenario), then by the number of years from exponential trendline (second scenario). The two variants are largely similar, thus we expected similar results.

Now we had all the necessary data to conduct the SOM prediction. As with the data in the period between 2001 and 2011, we imported data in the period between 2001 and 2030 into MATLAB. After data normalization, the map was trained.

In this case, there are several data values where related BMUs are not in each other's neighbourhood. The bigger the topographic error value is, the more complex the output space is. Cluster patterns and topology itself in these cases are more complex. A high number suggests errors in the training process. Quantization error had a value of 9 and 10 %, respectively, topographic error had a value of 0 and 3 %, respectively, thus we can say the model proved to be reliable.

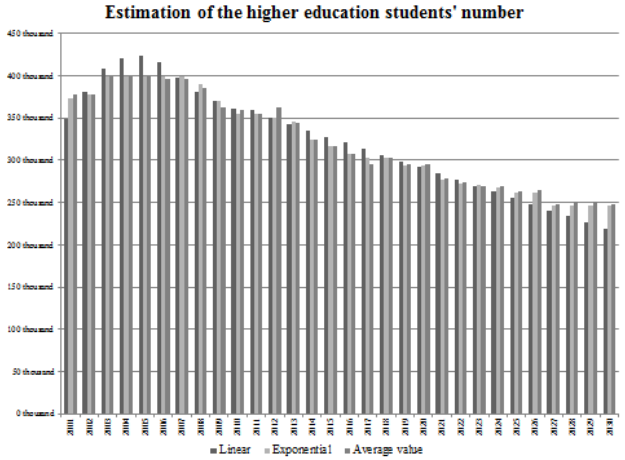

Fig. 2: Estimation of student numbers with linear trendline and with SOM (two different scenarios).

After creating the map we obtained the four neurons closest to each weight vector and looked up the values associated to them in the weight vector matrix. We obtained prediction results by calculating the mean of

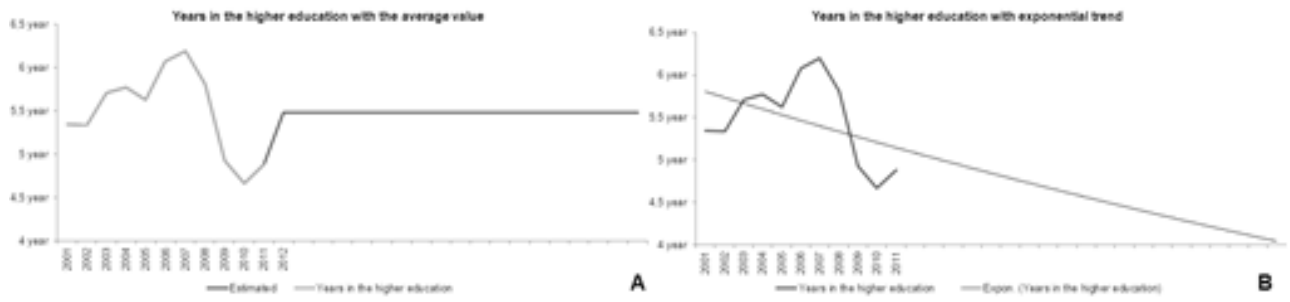

Fig. 3: The estimation of the years spent in the higher education. For the estimation data were used between 2001 and 2011. As seen on the graph the values changed between 4.5 and 6.5 years. A: On the graph the data was estimated with the average values. B: The exponential trend estimation shows that the number in 2030 might be between 4 and 4.5.

these values. Some recurrent values were also obtained because the method is extrema sensitive. This means that in some cases the same neurons are found to be the closest neurons, thus returns its values.

According to the SOM estimation, a decrease of 30% in student numbers is likely to occur in the first scenarios, whereas the second scenario predicts a decrease of 35% as compared to the situation in 2011. The linear trendline estimation thus gives a rough estimation as compared to the SOM prediction. Results only suggest one possible outcome, but many other should also be taken into account, such as students working and studying abroad. A larger amount of data is also necessary to validate the method.

IV. Conclusion

In the present study we tried to demonstrate the use of Kohonen's self-organizing map in the analysis of economic data through a case study. We estimated the number of higher education students twenty years ahead using SOM. According to data processing quality indicators, such as quantization error and topographic error, the model gives a good approximation of the observed tendencies. The present data quantity is not enough to test the validity of the method, but the method should be applied in case of larger quantities of historic time series data.

The decrease in student numbers are primarily caused by demographic tendencies, however, these were temporarily compensated by legal changes in the higher education. The tendency has been becoming more prevalent since 2008. Higher education produces researchers; significant R&D and innovation takes place in the sector itself, thus changes to higher education strongly determine the future situation of R&D as a sector of economy. Our study provides a basis for the research of these economic impacts.

In the future we would like to improve the prediction model based on several different scenarios. It should also be examined how the present rate of decrease of student numbers will affect the numbers of researchers and also the number of scientific publications and patents. In another scenario, an estimation should be given of the amount of financial assets necessary in order to maintain the present-day level of R&D activity while the number of research staff is decreasing.

Acknowledgment

We acknowledge the financial support of this work by the Hungarian State and the European Community under the TÁMOP-4.2.2.B-15/1/KONV-2015-0010 project entitled "Development of Alba Regia Technical Faculty's Scientific Workshops".

REFERENCES

[1] K. S. Hivatal, "Népesség, népmozgalom 1949-" 2013. [Online]. https://www.ksh.hu/docs/hun/xstadat/xstadat_hosszu/h_wdsd001a.html.

[2] K. S. Hivatal, "Oktatás (1960-)," 2013. [Online]. https://www.ksh.hu/docs/hun/xstadat/xstadat_hosszu/h_wdsi001b.html.

[3] Felvi.hu, "Elmúlt évek statisztikái." 2013. [Online]. http://www.felvi.hu/felveteli/ponthatarok_rangsorok/elmult_evek.

[4] Oktatási Hivatal, "Kezdő évfolyamos hallgatók száma képzési szintenként," 2013. [Online]. http://www.oktatas.hu/felsooktatas/felsooktatasi_statisztikak/!DARI_FelsooktStat/fir/fir_stat2010/stat2010_113.xls.

[5] J. L. Alba, A. Pujol, and J. J. Villanueva, "Separating geometry from texture to improve face analysis," in IEEE International Conference on Image Processing, 2001, vol. 2, pp. 673–676.

[6] C. H. Lee, D. S. Seong, and K. H. Park, "Face recognition using self-organizing map," J. Korea Inf. Sci. Soc., vol. 20, no. 11, pp. 1730–1738, Nov. 1993.

[7] Z. Q. Liu, "Adaptive subspace self-organizing map and its applications in face recognition.," International Journal of Image and Graphics Vol. 2, P. A. ; Garrido Frenich, A. ; Torres, J. A. ; Pulido Bosch, A., no. 4, pp. 519–540, 2002.

[8] I. K., N. S., and U. T., "A self-organizing map based navigation system for an underwater robot," in 2004 IEEE International Conference on Robotics and Automation IEEE Vol. 5, 2004, p. 5726.

[9] T. Kohonen, "Self-organizing formation of topologically correct feature maps," Biol. Cyb., vol. 43, no. 1, pp. 59–69, 1982.

[10] J. M. Pablo-Martí, F.; Arauzo-Carod, "Spatial distribution of economic activities : an empirical approach using self-organizing maps," in Rethinking the economic region. New possibilities of regional analysis from data at small scale, E. F.-V. F. Rubiera-Morollón, Ed. Springer, pp. 1–41.

[11] M. Collan, T. Eklund, and B. Back, "Using the Self-Organizing Map to Visualize and Explore Socio-Economic Development," EBS Rev. 22(1), 6-15, vol. 22, no. 22, 2007.

[12] C. Science, M. Po, T. Honkela, and T. Kohonen, Bibliography of self-organizing map (som) papers: 2002–2005 addendum. 2009, pp. 2002–2005.

[13] Q. K. Al-shayea, G. A. El-refae, and S. F. El-itter, "Neural Networks in Bank Insolvency Prediction," IJCSNS International Journal of Computer Sciences and Network Security VOL.10 No.5, vol. 10, no. 5, pp. 240–245, 2010.

[14] A. Ultsch and F. Röske, "Self-organizing feature maps predicting sea levels," Self-Organizing Feature Maps Prediciting Sea Levels, Inf. Sci. 144/Elsevier, pp 91 - 125, vol. 144, pp. 91–125, 2002.