Z. geol. Wiss., Berlin 40 (2012) 4/5: 253 – 266, 8 Abb.

Use of neural networks in the geochemical data interpretation

Anwendung neuronaler Netzwerke zur Interpretation geochemischer Daten

Uporaba nevronskih mrež za interpretacijo geokemičnih podatkov

Примена на невронски мрежи во интерпретација на геохемиски податоци

Gorazd Žibret, Ljubljana; Robert Šajn, Ljubljana; Jasminka Alijagić, Ljubljana; Trajče Stafilov, Skopje

Key words: neural networks, data interpretation, clustering, classification, function approximation, multilayer perceptron.

Abstract

Aim of the study is to represent some interpretation methods in geochemistry using an artificial neural networks technology. Four simple case studies are presented. The first case is the modelling of the Zn concentration in the attic dust in the Celje (Slovenia) area applying a multilayer perceptron with a back propagation learning algorithm. The second case study includes a classification problem. By means of a neural network, a classification of different sampling media (such as attic dust, soil, and alluvial sediments) will be made according to their chemical composition. The third case studies a clustering problem where the geochemical classification of chemical elements in Mežica (Slovenia) has been performed. In the last case study, the modelling of a trace element contamination in soil samples taken at Kavadarci (Macedonia) by the multilayer perceptrons method has been compared with the kriging interpolation method.

Zusammenfassung

Der Aufsatz behandelt vier Fallstudien für die Anwendung künstlicher neuronaler Netzwerke zur Interpretation geochemischer Daten. Im ersten Fall wird eine relativ einfache räumliche Interpolation zur Erklärung des Zn-Gehaltes in der Nachbarschaft der auflässigen Zinkhütte in Celje (Abb. 1) diskutiert. Die Interpolation erfolgte auf der Basis eines Multilayer-Perzeptrons, das über eine Recall-Lernphase an 118 Probenahmeorten für den Hüttenstaub „geeicht“ wurde. Das Ergebnis wurde mit der herkömmlichen Methode zur Konstruktion geochemischer Karten, dem Kriging, verglichen (Abb. 2). Auf den mittels neuronaler Netze konstruierten Karten ist erkennbar, dass die topographische Höhe einen entscheidenden Einfluss auf die Prognose des Zinkgehaltes hat. Das entspricht Beobachtungen, nach denen die Konzentrationen über der Zone der Temperaturinversion mit der Höhe extrem abnehmen. Dieses Problem kann mit Hilfe von Krigingverfahren nicht bearbeitet werden. Ein weiteres Phänomen, das aus den mit neuronalen Netzwerken erzeugten Karten erkennbar, aber durch Kriging-Interpolation nur schwer zu simulieren ist, besteht in der Berücksichtigung des vorherrschenden Westwindes. Einen Nachteil neuronaler Netzwerke stellt das Auftreten eines Rauschens in den fertigen Karten dar, welches möglicherweise nicht durch die geschilderten Phänomene bedingt ist.

Im zweiten Beispiel wird eine Klassifizierung untersucht. Verschiedene natürliche Kompartments des Celje-Gebietes (Abb. 1) wie Boden, alluviale Sedimente und Hüttenstaub wurden auf der Grundlage der in ihnen enthaltenen Spurenelemente mit Hilfe neuronaler Netzwerke klassifiziert. Die Lernphase des Netzwerkes (Abb. 3) basierte auf 121 Proben Hüttenstaub, 129 Boden- und 95 Sedimentproben. Ferner wurden der Lernmatrix 180 Muster mit Zufallswerten beigefügt. Die Typen der Probenahmemedien wurden durch die drei output-Neuronen repräsentiert. Die Ergebnisse hingegen sind mit Hilfe von 37 Mustern bewertet worden, die nicht in die Lernmatrix eingefügt worden waren. Das Netzwerk reflektiert die 37 Muster korrekt. Damit wurde belegt, dass das Netzwerk in der Lage ist, Muster mit zufälligen Werten zu erkennen.

Als drittes Beispiel wird die Bestimmung geochemischer Gruppen mittels selbst-organisierender Karten (SOM) behandelt. Dazu wurden die Daten einer geochemischen Probenahmekampagne für Boden und Hüttenstaub in Mežica (Abb. 1) verwendet. Es sollte ermittelt werden, welche Spurenelemente in der Umwelt Folge berg- und hüttenmännischer Aktivitäten in der Vergangenheit, eisenverarbeitender Tätigkeit und schließlich des natürlichen Untergrundes sind, also des Einflusses karbonatischer sowie von Eruptiv- und metamorphen Gesteinen. Die geochemischen Assoziationen lassen sich aus der Matrix des output-Netzes ablesen (Abb. 4), in dem verwandtere Elemente eine ähnlichere Verteilung gegenüber Elementen, die weiter voneinander entfernt sind, aufweisen (Abb. 5). Damit wurden fünf chemische Assoziationen identifiziert: eine Assoziation der anthropogenen Kontamination, eine solche der Karbonatgesteine, eine Gruppe spurenelementführender Minerale, eine Assoziation felsischer Eruptiv- und metamorpher Gesteine sowie die Gruppe der Kationen.

Im letzten Fall sind die geochemischen Karten des Gebietes von Kavadarci in Makedonien (Abb. 1) modelliert worden. Er ähnelt der ersten Fallstudie, aber er unterscheidet sich auch von ihr. Eine geochemische Assoziation der Elemente As, Sb und Tl wurde in den alluvialen Sedimenten der Flüsse Crna und Vardar gefunden. Sie ist Ergebnis der Erosion des anstehenden Gesteins wie auch bergbaulicher Aktivitäten in der Vergangenheit. Mittels Kriging lässt sich nur eine sehr ungenügende Interpretation konstruieren, in der extrem hohe Gehaltswerte „Bullaugeneffekte“ erzeugen. Hohe Werte besonderer Spurenelemente als Folge natürlicher Prozesse konnten nicht gefunden werden. In die neuronalen Netzwerke wurden folgende Parameter einbezogen: geographische Lage und topographische Höhe, geologische Situation, Flächennutzung und Hangneigung. Gegenüber dem Kriging ergaben die neuronalen Netzwerke eine deutlich bessere Verallgemeinerung (Abb. 6 bis 8).

Povzetek

V prispevku smo avtorji poskusili na kratko predstaviti 4 primere uporabe nevronskih mrež pri obdelavi geokemičnih podatkov. Prvi primer je precej enostavna prostorska interpolacija na primeru vsebnosti cinka v okolici opuščenih topilnic cinka v Celju (slika 1). Interpolacija je bila narejena na podlagi večslojnega perceptrona, kateri je bil naučen po metodi vzvratnega učenja na 118 vzorčnih točkah podstrešnega prahu. Narejena je tudi primerjava z običajno uporabljeno metodo za izris kart – krigiranjem (slika 2). Pri karti, narejeni z nevronsko mrežo vidimo, da ima nadmorska višina velik vpliv na napovedano vsebnost cinka, kar ustreza z opazovanji, saj vsebnosti cinka na višini, ki je višja, kot je običajna višina temperaturnega obrata, znatno padejo. S krigiranjem tega pojava ni moč zajeti. Drug viden pojav, ki se je odrazil pri izrisu karte z uporabo nevronskih mrež, in ga je zelo težko pravilno simulirati s krigiranjem, pa je vpliv prevladujočih zahodnih vetrov. Slaba stran uporabe nevronskih mrež pa je pojav šuma pri končni karti, ki najverjetneje ni povezan s pojavom.

Drugi primer je primer razvrščanja, kjer smo z nevronsko mrežo razvrščali vzorčne materiale Celjskega območja (slika 1), in sicer tla, aluvialni sediment ter podstrešni prah, na podlagi vsebnosti prvin v njih. Mrežo (slika 3) smo naučili na podlagi 121 analiz podstrešnih prahov, 129 analiz tal ter 95 analiz aluvialnih sedimentov. Dodatno smo v učno matriko vključili še 180 primerov naključnih vrednosti. Tip materiala so predstavljali trije izhodni nevroni, rezultate pa smo kontrolirali s 37 primeri, ki niso bili vključeni v učno matriko. Nevronska mreža je pravilno razporedila vseh 37 primerov. Sposobna je bila zaznati tudi primere z naključnimi vrednostmi.

Tretji primer je primer določevanje geokemičnih združb z uporabo samoorganizacijskih mrež. Pri tem smo uporabili podatke iz geokemičnega vzorčenja tal in podstrešnega prahu na območju Mežice (slika 1). Poizkušali smo ugotoviti, kateri elementi v okolju so posledica delovanja rudnika in topilnice, kateri elementi so posledica delovanja železarne, in katere so naravne združbe elementov (vpliv karbonatnih, magmatskih in metamorfnih kamenin). Geokemične združbe smo prebrali iz popolnoma naučene matrike izhodnih nevronov (slika 4), pri čemer imajo bližnji elementi bolj podobno razporeditev, kot elementi, ki so dlje narazen (slika 5). Na takšen način smo določili 5 geokemičnih združb, in sicer združbo onesnaženja, združbo karbonatnih elementov, združbo elementov težkih mineralov, združbo elementov kislih magmatskih in metamorfnih kamnin ter združbo kationov, ki tvorijo topne minerale.

Zadnji primer pa je izris geokemičnih kart na primeru Kavadarcev v Makedoniji (slika 1). Primer je sicer podoben prvemu primeru, kljub vsemu pa se zelo razlikuje. Ugotovljena geokemična združba As-Sb-Tl je zastopana v aluvialnih sedimentih reke Crne in Vardar, kar je posledica erozije matičnih kamnin in delovanja Sb-As rudnika Alšar južno od obravnavanega območja. Uporaba krigiranja je povzročila povsem neprimerno interpretacijo, saj so opazni učinki ekstremnih vrednosti, ki se odrazijo v t.i. »bikovih očeh«. Povišane vsebnosti elementov zaradi naravnih procesov ni vidna. Pri uporabi nevronskih mrež smo upoštevali položaj in nadmorsko višino, geologijo, rabo tal ter naklon pobočja. Rezultati, dobljeni z nevronskimi mrežami so podali bistveno boljšo generalizacijo, kot krigiranje (slike 6, 7 in 8).

Апстракт

Во трудот авторите се обидуваат на кратко да претстават 4 примери на примена на невронските мрежи при обработка на геохемиски податоци. Првиот пример е прилично едноставна просторна интерполација на примерот на содржината на цинк во околината на напуштената топилница на цинк во Цеље (слика 1). Интерполацијата е направена врз основа на повеќеслоен перцептрон, кој е научен по методот на повратно учење на 118 точки од кои се земени примероци од поткровна прашина. Исто така, применет е и вообичаените метод за припрема на карти –

со кригирање (слика 2). На картите подготвени со невронските мрежи, може да се види дека надморска височина има големо влијание врз предвидените содржини на цинк, што одговара со најденото, дека содржината на цинкот на повисоките новоа, каде вредноста на температурата е обратна, значително се намалуваат. Со кригирањето овој феномен не може да бидат забележан. Друга видлива појава која има влијание при изработката на картите со користење на невронските мрежи, и е многу тешко тоа правилно да се симулира со кригирање, е влијанието на доминантните западни ветрови. Слаба страна на невронските мрежи е појавата на шум на финалните карти, која најверојатно не е поврзана со некој феномен.

Друг пример е примерот на класификација, каде со невронски мрежи е извршена класификација на материјалите од земените примероци од областа на Цеље (слика 1), како што е почва, алувијален седимент и поткровна прашина, врз основа на содржината на елементите во нив. Мрежата (слика 3) е научена врз основа на анализа на 121 примерок од поткровна прашина, 129 примероци почва и 95 примероци од алувијални седименти. Покрај тоа, во матрицата се вклучени и 180 случајни вредности. Видот на материјалот е претставен со три излезни неврони, резултатите се проверени со 37 примери кои не се вклучени во матрицата. Невронската мрежа правилно ги распредели сите 37 примери. Беше во можност да ги детектира и примерите со случајните вредности.

Третиот пример е пример на утврдување на геохимиски групи со примена на самоорганизирани мрежи. При тоа се употребени податоците од геохимиските примероци на почви и поткровна прашина во околината на Межице (слика 1). Направен е обид да се утврди кои елементи присутни во животната средина се како резултат на работата на рудникот и топилницата, кои елементи се резултат на работата на железарницата, а кои се природни групи на елементи (влијание на карбонатните, магматските и метаморфните карпи). Геохимиските групи се одредени со пополнување на научената матрица на излезните неврони (слика 4), при што сличните елементи имаат посоодветна распределба од елементите кои се поразлични (слика 5). На овој начин се идентификувани пет геохимиски групи, антропогена група, група на карбонатни елементи, група на елементи на тешки минерали, група на елементи на киселите магматски и метаморфни карпи и група катјони кои формираат растворливи минерали.

Последниот пример се однесува на изготвување на геохимиски карти на примерот на Кавадарци во Македонија (слика 1). Овој пример е многу сличен на првиот случај, но сепак, се чини дека значително се разликува. Геохимиската асоцијација As-Sb-Tl е застапена во алувијалните седименти на реките Црна Река и Вардар и е како резултат на ерозија на матичните карпи и работата на Sb-As рудникот Алшар кој се наоѓа јужно од испитуваната област. Примената на кригирасето дава сосема несоодветна интерпретација, бидејќи се забележани ефекти на екстремни вредности, кои се рефлектираат во т.н. “бикови очи”. Зголемените вредности на содржините на елементи како резултат на природните процеси не се видливи. Со примена на невронските мрежи, усогласени се положбата и надморската височина, геологијата, уптребата на земјиштето и аголот на наклонот. Резултатите добиени со невронските мрежи даваат значително многу подобра генерализација отколку оние добиени со кригирање (слики 6, 7 и 8).

1. Introduction

This paper presents a definitive description of a neural network methodology applied to the geochemical data interpretation, and it provides an evaluation of its advantages and disadvantages compared with statistical procedures. The method is based on the principles of biological nervous systems where a basic unit is the neuron containing inputs with weights, activation function and outputs. These models emulate the neurophysical structure and decision making of the human brain mathematically. From a statistical point of view, they are closely related to generalized linear models. Artificial neural networks (ANNs), however, are nonlinear and they use a different estimation procedure (feed forward and back propagation) compared with traditional statistical models (West, Brockett & Golden, 1997). Neural networks can be applied for prediction and classification – fields where statistical methods have been used traditionally. Both the traditional statistical methods and neural networks are looked upon in the literature as competing model-building techniques (Paliwal & Kumar, 2009). The main difficulty of any pattern recognition system is the great amount of fuzzy and incomplete information to be dealt with. Moreover, the classification problem does not allow an exact solution, so statistical and artificial neural network techniques must be used in order to obtain results that offer an optimum degree of reliability (Hervas-Martinez et al., 1993).

Additionally, neural network models do not require some restrictive assumptions about relationships between independent and dependent variables, as statistical methods do normally. Researchers from many scientific disciplines are designing artificial neural networks to solve a variety of problems in pattern recognition, prediction, optimization, associative memory, and control (Du & Swamy, 2006). Consequently, these models have already been applied very successfully in many diverse disciplines, including biology, psychology, statistics, mathematics, business, insurance, and computer science (Razi & Athappilly, 2005). In this study, some interesting applications will be provided where neural networks have been used for geochemical data applications and where they exhibit superior performance in comparison to the classical statistical methods.

The success of the method can be laid down due to the following reasons: a) they can model extremely complex systems and they can be used to model non-linear natural systems (linearity in the sense of mathematical properties of additivity and homogeneity) due to their nature; we always have to make approximations and simplifications if using linear algebra (i. e. most of multivariate statistics) to describe non-linear systems; b) there are no limitations with the dimensionality of the problem; it can be arbitrary, depending on the CPU speed and memory; c) due to well developed learning algorithms, they are easy to apply (HAYKIN, 1999; KOHONEN, 2001). Another important issue of neural networks is their learning process. There are two main paradigms: supervised and unsupervised ones. In both cases its task is to adjust the weights and biases in every neuron and link of the network so that the network is capable of performing a specific task. In the case of a supervised algorithm, this is done in the way that the data input corresponds best to the appropriate outputs. The "back propagation algorithm" is frequently used. It has been applied for functional approximation problems or classification problems if we possess a dataset with known input-output pairs. The second basic learning process is called unsupervised learning, where neurons "compete" with each other and it is commonly called the competitive learning algorithm. This algorithm is mainly used in self organizing maps for clustering problems. In this type of learning, we do not need input-output data pairs in the dataset and the classification is made only according to known inputs (JAIN, MAO & MOHIUDDIN, 1996).

The present paper analyses the methodology used to compare some of these techniques. Our data are related to chemical analysis of different geochemical sampling media in order to determine the existence of possible contaminations nearby the assumed source of pollution such as metal smelter plants, ironworks and mines.

2. Materials, methods and results

2.1. Function approximation

As neural networks are universal approximators (they can approximate any function with arbitrary accuracy if there is enough available data and enough topologic complexity of neural network) they can be used for environmental modeling or for interpolation procedures. This case study will show and compare as an example the results of the interpolation of the Zn concentration around a former Zn smelter plant using neural networks and the digital elevation model with 25 m space resolution.

The town of Celje is the third largest city in Slovenia with 55,000 inhabitants (Fig. 1). The industry is concentrated in the eastern part of the town where chemical and metal industries (Zn smelter plant – Cinkarna) prevail. A further important industrial settlement is Štore in the eastern part of the area (Ironworks Štore). Particular studies concerning soil and atmospheric contamination with trace elements have been published recently (ŠAJN, 2005; ŽIBRET, 2008; ŽIBRET & ŠAJN, 2008a; 2008b; ŽIBRET & ROKAVEC, 2010). Ironworking and smelting activities as well as the urbanization of Celje left significant impacts on the environment. The maximum measured values of Cd (240 ppm) and Pb (3200 ppm) have been found in attic dust. The average content of Cd (7.5 ppm) in topsoil is even 15 times higher than the general Slovenian average.

We look at the Zn concentration in attic dust in Celje as a function of the spatial position of the sample (x and y coordinates) and elevation. This is a very simple and naive approximation but it will suit for the purpose of this demonstration. The "spatial position" has been expressed as a relative position of the sample according to the position of the mayor pollutant in Celje – the past Cinkarna Celje smelting furnaces. This can be expressed by the following equation:

$$C(Zn) = \tilde{A}(dx; dy; z)$$

where $dx = x_{\mathrm{sample}}$, $dy = y_{\mathrm{sample}}$, z = the elevation and $C(Zn)$ = the Zn concentration in attic dust.

Two data matrices have been prepared. The learning data matrix (Fig. 2a) contains the values of dx, dy, z and $C\,(Zn)$ from 118 sampling points in the Celje area (• IBRET, 2002; ŠAJN, 2005; ŽIBRET & ŠAJN, 2008b). All variables have been rescaled linearly to the interval [0; 1]. The second matrix, a recall matrix, is a modified digital elevation model for the Celje area (Fig. 2c) where the variables dx, dy and z have been introduced with 25 m of resolution in both, x and y, directions with the same rescaling as in the learning matrix. The topology of the used neural network (Fig. 2) is: three input neurons, representing dx, dy and z, a first layer with 30 hidden neurons, a second layer with 10 hidden neurons and output, representing $C(Zn)$. The network is fully connected and sigmoidal

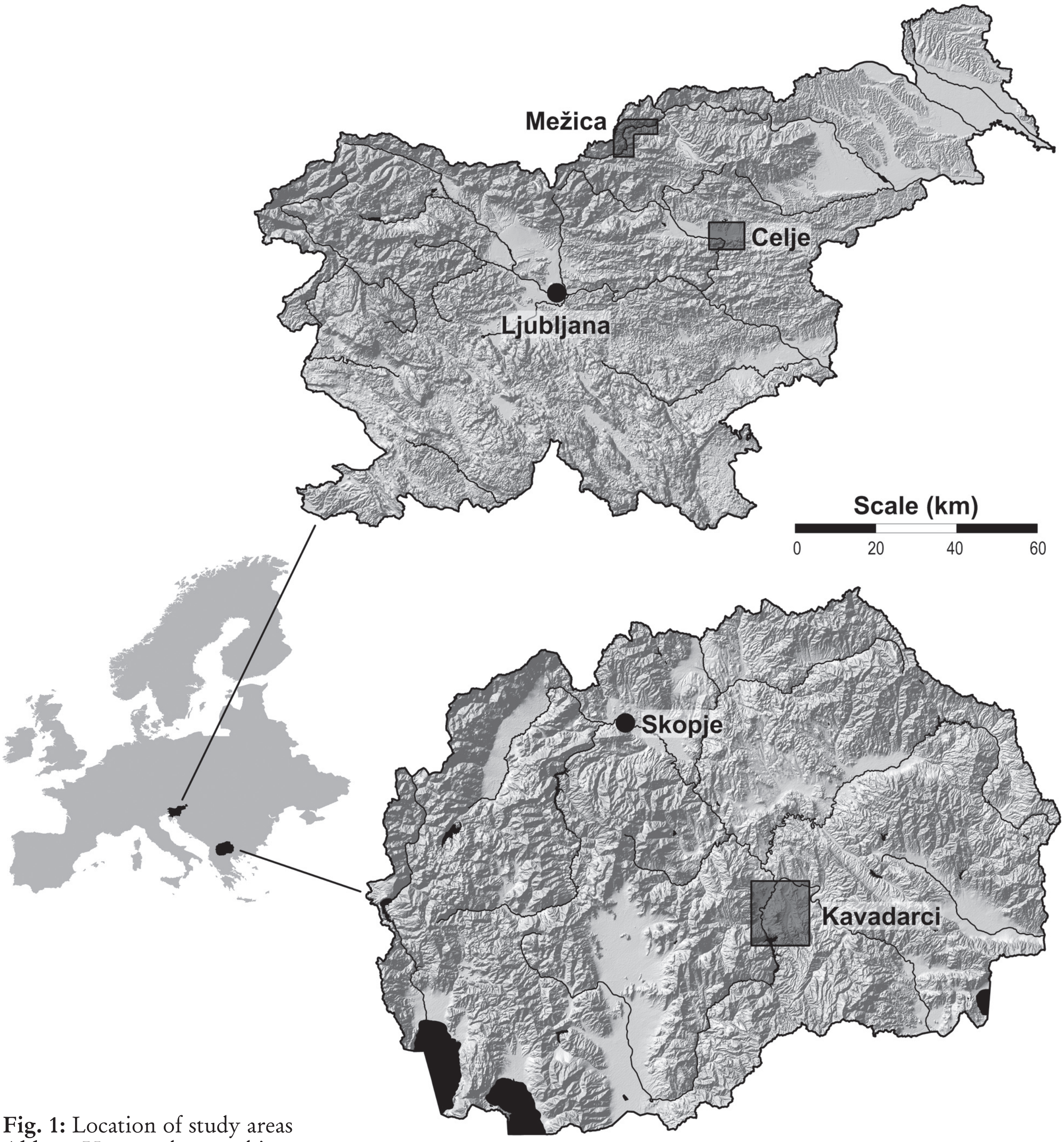

Fig. 1: Location of study areas
Abb. 1: Untersuchungsgebiete
Slika 1: Lokacije obravnavanih območij
Слика 1: Локации на испитуваните области

activation functions have been used. The learning phase has been realized on the basis of a supervised back propagation learning algorithm with gradual decreasing of learning speed. The recall phase has been made on the fully trained neural network. The obtained results are shown in Fig. 2d.

Comparing the interpolation results obtained from neural network and kriging method we see that the neural network interpolation yields better results, despite it has been made only on the basis of three spatial variables (*dx*, *dy* and *z*). Three results are visible which are better compared with the kriging method.

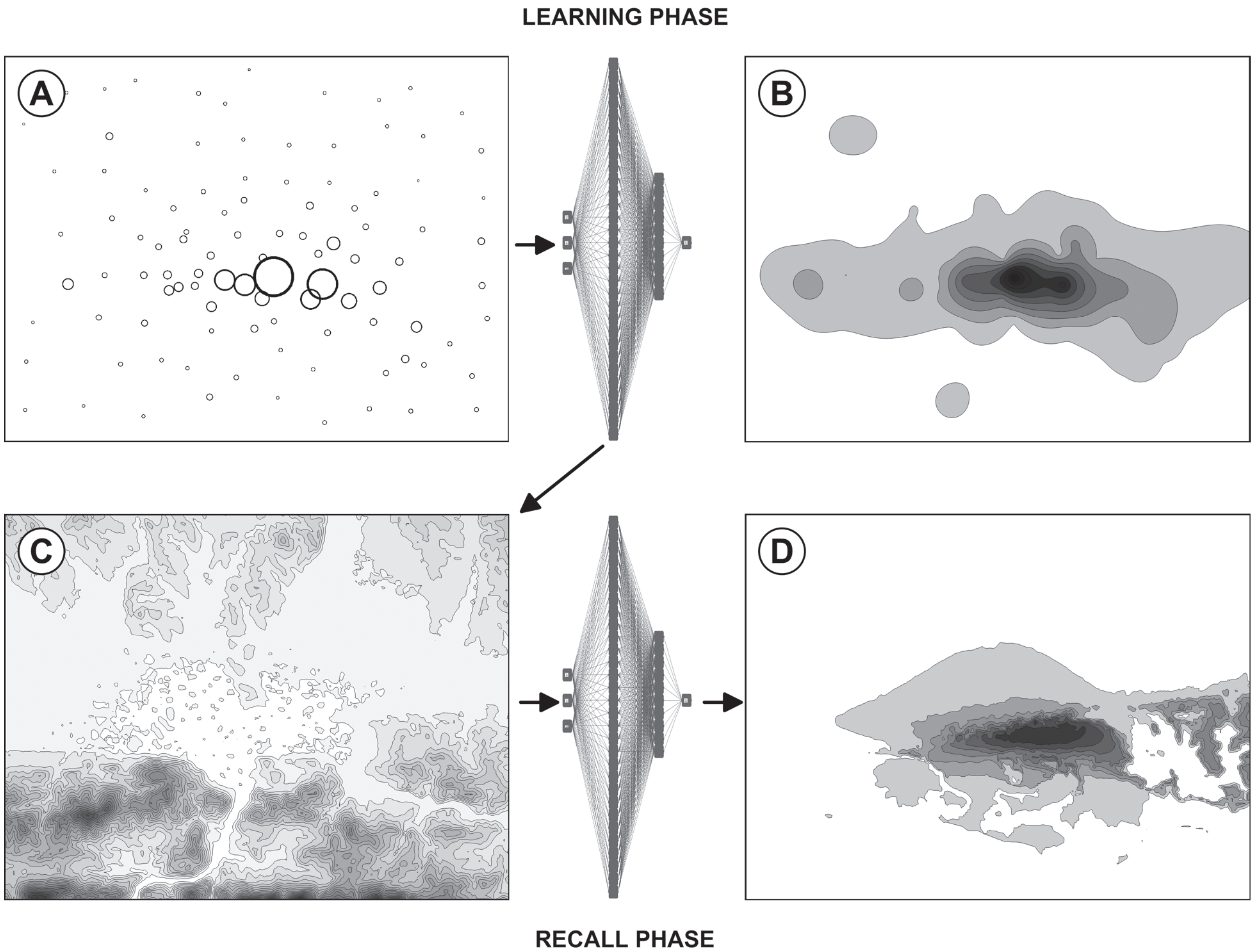

Fig. 2: Procedure of function approximation – Zn in Celje attic dust case study
A – visualization of learning data (Zn concentration); B – spatial distribution of Zn concentration, kriging interpolation method used (25 m spatial resolution); C – recall matrix (DEM of the Celje basin, 25 m spatial resolution); D – visualization of the neural network interpolation results.

Abb. 2: Funktionsanpassung – Zn in der Fallstudie Hüttenstaub Celje
A: Visualisierung der Lernphase (Zn-Konzentration). B:Räumliche Verteilung von Zn nach Kriging (räumliche Auflösung 25 m). C: Examensmatrix (DEM des Celje-Bassins, räumliche Auflösung 25 m). D: Visualisierung der Interpolationsergebnisse nach neuralem Netzwerk.

Slika 2: Proces funkcijske aproksimacije (vsebnost Zn v podstrešnem prahu na območju Celja)
A – vizualizacija učnih podatkov (vsebnost Zn); B – prostorska porazdelitev Zn, določena s krigiranjem (25 m ločljivost); C – vizualizacija vhodnih podatkov za model (DMR celjskega območja, 25 m ločljivost); D – vizualizacija interpoliranih vsebnosti izračunanih z nevronskimi mrežami

Слика 2: Постапка за функционална апроксимација, Zn во поткровна прашина од студијата за Целје
A – визуализација на податоците за учење (концентрацијата на Zn); B – просторна дистрибуција на концентрациите на Zn, применета е kriging метода на интерполација (25 m просторна резолуција); C – повикувачка матрица (модифициран DEM на басенот на Целје, 25 m просторна резолуција); D – визуализација на интерполацијата на резултатите со невронската мрежа

(1) The topographical elevation has a significant influence on the interpolated Zn concentration. This is very reasonable, as the temperature inversion is a very common phenomenon for this area. It prevents a spreading of the past pollution above the fog. (2) The dominant wind direction from W to E has been correctly suggested by neural network. (3) The neural network predicts certain elevation with increased Zn

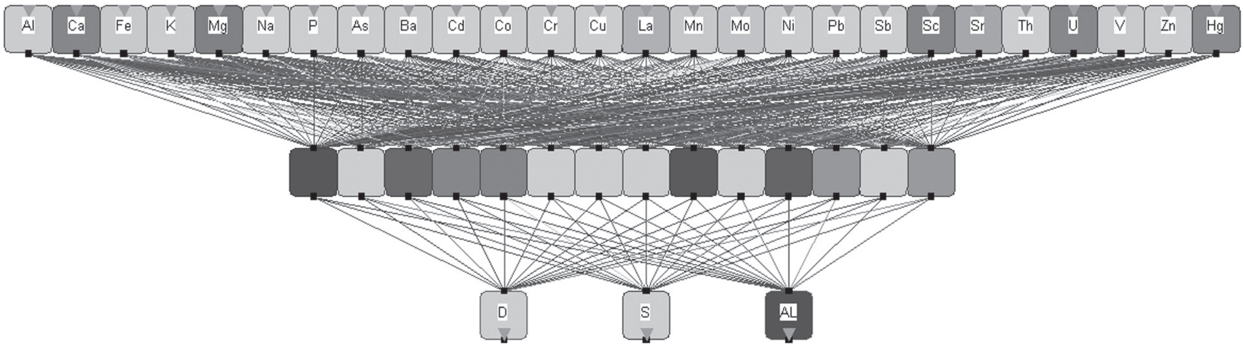

Fig. 3: The topology of neural network, used for classification of geochemical sampling materials in Celje area (attic dust case)
Abb. 3: Topologie des neuronalen Netzwerks für die Klassifizierung geochemischer Proben im Gebiet Celje (Fallstudie Hüttenstaub)
Slika 3: Topologija nevronske mreže, uporabljena za klasifikacijo geokemičnih vzorčnih sredstev na celjskem območju. Prikazan je primer za podstreљni prah
Слика 3: Топологија на невронската мрежа применета за класификација на геохемиските примероци од испитуваните материјали во регионот на Цеље. Прикажан е случајот со поткровната прашина

concentrations, well corresponding to the height of the winter temperature inversion, particularly visible hillsides S from Celje, which are facing towards the location of past Zn smelters. However, two predictions are visible that are not easily to explain –low concentrations in the Voglajna River valley (SE part) not corresponding to the actual data and a noise in the mostly polluted areas which can be caused by noise in the original learning dataset.

2.2. Classification

The classification problem will be addressed by the classification of different geological materials on the basis of their chemical composition. Attic dust, topsoil and alluvial sediment from different contaminated and unpolluted areas have been chosen as examples. The learning matrix was built up by chemical analyses of 26 different elements and of three index parameters which indicates the type of the material. If there is an analysis of attic dust (121 samples), the value of the first index is 1 and the other two indexes are 0. If there is the analysis of topsoil (129 samples), the second index value was set to 1 and first and third indexes are set to 0. The analogy is also given for the alluvial sediments (95 samples). Additionally 180 patterns with random values have been inserted to improve the performance of neural network. All three index parameters were set to 0 for the random value cases. All variables have been normalized to the interval [0; 1].

The topology of neural network is as follows (Fig. 3): there are 26 input neurons, representing the concentration of 26 chemical elements. The hidden layer contains 14 neurons with sigmoidal activation function. The output layer is composed of 3 neurons, representing attic dust, soil and alluvial sediments. Dark tones represent a high neuron activation (positive weight), bright tones – a low activation (links weighting close to 0 and negative weight). The neural network has been trained with approximately 200 learning cycles (back propagation) with a high learning rate. The performance of the neural network has been evaluated based on 37 patterns which were not included in the learning set. The evaluation set contains 9 analyses of attic dust, 7 analyses of topsoil, 8 analysis of alluvial sediment and 3 patterns with random values. The result shows that the completely trained neural network is capable to determine the type of material on the basis of the concentration of 26 elements with 100% accuracy. Such performance could be very useful for the automation of many industrial applications, where a production process depends on the composition of raw materials.

2.3. Clustering

The clustering of geochemical data using neural networks is realized by two stages. First stage is the reduction of the dimension using self organizing maps (SOM) which reduce a dimension of an input data set usually to two dimensions. At this process, a topolo-

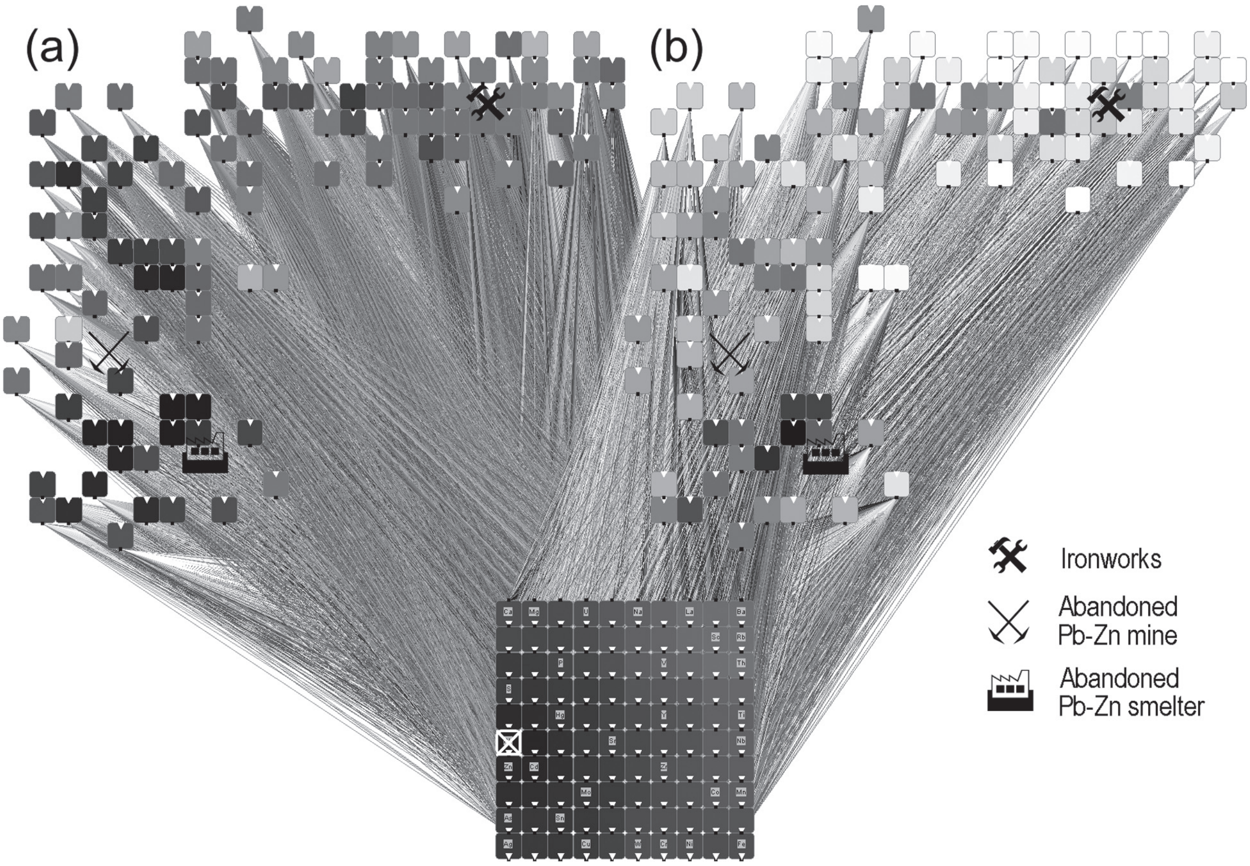

Fig. 4: Neural network (self organizing map), used for clustering case study – Mežica area
Abb. 4: Neuronales Netzwerk (selbstorganisierende Karte) für die Cluster-Fallstudie – Gebiet Mežica
Slika 4: Samoorganizacijska mreža, ki je bila uporabljena za potrebe združevanja (clustering) – Mežiško območje
Слика 4: Невронска мрежа (самоорганизирачка мапа), употребена за групирање на предметната студија – регионот на Межице

gical property of the input space is preserved. An unsupervised learning algorithm is used to train such neural networks. A SOM is composed of two layers – an input layer and an output lattice of neurons. Output neurons "compete for the domination" on the basis of neighborhood function and "winner takes it all" concept. The final result is that patterns with similar properties generate similar neural activation patterns in the output lattice. Second stage is the clustering of output neurons (and consecutively clustering of elements) by the k-means clustering algorithm.

This case study presents the clustering of elements in the Mežica area (Fig. 1) where four groups of elements are expected: a carbonate group, an ironworking group, the Pb–Zn mining and smelting group and a group of elements representing metamorphic rocks (ŠAJN, 2006). The input dataset consists of chemical analyses of soils and attic dust sampled at 106 locations (212 samples). 37 elements have been analyzed. All variables have been linearly remapped to the interval [–1, +1]. The topology of the neural network is presented in Fig. 4. Input neurons (upper boxes) are directly linked to the output lattice (bottom-down 10 * 10 squares of output neurons). The Pb array of the data is presented as an example with the indication of best matching unit inside output grid. A bright tone of neurons indicates low activation and dark colors indicate high neuron activation. The locations of input neurons roughly correspond to the position of the sample represented by particular input neuron but does influence to the performance of SOM. Group (a) corresponds to the attic dust samples and group (b) to the soil samples. The first group approximately shows the position of ironworks, the second group the abandoned Pb–Zn–Mo mine and the third group the abandoned Pb–Zn smelter.

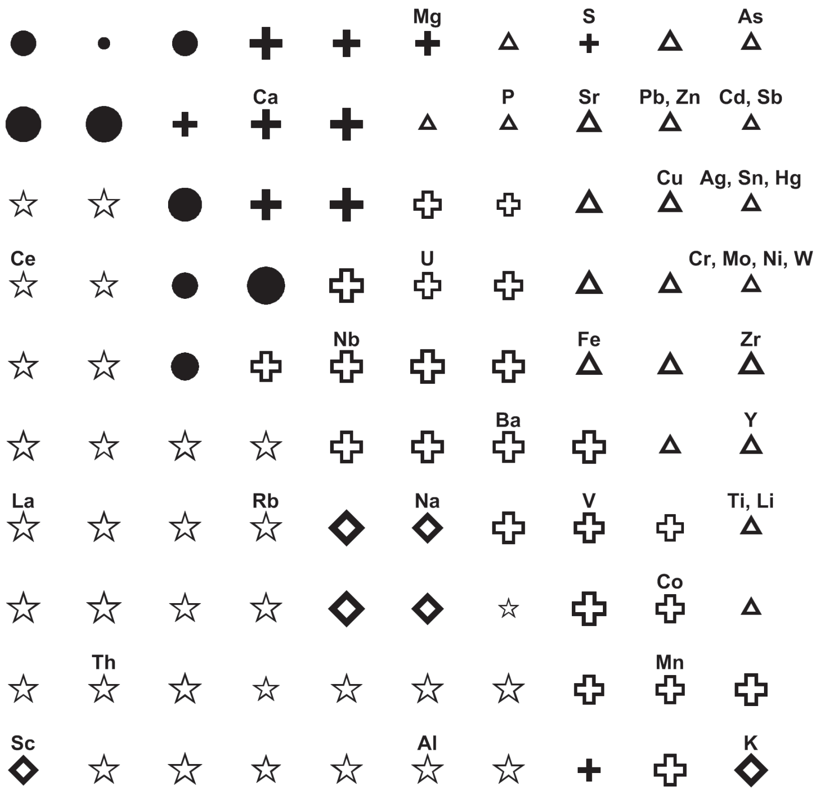

Fig. 5: Results of determination of geochemical associations using SOM and k-means clustering – Mežica area
Abb. 5: Geochemische Assoziationen nach SOM und Clusterung nach k-Mittelwerten – Gebiet Mežica
Slika 5: Rezultati določanja geokemičnih združb z uporabo samoorganizacijskih mrež in k-means algoritma – Mežiško območje
Слика 5: Резултати од определувањето на геохемиските асоцијации со примена на SOM и групирање со k-значење – регион на Межице

212 input neurons represent the result of chemical analyses of all 212 samples. The learning set contains 37 patterns (each pattern for every element). The output lattice consists of 100 output neurons. The Euclidean distance measure was used to calculate neural activations. The learning process has been performed on the basis of a competitive learning algorithm. The starting adaptation radius was set to 10 neurons, the final adaptation radius to 1 neuron. The starting learning rate was set to 0.7%, the final learning rate to 0.1% of initial one. There were 10,000 repetitions per learning cycle. The Mexican hat function (second derivate of Gaussian function) was used as neighborhood function.

To draw boundaries between groups of elements, output neuron activations further have been processed with the k-means clustering algorithm. Five clusters have been extracted. The results are shown in Fig. 5 where different symbols indicate different group of elements in the output lattice of SOM. Positions of best matching units for every element are indicated above the symbol. The size of the symbol corresponds to the distance to the k-means cluster centre (larger is the symbol; shorter is the distance to the group centre).

A first group, representing anthropogenic pollution, consists of As, Sr, Pb, Zn, Cd, Sb, Cu, Ag, Sn, Hg, Cr, Mo, Ni, W, Fe, Zr, Y, Ti and Li. The second group of elements links Ca, Mg and S and represents the carbonatic part of the research area (the smelter plant was located at the area with mainly carbonatic geological settings). A third group of elements contains U, Nb, Ba, V, Co, and Mn, and it represents elements with intermediate ionic potential thus forming hardly soluble minerals. The fourth group of elements is built up by the easily soluble elements Sc, Na and K and thus it has a lower abundance in attic dust. The fifth group of elements consists of Ce, La, Rb, Th and Al and represents elements of metamorphic origin.

The obtained results confirm the expectations. Anthropogenically induced elements to the environment are grouped together and elements representing natural associations are separated in four different groups. Among anthropogenically induced elements, two subgroups can be recognized. The first subgroup consists of Pb, Zn, Cd, Sb and As, which is a consequence of the Pb-Zn mining and smelting processes, and the second group includes Cr, Mo, Ni and W, which is a consequence of former ironworks.

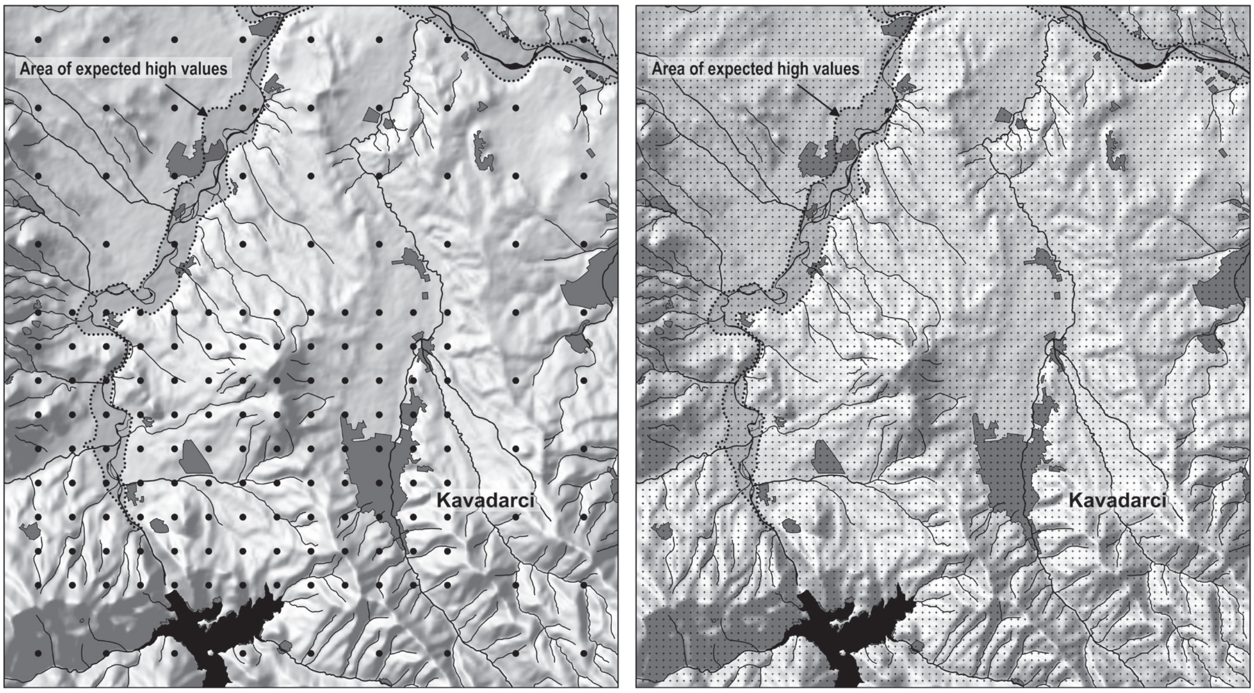

Fig. 6: Soil sampling grid of Kavadarci area (left), distribution sites for the modeling of spatial distribution of geochemical group of As–Sb–Tl (right)
Abb. 6: Probenahmeschema des Gebietes Kavadarci (links), Anordnung der Teilflächen für die Modellierung der räumlichen Verteilung der Gruppe As–Sb–Tl (rechts)
Slika 6: Vzorčna mreža tal na območju Kavadarcev (levo) in razporeditev točk za modeliranje prostorske razporeditve geokemične skupine As–Sb–Tl (desno)
Слика 6: Мрежа за земање проби во областа на Кавадарци (лево) и распределба на точките за моделирање на просторната распределба на геохемиската група As–Sb–Tl (десно)

2.4. Multilayer perceptrons

The advantage of the application of ANN vs. kriging method is explained in the test region Kavadarci, Republic of Macedonia. This case study is similar to the Celje case study. As in Celje, main factors for spreading the pollution around smelter are the topography and the dominant wind direction (attic dust is a very stable geochemical medium). But the situation in Kavadarci is much more complicated and the input data includes more noise than in Celje because in this case study, soil samples have been collected.

The town Kavadarci is located in the Tikveš valley, about 100 km south from the capital Skopje (Fig. 1). The region is known for its ferronickel industrial activity in the nearest past but it is also the main vine producing region in Macedonia. There were several investigations of the atmosphere, subsoil, vegetables and fruits produced in this region, mainly concerning a contamination by trace elements (BOEV, ŽIVANOVIĆ & LIPITKOVA, 2005; BARANDOVSKI, CEKOVA, FRONTASYEVA et al., 2006; BARANDOVSKI, CEKOVA, FRONTASYEVA et al., 2008; STAFILOV, ŠAJN, BOEV et al., 2008; STAFILOV, ŠAJN, BOEV et al., 2010). Very high concentrations of As, Sb and Tl have been detected in the Holocene alluvium of the Crna Reka river. The enrichment of polluting elements in the Holocene alluvium of the Crna Reka river is suppose to be as a consequence of natural erosion from the Alšar mine deposits (As–Sb–Tl) at the Kožuf Mountain, but also from mining activities in the past.

In this case study, the results of soil contamination have been used to train the neural network. The complete investigated region (360 km^2) is covered by a sampling grid of 2 × 2 km^2; in the urban zone and around the ferronickel smelter plant (117 km^2) the sampling grid is denser: 1 × 1 km^2. The regular soil sampling grid of the Kavadarci region is provided in Fig. 6 (left side). The distribution of points for the recall grid used to model the spatial distribution of geochemical associations As–Sb–Tl is also shown in Fig. 6 (right side). Each sampling site is defined by

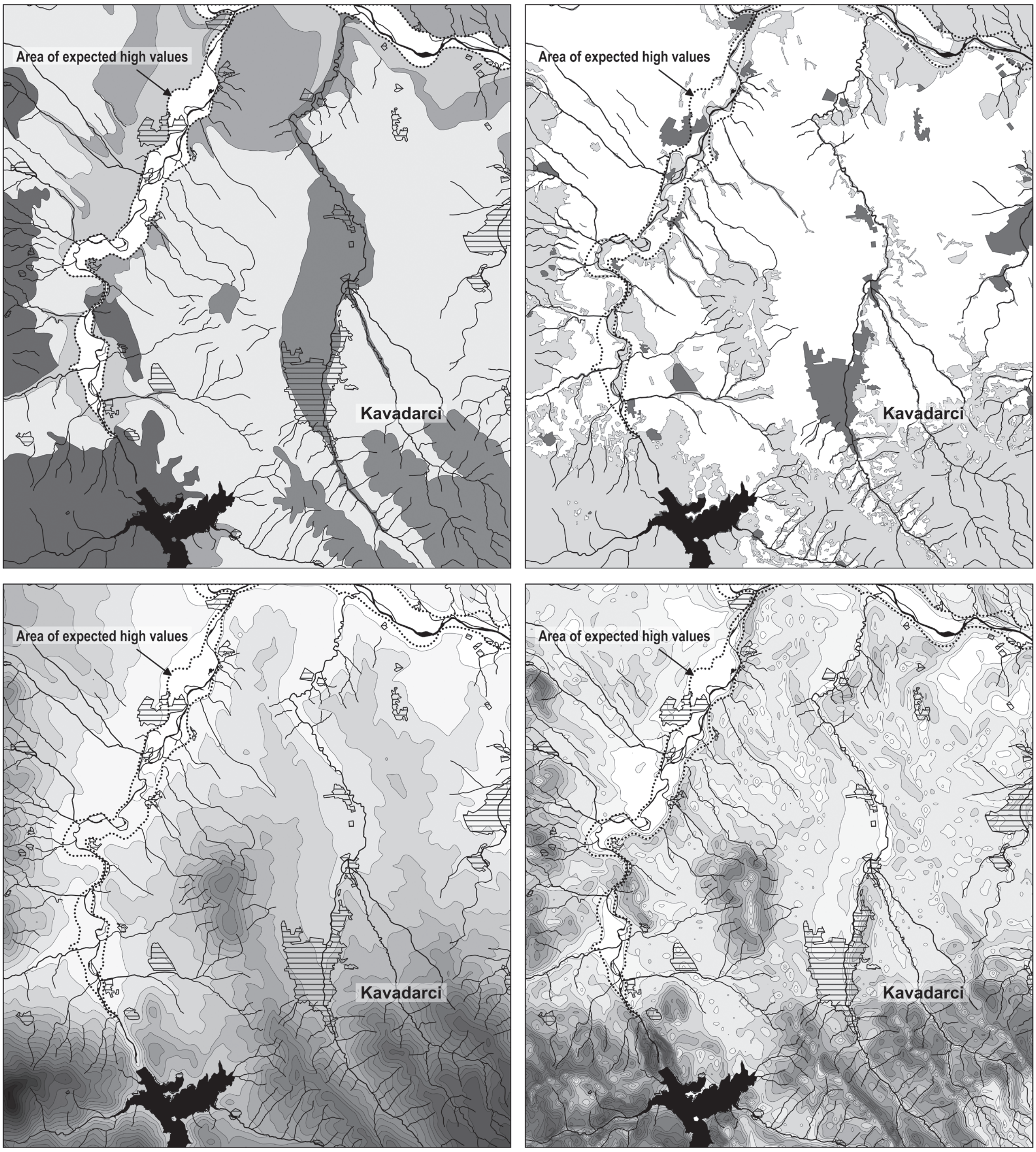

Fig. 7: Spatial data (lithology, land use, position and elevation and slope angle) used for the construction of the learning and recall data matrix for the modeling As–Sb–Tl
Abb. 7: Raumdaten (Lithologie, Flächennutzung, Position, Höhe und Hangneigung) für die Erzeugung der Lern- und Examensmatrix zur Modellierung von As–Sb–Tl
Slika 7: Prostorni podatki (litologija, raba tal, nadmorska višina in naklon pobočja), uporabljeni za učenje in modeliranje prostorske porazdelitve geokemične skupine As–Sb–Tl
Слика 7: Просторни податоци (употреба на земјиштето, надморска висина и агол на наклонот) применети за учење и моделирање на просторната распределба на геохемиската група As–Sb–Tl

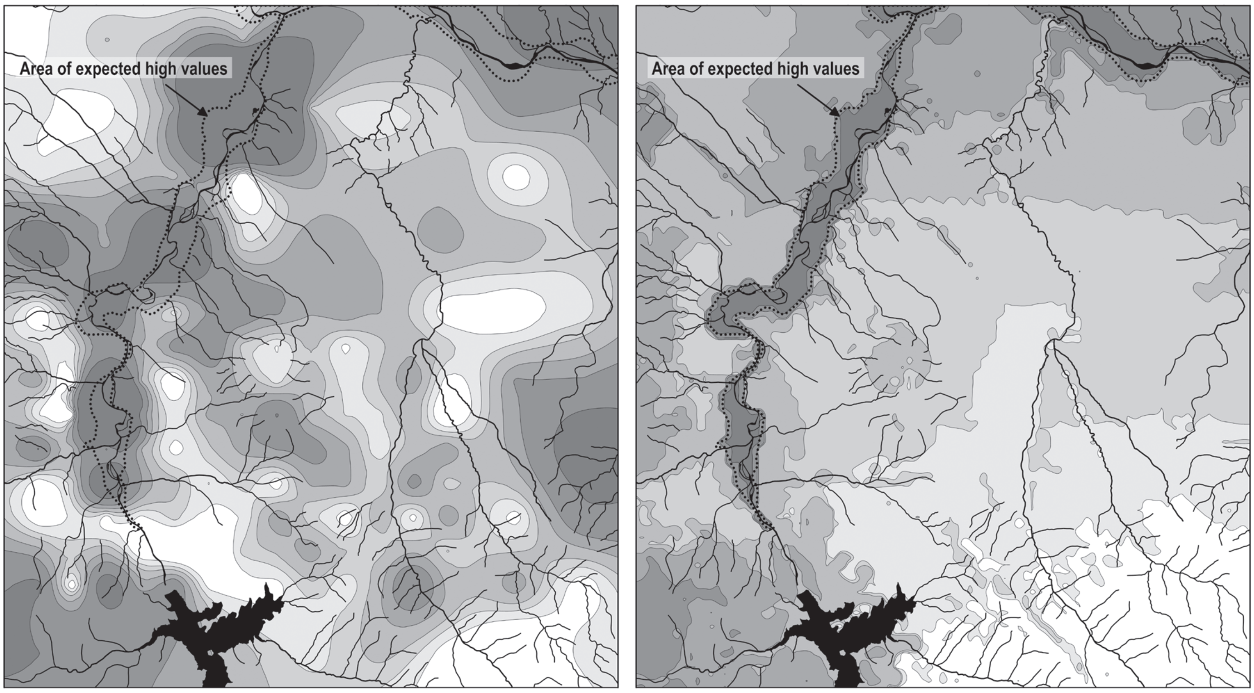

Fig. 8: The comparison of the geochemical maps, made by Kriging (left) and ANN-multilayer perceptrons (right)
Abb. 8: Vergleich der geochemischen Karten nach Kriging (links) und ANN in Multilayer-Projektion (rechts)
Slika 8: Primerjava geokemičnih kart, izdelanih na podlagi krigiranja (levo) in uporabe nevronskih mrež – veиslojni perceptron (desno)
Слика 8: Споредба на геохемиските карти, припремени на Kriging подлоги (лево) и со примена на невронски мрежи – повеќеслоен перцептрон (десно)

some parameters such as longitude, altitude and concentration of chemical elements. All together, 344 soil samples were collected in 172 locations. At each sampling point, soil samples were collected at two depths: topsoil (0 to 5 cm below surface) and bottom (20 to 30 cm) soil. The determination of 36 elements (Ag, Al, As, Au, B, Ba, Bi, Ca, Cd, Co, Cr, Cu, Fe, Ga, Hg, K, La, Mn, Na, Mg, Mo, Ni, P, Pb, S, Sb, Sc, Se, Sr, Th, Tl, Ti, U, V, W, and Zn) was performed by inductively coupled plasma–mass spectrometry (ICP-MS).

Using kriging as interpolation method (DAVIS, 1986; WEBSTER & OLIVER, 2007), the distribution of particular chemical elements or group of elements can be represented in 2D maps. In such maps there are some errors called bull's-eye effect causing an elongated division in the isotropic space. Possibly this effect can be solved only by use of a denser sampling grid. We are trying to avoid those problems and improve the maps applying ANN (ŽIBRET & ŠAJN, 2010) and exploiting their compatibility of processing different types of independent variables (normally distributed, simplex, attributive data, such as land use or geology). Each input besides the standard position parameters (x and y) is described by some new parameters such as elevation, slope angle, geological composition, land use etc. (Fig. 7). These parameters have been determined for all of the 172 sampling locations which have been used for the learning dataset, as well as for all recall points (200 m × 200 m grid).

In the study, a multilayer perceptron (MLP) has been applied to approximate the function, similar than in Celje. However, the Kavadarci study differs in the way that the investigated concentrations are connected with a specific geological unit (alluvial deposits of the rivers Crna and Vardar). Thus, the modeling space is far from being homogeneous. Many other different variables have been used in this case all contributing their influence to the final product – a generalized geochemical map. All elements were modeled using this way but only the As–Sb–Tl association is presented in this contribution. The applied method of ANN in the test region shows the distribution much better than the kriging method (Fig. 8). Areas of high concentrations are connected only with the areas where such concentrations can be expected – these are alluvial deposits and hills in the SW part of the research area. The bull's-eye effect is removed completely and, moreover, the map created by ANN contains much less noise than the map prepared by kriging. Some geochemists or mathematicians might argue that the ANN map is "too generalized", but the ANN map is certainly much better than the kriging map looking from the geological point of view.

3. Conclusions

Four case studies illustrate some of the possibilities to interpret geochemical and chemical data using neural networks. Despite this paper deals only with the potential of neural network technology there are also some limitations (local minima, wrong learning methodology, wrong generalization etc.) which are not discussed in details here. But nevertheless neural networks with their endless capacities and customization options offer a solution for very broad types of problems to interpret geochemical, geological, environmental and related data.

References

BOEV, B., ŽIVANOVIČ, J., LIPITKOVA, S. (2005): Selenium and other trace elements in the soil of the Tikveš region. - In: B. BOEV & T. SERAFIMOVSKI [eds.]: Proceedings on the 3rd International Workshop on the Anthropogenic Effects on the Human Environment in the Tertially Basins in the Mediterranean.- 23–35, Štip.

BARANDOVSKI, L., CEKOVA, M., FRONTASYEVA, M. V., PAVLOV, S. S., STAFILOV, T., STEINNES, E., URUMOV, V. (2006): Air Pollution Studies in Macedonia Using the Moss Biomonitoring Technique. - NAA, AAS and GIS Technology, JINR Preprint, E18: 160 p., Dubna.

BARANDOVSKI, L., CEKOVA, M., FRONTASYEVA, M. V., PAVLOV, S. S., STAFILOV, T., STEINNES, E., URUMOV, V. (2008): Atmospheric deposition of trace element pollutants in Macedonia studied by the moss biomonitoring technique. - Environmental Monitoring and Assessment, **138**: 107–111, Dordrecht etc.

DAVIS, J. C. (1986): Statistic and data analysis in geology. - John Wiley & Sons: 646 p., New York.

DU, K. L., SWAMY, M. N. S. (2006): Neural Networks in a Softcomputing Framework. - Springer Verlag: 566 p., London.

HAYKIN, S. (1999): Neural Networks. A comprehensive foundation. - 2nd edition, Pearson: 842 p., Singapore.

HERVAS-MARTINEZ, C., ROMERO-SOTO, E. J., GARCIA-PEDRAJAS, N., MEDINA-CARNICER, R. (1993): Comparison between artificial neural networks and classical statistical methods in pattern recognition. - Lecture Notes in Computing Science, **682**: 351–360, Berlin etc.

JAIN, A. K., MAO, J., MOHIUDDIN, K. M. (1996): Artificial Neural Networks: A Tutorial. - Computer, **29** (3): 31–44, IEEE Computer Soc. Publ., Los Alamitos, CA (US).

KOHENEN, T. (2001): Self-Organizing Map. - 3rd ed., Springer Verlag: 501 p., Berlin etc.

PALIWAL, M., KUMAR, U. A. (2009): Neural networks and statistical techniques: A review of applications. - Expert System with Applications, **36**: 2–17, Amsterdam etc.

RAZI, A. M., ATHAPPILLY, K. (2005): A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. - Expert System with Applications, **29**: 65–74, Amsterdam etc.

ŠAJN, R. (2005): Using attic dust and soil for the separation of anthropogenic and geogenic elemental distributions in an old metallurgic area (Celje, Slovenia). - Geochemistry: Exploration, Environment, Analysis, **5**: 59–67, London.

ŠAJN, R. (2006): Factor analysis of soil and attic-dust to separate mining and metallurgy influence, Meza Valley, Slovenia. - Mathematical Geology, **38**: 735–747, New York etc.

STAFILOV, T., ŠAJN, R., BOEV, B., CVETKOVIČ, J., MUKAETOV, D., ANDREEVSKI, M. (2008): Geochemical atlas of Kavadarci and the environment. - Sts. Cyril and Methodius University, Faculty of Natural Science: 109 p., Skopje.

STAFILOV, T., ŠAJN, R., BOEV, B., CVETKOVIČ, J., MUKAETOV, D., ANDREEVSKI, M., LEPITKOVA, S. (2010): Distribution of some elements in surface soil over the Kavadarci region, Republic of Macedonia. - Environmental Earth Sciences, **61** (7): 1515–1530, Berlin etc.

WEST, P. M., BROCKETT, P. L., GOLDEN, L. L. (1997): A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice. - Marketing Science, **16** (4): 370–391, Baltimore etc.

WEBSTER, R., OLIVER, M. A. (2007): Geostatistics for environmental scientists. - John Wiley and Sons: 315 p., New York.

ŽIBRET, G. (2002): Geokemiine lastnosti tal in podstrešnega prahu na obmoiju Celja [Geochemical properties of soil and attic dust in Celje area]. - BsCh thesis, University of Ljubljana: 76 p., Ljubljana.

ŽIBRET, G. (2008): Determination of historical emission of heavy metals into the atmosphere: Celje case study. - Environ. - Geology, **56**: 189–196, Berlin etc.

ŽIBRET, G., ŠAJN, R. (2008a): Impacts of the mining and smelting activities to the environment - Slovenian case studies. - In: SÁNCHEZ, M.L. [eds.] Causes and effects of heavy metal pollution. - Nova Science Publishers: 1–80, New York.

ŽIBRET, G., ŠAJN, R. (2008b): Modelling of atmospheric dispersion of heavy metals in the Celje area, Slovenia. - Journ. geochem. explor., **97**: 29–41, Amsterdam etc.

ŽIBRET, G., ŠAJN, R. (2010): Hunting for geochemical associations of elements: factor analysis and self-organising maps. - Math. Geology, **42** (6): 681–703, New York etc .

ŽIBRET, G., ROKAVEC, D. (2010): Household dust and street sediment as an indicator of recent heavy metals in atmospheric emissions: a case study on a previously heavily contaminated area. - Environmental Earth Sciences, **61** (3): 443–453, Berlin etc.

Manuscript received: January 03, 2012
Manuscript accepted: March 30, 2012

Addresses of authors
Gorazd Žibret, Geological survey of Slovenia, Ljubljana, Slovenia
gorazd.zibret@geo-zs.si
Robert Šajn, Geological survey of Slovenia, Ljubljana, Slovenia
robert.sajn@geo-zs.si
Jasminka Alijagić, Geological survey of Slovenia, Ljubljana, Slovenia
jasminka.alijagic@geo-zs.si
Trajče Stafilov, Institute of Chemistry, Faculty of Science, Sts. Cyril and Methodius University, Skopje, Macedonia
trajcest@iunona.pmf.ukim.edu.mk