Scientific paper

X-ray Powder Diffraction and Supervised Self-Ogranizing Maps as Tools for Forensic Classification of Soils

Hirijete Idrizi,[1,2] Mile Markoski,[3] Metodija Najdoski[1] and Igor Kuzmanovski[1,*]

[1] *Ss Cyril and Methodius University, Institute of Chemistry, Faculty of Natural Sciences and Mathematics, Str. Arhimedova 5, Skopje 1000, Republic of North Macedonia*

[2] *State University of Tetovo, Faculty of Natural Sciences and Mathematics, Bul. Ilinden bb, Tetovo 1200, Republic of North Macedonia*

[3] *Ss Cyril and Methodius University, Faculty of Agricultural Sciences and Food, Str. 16-ta Makedonska Brigada 3, 1000 Skopje, Republic of North Macedonia*

* *Corresponding author: E-mail: shigor@pmf.ukim.mk*

Received: 05-01-2023

Abstract

Due to its transferability, the soil has been commonly used as evidence in criminal investigations. In this work, 172 soil samples were taken from five urban parks from the town of Tetovo (North Macedonia) and from additional four rural locations in its vicinity. The soil samples were examined using X-ray powder diffraction. The collected diffractograms were used for development of classification models based on supervised self-organizing maps for determination of their origin. The examination of generalization performances of the developed models showed that they were able to correctly classify between 95.6 and 97.8% of the samples from the independent test set. The influence of the weather and the seasonal changes on the composition of the soil was also examined. For this purpose, three years after the initial soil samples were collected, additional 28 samples were analyzed from different locations. The best models presented in this work were able to successfully classify 27 of these additional samples.

Keywords: Chemometrics, soil analysis, forensic analysis, X-ray powder diffraction

1. Introduction

An aerial view, on the agricultural land, in the right season, could reveal plots of land with variety of colors. It is amazing how relatively close these plots of land could be but their color can still differ. For a chemist, the difference in the soil color or nuance of the soil color is the first clue that could lead to a conclusion about possible difference in its chemical composition. The soil color is influenced by the minerals, the water and the organic matter present in it. For example, the soils with high concentration of calcium tend to be white, those with high concentration of iron are reddish, and those high in humus are dark brown to black.[1] The soil color is significant indicator of the chemical composition and a Munsell color chart could be enough for classification of soils for agricultural purposes. However, this approach is not enough for forensic investigations. Due to its high mineral content X-ray powder diffraction analysis of the soils can provide additional and sufficient data which could be used as forensic evidence. The idea behind the soil as a forensic evidence comes from its divisibility and transferability.[2] Namely, the soil taken from a perpetrator's shoes, car tires or tools, can be linked to a crime scene.[3]

The soil samples have specific chemical and physical composition that has been analyzed with a variety of analytical methods. Scanning electron microscopy has been applied to identify unusual particles. This technique has also been coupled with EDS.[4–6] The potential use of the soil as forensic evidence has been studied with atomic absorption spectrometry,[7,8] inductively coupled plasma, with mass spectrometry,[9,10] gas chromatography coupled with mass spectrometry,[11,12] Raman spectroscopy,[13–15] infrared absorption spectroscopy and infrared reflectance spectroscopy.[16–23] The IR spectroscopy has been used for determination of both (1) organic and (2) mineral components of the soil.[18,24,25]

X-ray powder diffraction (XRD) is a nondestructive technique that can provide a rapid and accurate miner-

alogical analysis of multicomponent mixtures without a need for extensive sample preparation.[26] In addition to this, we have to state here that, for forensic purposes, X-ray powder diffraction has been previously used.[19,20,27,28]

In this work, we present our efforts for the development of chemometric method based on supervised self-organizing maps (SOM) used for classification of soil samples for determination of their origin for forensic purposes.[29] Chemometrics by itself has already found its application in forensic science.[30–39] In our previous work, we successfully developed models for classification of urban soils for forensic analysis from five locations using infrared spectroscopy as an experimental technique.[40] However, the signals (the bands) in infrared spectra are highly overlapped most of the time. That was the reason why, in order to obtain successful classification of the samples, we used one-against-the-rest approach. The performances of the one-against-the-rest approach were considerably better compared to a single model approach used for classification of all five types of urban soil samples.[40]

Compared to the signals in the infrared spectra, the signals obtained by X-ray powder diffraction are less overlapped. Having this information in mind, in this work we decided to use X-ray powder diffraction as an experimental technique for classification of samples from nine locations. Five urban location from the town of Tetovo, and four rural locations.

2. Experimental

For this purpose, the soil samples were collected from (1) five different parks from the town of Tetovo (North Macedonia) and from (2) four additional rural locations in its vicinity. These locations are presented on Table 1 and in Figure 1.

Table 1. Locations from which the soil samples were collected, the description of the locations and their labels.

Label	Location
A	Intercity Bus Station Park
B	House of Culture Park
C	Colorful Mosque Park
D	State University of Tetovo Park
E	Moša Pijade High School Park
F	Village of Džepčište (north exit of the town)
G	Near the tollbooths on the highway Skopje–Tetovo (east exit)
H	Near the village of Gajre (west exit of the town)
I	Near the village of Dolno Palčište (south exit of the town)

Three of the five parks (locations: A, B and C) are located at approximate distances between 1 and 1.5 km. The distance between these three parks from the remaining two (locations: D and E) is about 2.5 km. It is also important to note that the distance between the remaining two parks (D – State University of Tetovo Park and E – Moša Pijade High School Park) is about 250–300 m. These two parks were selected in this way in order to examine whether the smaller distance will have influence on the performances of the classification models due to the possible similarities of the composition of the soils. The distances between center of the town and the remaining four rural locations (F, G, H and I) are between 4.5 and 5 km.

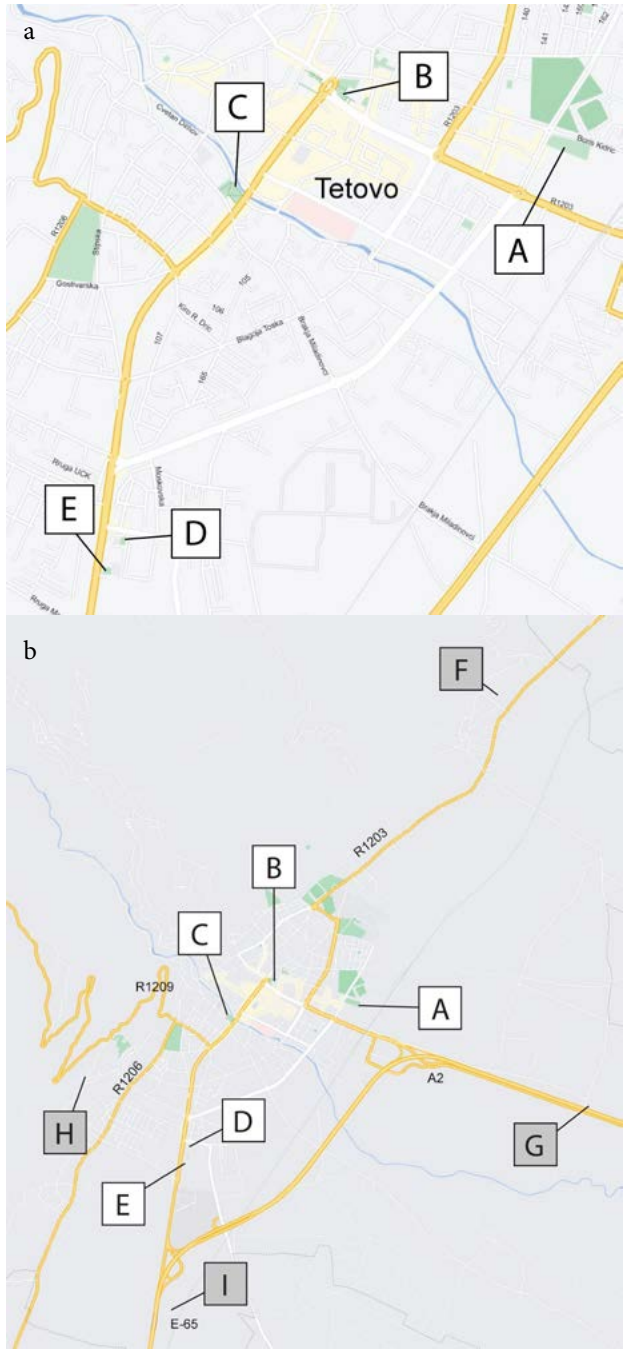


Figure 1. The nine locations (a) inside and (b) around town of Tetovo. Locations A, B, C, D and E represent the five parks located in the town, while locations F, G, H and I represent locations from which rural samples were collected.

Total number of 144 soil samples were collected from all nine locations. Sixteen samples were collected from each location. Each of the sixteen samples was taken from the predetermined square grid with area of about 9 m^2 (Figure 2). The distance between the sampling positions on the grid was about 1 m. The samples were collected from top soil layer (10 cm).

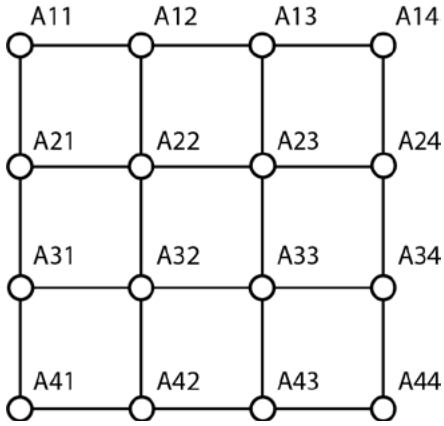


Figure 2. The grid used for soil sampling. The indices which are used on this grid are also used for labeling the different samples taken from the same location.

In order to properly analyze the results, it was important to label the collected samples in a systematic manner. For this purpose, each of the samples were labeled as shown in Figure 2. On this figure each of the nodes, which correspond to different samples taken on location A, were labeled as: A11, A12, A13, A14; A21, A22 … A44.

Three years later (in the autumn of 2019), additional 28 new soil samples were collected from seven of the nine locations. The selected locations were B, D, E (from three parks in Tetovo) as well as all four rural locations (F, G, H and I). This was performed in order (1) to validate our models with new data, but also (2) to examine the influence of the seasonal changes and the weather on the composition of the soil in these parks.

In addition to this, we did not get any information form the local Police Department that these locations were scene of the crime in order to validate our models with their data.

Also, it has to be stated that at this point of our experimental work (in the autumn of 2019), we did not take new samples from the locations A and C because, during these three years, larger horticultural interventions were performed in these two parks by the Municipality of Tetovo.

The samples of the soil were dried at ambient conditions for few days. They were sieved with 20 mesh Tyler sieve. The material that passed the sieve was collected and marked. Collected samples were dried at temperature of 110 °C. The dried samples were kept in desiccator. In addition to this, before the diffractograms were recorded the samples were powdered in a mortar with a pestle.

The X-ray diffractograms of all samples were recorded using the Rigaku Ultima IV powder X-ray diffractometer in the Bragg-Brentano geometry with CuK_α radiation (λ = 1.54178 Å) at room temperature. The sample holder was a 2 mm thick glass plate with dimensions 60 mm × 35 mm and 20 mm × 20 mm depression for the sample. The depression in the holder was filled with sample and was flattened. The mass of the analyzed samples was approximately 200 mg.

Diffraction patterns were measured in the 2ϑ range from 5° to 60° with a step size of 0.02° and scanning speed of 20° per minute. The accelerating voltage and the electric current were set to 40 kV and 40 mA, respectively. The divergence slit parameter (DivSlit) was 2/3 degrees, the height limiting slit parameter (DivH.L.Slit) was 10 mm and the anti-scatter slit parameter (SctSlit) was 8.0 mm.

2. 1. Data Pre-processing and Algorithms Used

In order to properly prepare the data for optimization of the SOMs, it was necessary to pre-process the obtained diffractograms. The experimentally collected diffractograms of the soil samples were stored in a single data matrix. The data matrix was composed of 172 diffractograms (rows) and 2749 intensities at different 2θ values (columns).

In this study, the first step in the pre-processing (see Figure 3) was the baseline correction of the diffractograms. After that, baseline corrected diffractograms were normalized to unit area under the curve. Further, in order to make the (1) optimization faster, (2) to reduce the noise in the diffractograms as well as (3) to reduce the number of data points because most of the intensities on different 2θ values are correlated, data reduction was performed by averaging each consecutive non-overlapping interval composed of 11 intensities using the following formula:

$$d'_{im} = \frac{\sum_{j=(m-1)\cdot 11+1}^{m\cdot 11} d_{ij}}{11} \quad (1)$$

d_{ij} in the equation (1) represents the data point from pre-processed matrix consisting of diffractograms, i – is the sample number, j – represent the intensity values at different 2θ values, whereas d_{im} is data point from i-th sample and m-th column in the reduced data matrix. Using this approach, the number of intensities were reduced from 2749 down to 259 (Figure 3). The diffractograms obtained using this data pre-processing procedure were stored in single data matrix (***D***).

The previously obtained data matrix (***D***) was further reduced using principal component analysis (PCA). In order to perform PCA the variables (the columns in ***D***) were auto-scaled. Using PCA we were able to extract the largest fraction of the information stored in ***D*** into small number of principal components. Finally, the obtained princi-

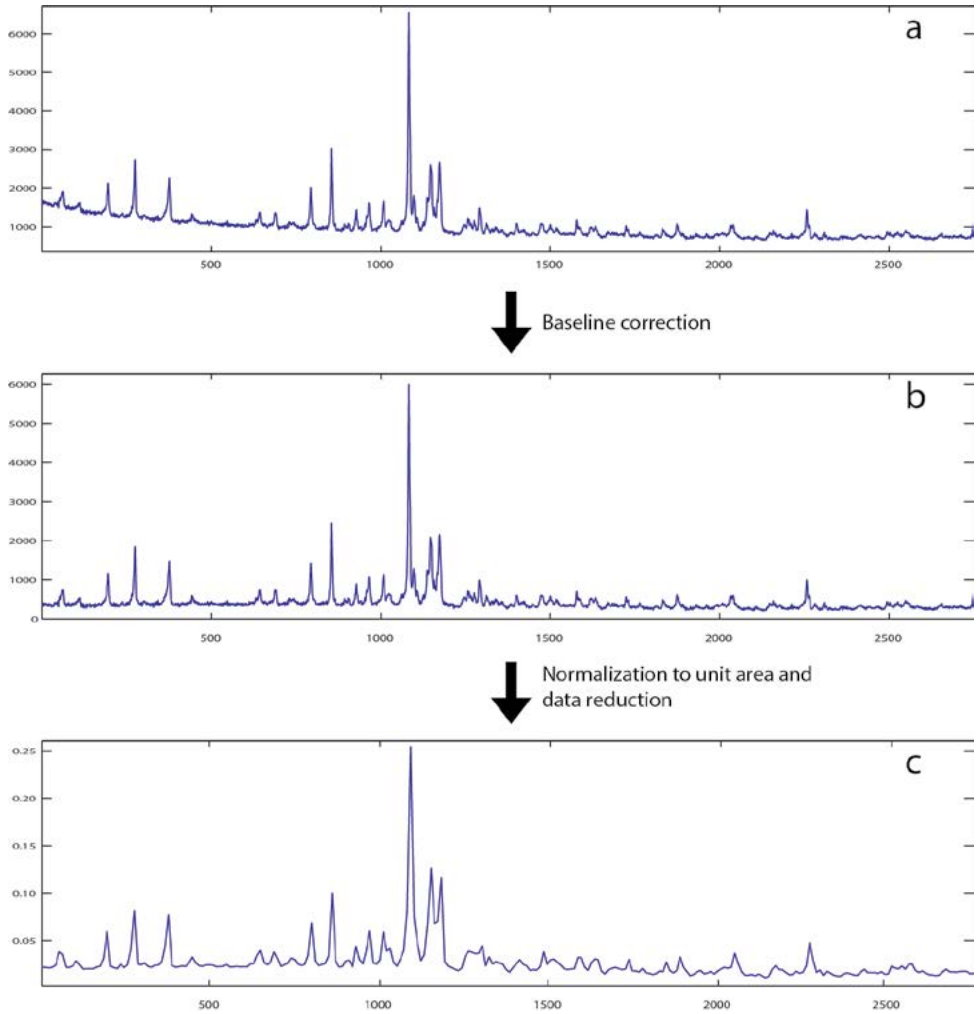


Figure 3. Illustration of the main steps of the preprocessing of the experimentally obtained diffractograms. a – original diffractogram; b – baseline corrected diffractogram; c – normalized diffractogram to unit area under the curve with data points reduced down to 259 using equation (1).

pal components were used for training of the supervised self-organizing maps.

2. 2. Supervised Self-organizing Maps

According to its inventor, Teuvo Kohonen, self-organizing maps (alternative names: Kohonen maps or Kohonen neural networks) were originally developed as algorithm for unsupervised learning.[29,41] In chemistry and related sciences, most often the unsupervised version of this algorithm is used.[42,43] Today, the unsupervised variant of the algorithm is simply called *self-organizing maps* or *Kohonen neural networks*. While the supervised version of the algorithm, which is not used as frequently as the previously mentioned version, is called *supervised self-organizing maps.* The supervised version of SOM is used in cases when there is not a clear separation among different types of samples.[29] In order to adapt the SOM algorithm for classification purposes (see Figure 4), it is necessary to augment each training vector ($\boldsymbol{d}_s$) with unit vector ($\boldsymbol{d}_u$) assigned into one of the nine classes of samples in our case (Figure 4a). This augmentation of the training set vectors (samples) with $\boldsymbol{d}_u$ helps in better separation of the different types of samples during the training.

During the prediction phase the weight levels which correspond to the unit vectors ($\boldsymbol{w}_u$) are removed (Figure 4b). In other words, for each sample in the training set $\boldsymbol{d}_s$ the corresponding $\boldsymbol{d}_u$ must be used during training. While during the prediction phase, for the unknown samples – $\boldsymbol{x}$ only, $\boldsymbol{x}_s$ part is compared with the corresponding part of the weight vectors ($\boldsymbol{w}_u$) of the trained supervised SOM.

Supervised self-organizing maps were implemented in Matlab[44] programming language using SOM Toolbox developed by J. Vesanto[45–47] on a Windows computer.

2. 3. Genetic Algorithms

In this work, the optimization of the supervised SOM models was performed in automated manner using genetic algorithms. Genetic algorithms have been used successfully for solving different problems in the field of chemistry and related sciences since the beginning of the last decade of

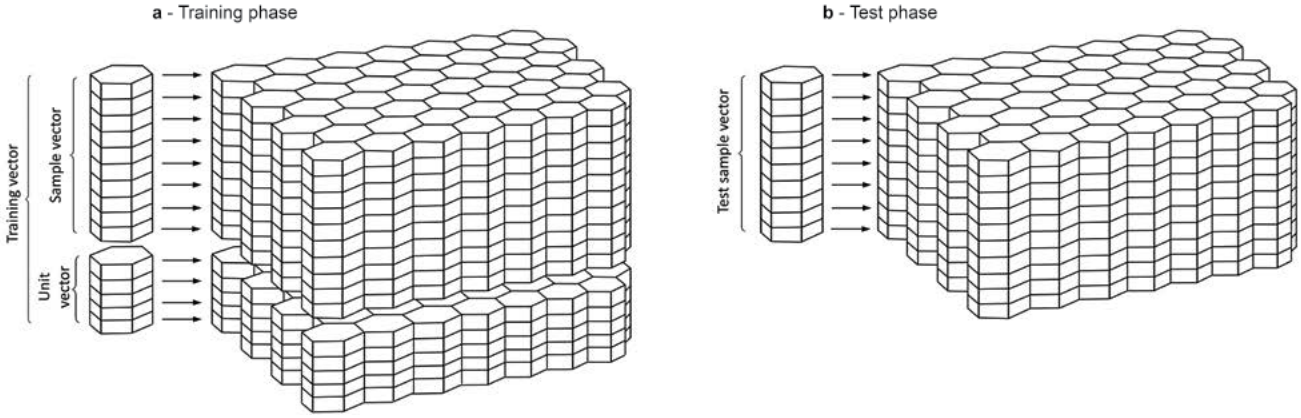


Figure 4. Illustration of the structure of the supervised self-organizing maps during the phases of (a) training and (b) prediction. As shown on this figure in the training phase the vector which represents samples is augmented with different unit vector for five different classes of samples. While using the supervised SOM for prediction purposes, the weight levels that correspond to unit vectors are removed.

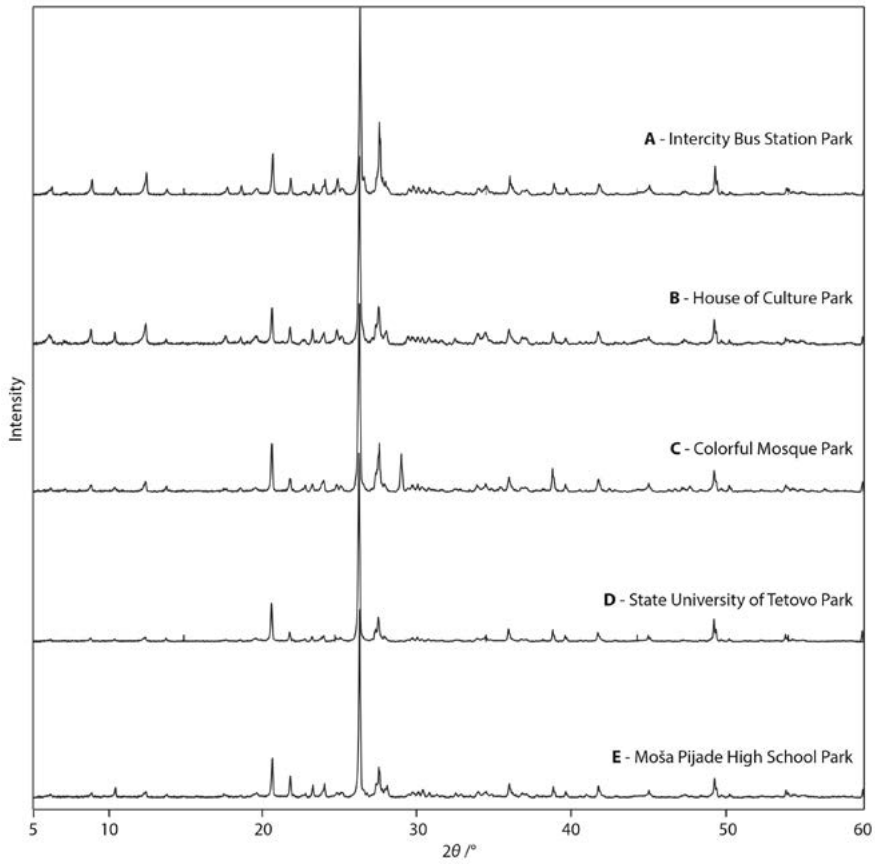
the 20th century.[48–51] The theory of genetic algorithms has been described several times in the chemometric literature during the same decade.[52–54] We have to mention here that, most often, in chemometrics GAs have gave been used for selection of variables.[52–54] In our work, we use GAs not only as a variable selection tool but also in order to find optimal parameters of the developed models.[40,55,56]

Genetic algorithms were also implemented in Matlab programming language. For this purpose, Genetic Algorithm Toolbox developed at University of Sheffield was used.[57]

3. Results and Discussion

3. 1. Main Mineralogical Components

It is important to state that diffractograms from different locations are similar (see Figure 5 for the urban samples and Figure 6 for the rural samples). For more detailed comparison all diffractograms are available in the



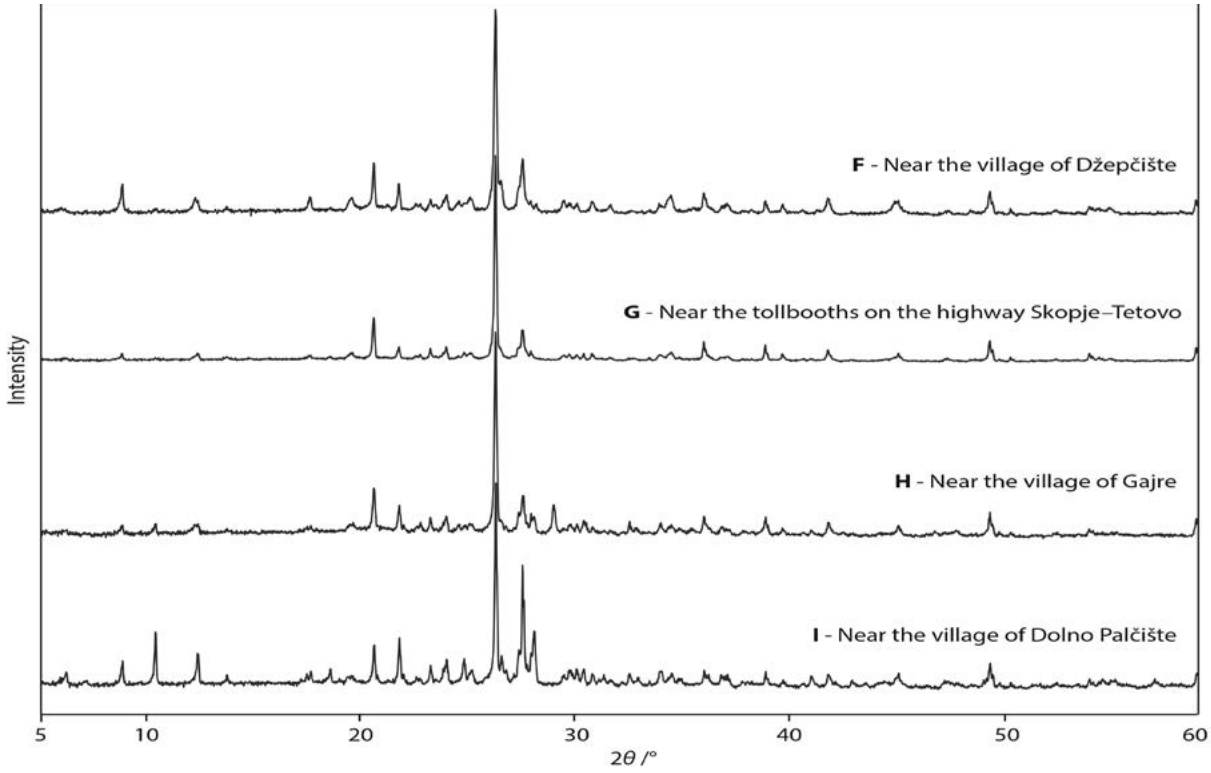


Figure 6. X-ray powder diffractograms for the selected samples from four rural locations. (The files with all samples from these five locations are given as Supplementary Material.)

Supplementary Material together with some additional figures. One can notice that the main differences in the diffractograms are in the relative intensities of the major components of the soil samples.

Main mineralogical components of the analyzed soil samples were found by comparing the obtained diffractograms with the diffractograms stored in COD[58–64] and PDF-2 databases[65] using Match! software.[66] The results show that the main component in the samples is SiO_2 (silicon dioxide) in a form of quartz. Three additional minerals with lower mass fractions were also detected as possible constituents of the samples. It is interesting to state that they all have an empirical formula $MAlSi_3O_8$, where M represents potassium or sodium. When M represents sodium, the mineral is known as albite. In the case when M is potassium, the mineral is orthoclase (which can make solid solutions with albite). The third mineral present in our samples is, probably, the high-temperature polymorph of albite known as sanidine.

It is important to state here that there are diffraction peaks in the recorded diffractograms which do not correspond to the previously mentioned minerals. These signals probably belong to the additional mineral components. However, due to the fact that their mass fraction and consequently the intensities of their signals are weak we were not able to identify them using previously described approach. Also, earlier we pointed out that the diffractograms from all locations are similar (Figure 5 and Figure 6). Most of the differences among the diffractograms are in the regions with diffraction peaks that have smaller intensities. This is one of the reasons why genetic algorithm was used for variable selection. In cases like this, optimization using GA performs selection of the intensities at 2θ values which can help in finding better classification models and, at the same time, it eliminates the intensities at 2θ values which are similar for all samples.

3. 2. Principal Component Analysis

The principal component analysis (PCA) which was performed on the auto-scaled data matrix showed us that about 92% of the variance in the pre-processed diffractograms was captured by first 16 principal components (PCs). As previously stated, these PCs were used for development of the SOM models which will be able to classify the samples according to their geographic origin. However, before the development of the models started, we used PCA as an auxiliary tool in order to evaluate whether the samples from different location were at least partially separated. This is important since the main goal of this work is determination of the origin of the soil samples. The first two principal components (labeled as: PC1 and PC2) are presented on Figure 7a. The first (PC1) and third (PC3) principal components are presented on Figure 5b. A careful examination of these two figures (Figure 7a and b) shows that all samples taken from the rural locations (F, G,

H and I) are grouped in the first and fourth quadrant. Having in mind that these two figures are projections of the three-dimensional space defined by PC1, PC2 and PC3, it is easy to see that the samples from these four locations form well separated clusters. The reason for this might be the fact that most of the rural locations from which the samples were collected are at distances larger than 4 km. Only the distance between locations H and I is smaller than 4 km.

In the second and third quadrant, the remaining samples from the urban location are grouped. Here it could be seen that there is a partial overlap among the samples from all urban locations. Also, most of the samples from location C are well separated from the remaining samples (see Figure 7b). Likewise, the reason for larger overlap between these clusters might be the fact that these locations are closer one to another. As previously stated, the distances between the parks (1) next to the Intercity Bus Station (labeled as A), (2) House of Culture Park (labeled as B) and (3) Colorful Mosque Park (labels: C) vary between 1 and 1.5 km. While the distance between the park labeled as D (State University of Tetovo Park) and park labeled as E (Moša Pijade High School Park) is only about 300 m.

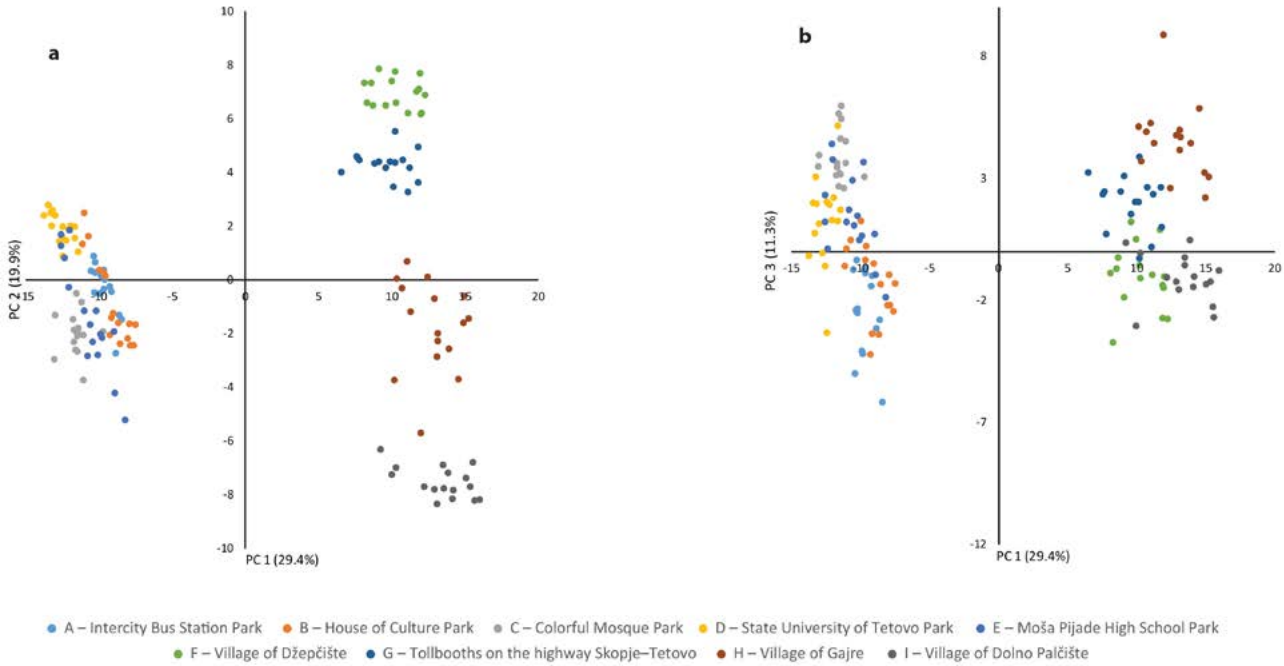


Figure 7. The soil samples presented in the space defined by first (PC1) and second (PC2) as well as the first (PC1) and third (PC3) principal component obtained from autoscaled matrix.

3. 3. Optimization of the SOM Models Using Genetic Algorithms

As earlier stated, the optimization of the models based on supervised self-organizing maps was performed using GAs. For optimization purposes, we used populations composed of 100 binary chromosomes. The initial values of genes in the chromosomes were randomly generated. Different parts of these chromosomes were responsible for decoding different parameters for the supervised SOMs. In this case 259 genes were used for variable selection. After the variable selection was performed, PCA was applied on the selected intensities. The number of PCs used during the optimization varied between 1 and 16. For this purpose, four additional genes were allocated in the chromosomes. Eight more genes were allocated for decoding the width and length of the supervised SOM. These two parameters were changed in the interval between 7 and 22. Four more genes were dedicated for selection of the number of epochs in the rough training phase, which is performed in larger neighborhood and larger learning rate. The number of epochs here was changed in the interval between 10 and 25. Finally, for the number of epochs in the fine-tuning phase, additional seven genes were selected. The use of seven binary genes could produce the maximal number of epochs 127 and the minimal number of epoch zero. In order to avoid the zero which appears here, but also to be sure that number of epochs in the fine-tuning phase is larger than that obtained in the rough training phase, the number of epochs obtained by these genes was increased by the number of epochs obtained for the rough training phase.

Before the optimization with GA started, we had to properly separate the data set into training and test data sets in order to obtain good generalization performances. The training data set was used for the optimization of

Table 2. Six selected models with best performances (misclassified samples). The size of the network, the training parameters, the number of principal components used for training of supervised SOMs, as well as the labels of the misclassified samples are also presented.

Model	Size of the SOM		Training epochs		No of principal components	Misclassified samples				Labels for the misclassified samples					
										Test set			Real samples		
	Width	Length	Rough	Fine		Training	Cross validation	Test	Real	A42	D41	E11	E3	E4	H4
1	14	11	16	228	8	0	0	2	1		E	C	D		
2	15	9	17	131	9	0	1	2	1		E	C			G
3	17	9	13	260	6	0	1	2	2	B	E		D	A	
4	8	20	16	101	7	0	3	2	1		E	C			G
5	12	13	15	91	6	0	1	1	2	B				B	G
6	21	8	17	163	7	0	1	1	1			C	D		

the models. During the optimization, the generalization performances of the models were controlled by cross-validation. After the optimization was finished, additional validation of the models was performed using the test set composed of samples which were not used during the training of the models. The original data set (***D***) was divided into training and test set using Kennard-Stone algorithm separately for each location.[67] Using this algorithm, five of the sixteen diffractograms from each location were selected to be part of the test set. The remaining eleven samples were stored into the training data set. As a result, our training set was composed of 99 diffractograms and the test set was composed of the remaining 45 diffractograms.

The entire search for optimal classification model performed by GA was repeated several times. Using this approach, we were able to obtain more than 100 models with good generalization performances. Some of the best models are presented on Table 2. The criteria for selection of these models for presentation were: (1) The size of the SOMs should be different; (2) If the only difference between the models were in number of the training epochs, then the one with smaller number of epochs was selected; (3) Finally, the most important criteria was the performances on the independent test set.

The examination of the results obtained for the test set (presented in Table 2) shows that three soil samples are most often misclassified. Two of these samples (labeled as D41 and E11) are from the locations that are at a distance of about 300 m. These two locations are: D – State University of Tetovo Park and E – Moša Pijade High School Park. As an illustration, the trained supervised SOM, which corresponds to model 1 in Table 2, is presented on Figure 6. Percentage of incorrectly classified samples from the test set which was used for examination of the generalization performances of the trained SOM vary between 2.2% for the model number 6 up to 4.4% for all other models presented in Table 2.

In our previous work, when we developed different models for classification of the urban soils based on infrared spectroscopy, the samples from these two parks were most often misclassified, probably due to smaller difference in the composition of the soils on these two locations.[40] In this case, the sample D41 is misclassified as a sample from Moša Pijade High School Park (label: E). However, the second of these samples (label: E11) is classified together with the samples taken from the park which is in the neighborhood of the Colorful Mosque (labels: C).

The third misclassified sample was taken from Intercity Bus Station Park (labels: A). This sample was classified together with the samples from House of Culture Park (labels: B) by two of the presented models.

Compared to the results which we obtained using infrared spectra of the urban soils, where only one sample was misclassified, we have to state that, in that case, due to higher overlap between signals in the infrared spectra there we were not able to develop good classification models which will be able to classify all five urban soils samples.[40] In that case, as stated earlier, we were forced to use *one-against-the-rest* approach and, consequently, we developed five separate models in order to obtain good classification models for all five locations.

Due to exposure to (1) seasonal changes, (2) biological processes and (3) the changing weather conditions, the composition of the soil is slowly changing. Sometimes, during the forensic investigations, by order of court or in the cases where the crime has been detected few years after it has been committed, in order to perform reliable detection of the origin of the soils, it is important to know how these variations in composition of the soils could influence the results. In order to examine the influence of the previously mentioned factors on the performances of the models developed here, three years after the initial samples were collected four more samples were collected from seven of nine original locations used in this study. The seven selected locations are labeled: B, D, E, F, G, H and I (see Table 1 for more details). Total number of newly collected samples was 28. These samples were treated in a same way as the initial 144 samples from all nine locations (see Experimental part).

As shown in Figure 8 which represents model 1 (presented in Table 2), only one sample from location E (with label E3) is misclassified as a sample from the nearby park (State University of Tetovo Park).

For the remaining models, the misclassified samples are also presented in Table 2. Here one can see that only

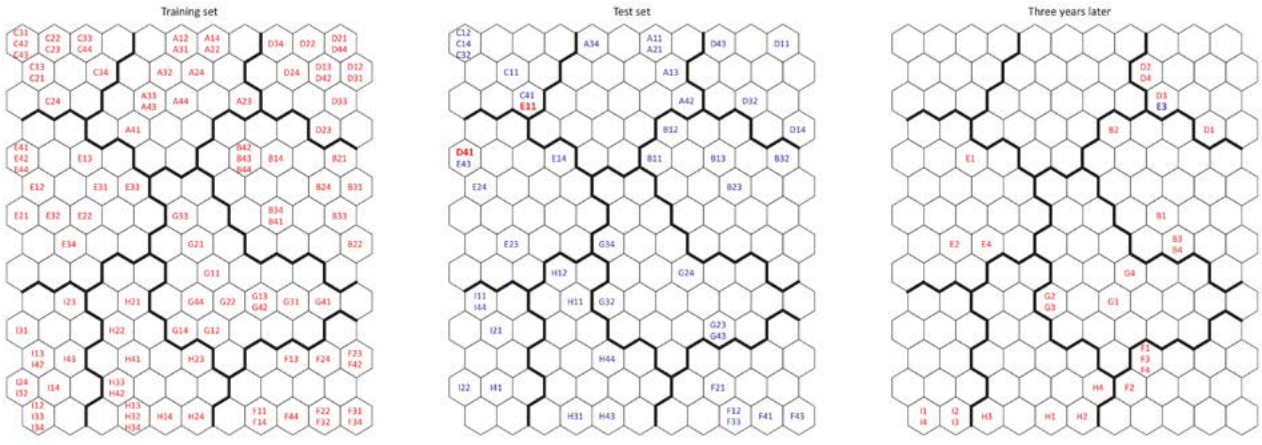


Figure 8. Supervised self-organizing map which corresponds to model 1 in Table 2 labelled with (a) samples from the training set, (b) samples from the test set and (c) with samples collected three years later in order to examine the influence of the weather and the seasonal changes on the composition of the soil on different locations. The misclassified samples are labelled with bold characters in different color than the remaining labels.

three of these samples are misclassified by more than one model. One of these samples (E3) was already discussed. The remaining two samples are labeled as E4 and H4. Two of the samples are from Moša Pijade High School Park (E3 and E4). Here the sample labeled as E4 is mapped once in the part of the supervised SOM that serves for recognition of the samples from location A and the second time it is misclassified together with the samples from location B.

4. Conclusion

In this study 144 samples from five urban and four rural locations were analyzed using supervised self-organizing maps. As a tool for automated search for the best models, genetic algorithms were used. Performances of the models during the optimization were controlled using cross-validation. Further, the generalization performances were examined using the test set which was not used during the training. The best models obtained and presented in this study were able to correctly classify between 95.6 and 97.8% of the test samples.

Having in mind that in our previous work, where the classification was performed using infrared spectra of the analyzed samples, due to the highly overlapping signals we had to develop five different models for successful classification of the samples from five urban locations. In this study, probably because the signals from the minor components of the analyzed soils in the X-ray diffractograms were well separated and, with the help from GA, we were able to select intensities which could help in better discrimination of the soil sample we were able to successfully classify most of the samples from all nine locations with a single model.

The performances of the models obtained here are comparable to those obtained in our previous work.[40] However, there in order to perform successful classification of the samples from the five locations there we had to develop separate models for each location. While here, using X-ray diffractograms of the samples, with only one model, we were able to correctly classify between 95.6 and 97.8% of the samples.

As previously stated in this study, it is also important to check the robustness of the model on the changes of the composition of the samples due to the changes of the environment. For this purpose, additional 28 samples were collected from seven locations (B, D, E, F, G, H and I). The best models presented here are capable of correctly classifying 27 of 28 samples collected three years after the initial soil samples were analyzed.

5. References

1. P. R. Owens, E. M. Rutledge, *Morphology,* chapter in *Encyclopedia of soils in the environment*, D. Hillel (Ed.), Elsevier, Amsterdam, **2005**.
2. K. Inmana, N. Rudin, *Forensic Sci. Int.* **2002**, *126*, 11–16. **DOI:**10.1016/s0379-0738(02)00031-2.
3. D. Werner, C. Burnier, Y. Yu, A. R. Marolf, Y. Wang, G. Massonnet, *Sci. Justice* **2019**, 59, 643–653. **DOI:**10.1016/j.scijus.2019.07.004
4. K. Pye, D. Croft, *Forensic Sci. Int.* **2007**, *165*, 52–63. **DOI:**10.1016/j.forsciint.2006.03.001.
5. B. G. Rawlins, S. J. Kemp, E. H. Hodgkinson, J. B. Riding, C. H. Vane, C. Poulton, K. Freeborough, *J. Forensic Sci.* **2006**, *51*, 832–845. http://doi.10.1111/j.1556-4029.2006.00152.x
6. S. Cengiza, A. C. Karaca, I. Çakır, H. B. Üner, A. Sevindik, *Forensic Sci. Int.* **2004**, *141*, 33–37. **DOI:**10.1016/j.forsciint.2003.12.006.
7. P. A. Bull, A. Parker, R. M. Morgan, *Forensic Sci. Int.* **2006**, *162*, 6–12. **DOI:**10.1016/j.forsciint.2006.06.075.
8. R. M. Morgan, P. A. Bull, Environ. *Forensics* **2006**, *7*, 325–334. **DOI:**10.1080/15275920600996248.
9. L. Arroyo, T. Trejos, T. Hosick, S. Machemer, J. R. Almirall, P.

R. Gardinali, *Environ. Forensics* **2010**, *11*, 315–327. **DOI:**10.1080/15275922.2010.494949.
10. L. Reidy, K. Bu, Murrell, G. James, V. Cizdziel, *Forensic Sci. Int*. **2013**, *233*, 37–44. **DOI:**10.1016/j.forsciint.2013.08.019.
11. J. M. L. Mazzettoa, V. F. Meloa, E. J. Bonfleura, P. Vidal-Torradob, J. Dieckowa, *Sci. Justice* **2019**, *59*, 635–642. **DOI:**10.1016/j.scijus.2019.07.003.
12. P. M. Medeiros, B. R. T. Simoneit, *J. Sep. Sci*. **2007**, *30*, 1516–1536. **DOI:**10.1002/jssc.200600399.
13. S. C. Jantzi, J. R. Almirall, *Anal. Bioanal. Chem*. **2011**, *400*, 3341–3351. **DOI:**10.1007/s00216-011-4869-7
14. B. W. Kammrath, A. Koutrakos, J. Castillo, C. Langley, D. H. Jones, *Forensic Sci. Int*. **2018**, *285*, e25–e33. **DOI:**10.1016/j.forsciint.2017.12.034
15. K. Ritz, L. Dawson, D. Miller (Eds.), *Criminal and Environmental Soil Forensics*. Soil Forensics International, Edinburgh Conference Centre, Springer, **2008**.
16. X. Xu, C. Du, F. Ma, Y. Shen, J. Zhou, *Forensic Sci. Int*. **2020**, *310*, 110222. **DOI:**10.1016/j.forsciint.2020.110222
17. R. J. Cox, H. L. Peterson, J. Young, C. Cusik, E. O. Espinoza, *Forensic Sci. Int*. **2000**, *108*, 107–116. **DOI:**10.1016/S0379-0738(99)00203-0
18. B. A. Weinger, J. A. Reffner, P. R. De Forest, *J. Forensic Sci*. **2009**, *54*, 851–856. **DOI:**10.1111/j.1556-4029.2009.01064.x
19. L. A. Dawson, S. Hillier, *Surf. Interface Anal*. **2010**, *42*, 363–377. **DOI:**10.1002/sia.3315
20. M. D. Suarez, R. J. Southard, S. J. Parikh, *J. Forensic Sci*. **2015**, *60*, 894–905. **DOI:**10.1111/1556-4029.12762.
21. J. M. Soriano-Disla, L. J. Janik, R. A. V. Rossel, L. M. Macdonald, M. J. McLaughlin, *Appl. Spectrosc. Rev.* **2014**, *49*, 139–186. **DOI:**10.1080/05704928.2013.811081
22. V. Sharma, J. Yadav, R. Kumar, D. Tesarova, A. Ekielski, P. K. Mishrad, *Vib. Spectrosc*. **2020**, *110*, 103097. **DOI:**10.1016/j.vibspec.2020.103097.
23. V. Sharma, S. Bhardwaj, R. Kumar, *Vib. Spectrosc*. **2019**, *101*, 81–91. **DOI:**10.1016/j.vibspec.2019.02.006.
24. R. R. E. Artz, S. J. Chapman, A. H. J. Robertson, J. M. Potts, F. Laggoun-Défarge, S. Gogo, L. Comont, J. R. Disnar, A. J. Francez, *Soil Biol. Biochem.* **2008**, *40*, 515–527. **DOI:**10.1016/j.soilbio.2007.09.019.
25. Š. Matějková, T. Šimon, *Plant Soil Environ*. **2012**, *58*, 192–195. **DOI:**10.17221/317/2011-PSE
26. S. Bashir, J. Liu, *Advanced nanomaterials and their applications in renewable energy,* Elsevier, Amsterdam, **2015**.
27. L. V. Prandel, V. F. Melo, A. M. Brinatti, S. C. Saab, F. A. S. Salvador, *J. Forensic Sci*. **2018**, *63*, 251–257. **DOI:**10.1111/1556-4029.13476.
28. A. Ruffella, P. Wiltshireb, *Forensic Sci. Int*. **2004**, *145*, 13–23. **DOI:**10.1016/j.forsciint.2004.03.017.
29. T. Kohonen, *Self-Organizing Maps*, 3rd Edition, Springer, Berlin, **2001**.
30. R. Chauhan, R. Kumar, V. Sharma, *Microchem. J*. **2018**, *139*, 74–84. **DOI:**10.1016/j.microc.2018.02.020
31. V. Sharma, R. Kumar, *TrAC – Trends Anal. Chem.* **2018**, *107*, 181–195. **DOI:**10.1016/j.trac.2018.08.006
32. R. Kumar, V. Sharma, *TrAC – Trends Anal. Chem.* **2018**, *105*, 191–201. **DOI:**10.1016/j.trac.2018.05.010
33. M. Baron, J. G. Rodriguez, R. Croxton, R. Gonzalez, R. Jimenez, *J. Appl. Spectrosc*. **2011**, *65*, 1151–1161. **DOI:**10.1366/10-06197
34. R. Chauhan, R. Kumar, P. K. Diwan, V. Sharma, *Forensic Chem*. **2020**, 100191. **DOI:**10.1016/j.forc.2019.100191
35. M. E. Sigman, M. R. Williams, *WIREs Forensic. Sci*. **2020**, *2*, e1368. **DOI:**10.1002/wfs2.1368
36. C. S. Leea, T. M. Sung, H. S. Kim, C. H. Jeon, *J. Anal. Appl. Pyrol.* **2012**, *96*, 33–42. **DOI:**10.1016/j.jaap.2012.02.017
37. M. I. Kaniu, K. H. Angeyo, *Geoderma* **2015**, *241–242*, 32–40. **DOI:**10.1016/j.geoderma.2014.10.014.
38. N. C. Thanasoulias, E. T. Piliouris, M. S. E. Kotti, N. P. Evmiridis, *Forensic Sci. Int*. **2002**, *130*, 73–82. **DOI:**10.1016/s0379-0738(02)00369-9.
39. M. Baron, J. Gonzalez-Rodriguez, R. Croxton, R. Gonzalez, R. Jimenez-Perez, *Appl. Spectrosc.* **2011**, *65*, 1151–1161. **DOI:**10.1366/10-06197.
40. H. Idrizi, M. Najdoski, I. Kuzmanovski, *J. Chemom*. **2021**, *35*, e3328. **DOI:**10.1002/cem.3328
41. T. Kohonen, *Computer*, **1988**, *21*, 11–22.
42. J. Zupan, M. Novic, I. Ruisánchez, *Chemometr. Intell. Lab. Syst*. **1997**, *38*, 1–23. **DOI:**10.1016/S0169-7439(97)00030-0
43. J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*. WCH: Weinheim, **1999**.
44. MATLAB 7.12 (R2011a), 1984–2011, MathWorks.
45. J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, *SOM Toolbox for Matlab 5*, Technical Report A57, Helsinki University of Technology, **2000**.
46. http://www.cis.hut.fi/projects/somtoolbox/ (assessed: July 15, 2023).
47. J. Vesanto, *Intell. Data Anal*. **1999**, *6*, 111–126.
48. C. B. Lucasius, S. Werten, A. Vanaert, G. Kateman, M. J. J. Blommers, *Lect. Notes Comput. Sci*. **1991**, *496*, 90–97.
49. D. Jouan-Rimbaud, D. L. Massart, R. Leardi, O. E. de Noord, *Anal. Chem*. **1995**, *67*, 4295–4301. **DOI:**10.1021/ac00119a015
50. C. B. Lucasius, G. Kateman, *TrAC – Trends Anal. Chem.* **1991**, *10*, 254–261.
51. M. Bos, H. T. Weber, *Anal. Chim. Acta* **1991**, *247*, 97–105. **DOI:**10.1002/bip.360320107
52. R. Leardi, A. Lupiañez Gonzales, *Chemometr. Intell. Lab. Syst*. **1998**, *41*, 195–207.
53. H. Handels, T. Roß, J. Kreusch, H. H. Wolff, S. J. Pöppl, *Artif. Intell. Med*. **1999**, *16*, 283–297.
54. S. S. So, M. Karplus, *J. Med. Chem.* **1996**, *39*, 5246–5256.
55. I. Kuzmanovski, M. Trpkovska, B. Šoptrajanov, *J. Mol. Struct*. **2005**, *744–747*, 833–838.
56. N. Stojić, S. Erić, I. Kuzmanovski, *J. Mol. Graph. Model.* **2010**, *29*, 450–460.
57. A. Chipperfield, P. Fleming, H. Pohlheim, C. Fonseca, *Genetic algorithm toolbox user's guide*, University of Sheffield, Sheffield, **1994**.
58. A. Merkys, A. Vaitkus, A. Grybauskas, A. Konovalovas, M. Quirós, S. Gražulis, *J. Cheminformatics* **2023**, 15(25). **DOI:**10.1186/s13321-023-00692-1

59. A. Vaitkus, A. Merkys, S. Gražulis, *J. Appl. Crystallogr.* **2021**, 54(2), 661–672. **DOI:**10.1107/S1600576720016532
60. M. Quirós, S. Gražulis, S. Girdzijauskaitė, A. Merkys, A. Vaitkus, *J. Cheminformatics* **2018**, 10(23). **DOI:**10.1186/s13321-018-0279-6
61. A. Merkys, A. Vaitkus, J. Butkus, M. Okulič-Kazarinas, V. Kairys, S. Gražulis, *J. Appl. Crystallogr*. **2016**, 49. **DOI:**10.1107/S1600576715022396
62. S. Gražulis, A. Merkys, A. Vaitkus, M. Okulič-Kazarinas, *J. Appl. Crystallogr*. **2015**, 48, 85–91. **DOI:**10.1107/S1600576714025904
63. S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs, A. LeBail, *Nucleic Acids Res.* **2012**, 40, D420–D427. **DOI:**10.1093/nar/gkr900
64. S. Grazulis, D. Chateigner, R. T. Downs, A. T. Yokochi, M. Quiros, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, A. Le Bail, *J. Appl. Crystallogr*. **2009**, 42, 726–729. **DOI:**10.1107/S0021889809016690
65. ICDD, PDF-2 2011 (Database), S. Kalakkodu (Ed.), Newtown Square: Int. Centre Diffraction Data, **2011**.
66. Match! - Phase Analysis using Powder Diffraction, Version 3.x, Crystal Impact - Dr. H. Putz & Dr. K. Brandenburg GbR, Kreuzherrenstr. 102, 53227 Bonn, Germany, https://www.crystalimpact.de/match.
67. R. W. Kennard, L. A. Stone, *Dent Tech*. **1969**, *11*, 137–148.

Povzetek

Zaradi svoje prenosljivosti se tla pogosto uporabljajo kot dokazni material v kriminalnih preiskavah. V tej raziskavi smo odvzeli 172 vzorcev tal iz petih urbanih parkov v mestu Tetovo (Severna Makedonija) in iz dodatnih štirih podeželskih lokacij v njegovi bližini. Vzorce tal smo preiskali z uporabo X-žarkovne praškovne difrakcije. Zbrane difraktograme smo uporabili za razvoj klasifikacijskih modelov za določitev njihovega izvora, ki temeljijo na nadzorovanih samoorganiziranih mapah. Preiskava generalizacijske sposobnosti razvitih modelov je pokazala, da so bili zmožni pravilno klasificirati med 95,6 in 97,8 % vzorcev iz neodvisnega testnega niza. Preučili smo tudi vpliv vremenskih in obdobnih sprememb na sestavo tal. Za ta namen smo tri leta po začetnem zbiranju vzorcev tal analizirali dodatnih 28 vzorcev iz različnih lokacij. Najboljši modeli, predstavljeni v tej raziskavi, so bili zmožni uspešno klasificirati 27 od teh dodatnih vzorcev.